European Journal of Radiology 174 (2024) 111399

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Meta-analyses

Deep learning performance compared to healthcare experts in detecting wrist fractures from radiographs: A systematic review and meta-analysis

V. Hansen [a], J. Jensen [b,c], M.W. Kusk [a,d,e,f], O. Gerke [g,h], H.B. Tromborg [h,i], S. Lysdahlgaard [a,d,e,*]

[a] Department of Radiology and Nuclear Medicine, Hospital of South West Jutland, University Hospital of Southern Denmark, Esbjerg, Denmark
[b] Department of Radiology, Odense University Hospital, Odense, Denmark
[c] Research and Innovation Unit of Radiology, University of Southern Denmark, Odense, Denmark
[d] Department of Regional Health Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark
[e] Imaging Research Initiative Southwest (IRIS), Hospital of South West Jutland, University Hospital of Southern Denmark, Esbjerg, Denmark
[f] Radiography and Diagnostic Imaging, School of Medicine, University College Dublin, Belfield 4, Dublin, Ireland
[g] Department of Nuclear Medicine, Odense University Hospital, Odense, Denmark
[h] Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[i] Department of Orthopedic Surgery, Odense University Hospital, Odense, Denmark

ARTICLE INFO

Keywords:
Digital radiography
Wrist fractures
Convolutional neural network
Systematic review
Meta-analysis

ABSTRACT

Objective: To perform a systematic review and meta-analysis of the diagnostic accuracy of deep learning (DL) algorithms in the diagnosis of wrist fractures (WF) on plain wrist radiographs, taking healthcare experts consensus as reference standard.

Methods: Embase, Medline, PubMed, Scopus and Web of Science were searched in the period from 1 Jan 2012 to 9 March 2023. Eligible studies were patients with wrist radiographs for radial and ulnar fractures as the target condition, studies using DL algorithms based on convolutional neural networks (CNN), and healthcare experts consensus as the minimum reference standard. Studies were assessed with a modified QUADAS-2 tool, and we applied a bivariate random-effects model for meta-analysis of diagnostic test accuracy data.

Results: Our study was registered at PROSPERO with ID: CRD42023431398. We included 6 unique studies for meta-analysis, with a total of 33,026 radiographs. CNN performance compared to reference standards for the included articles found a summary sensitivity of 92% (95% CI: 80%–97%) and a summary specificity of 93% (95% CI: 76%–98%). The generalized bivariate I-squared statistic indicated considerable heterogeneity between the studies (81.90%). Four studies had one or more domains at high risk of bias and two studies had concerns regarding applicability.

Conclusion: The diagnostic accuracy of CNNs was comparable to that of healthcare experts in wrist radiographs for investigation of WF. There is a need for studies with a robust reference standard, external data-set validation and investigation of diagnostic performance of healthcare experts aided with CNNs.

Clinical relevance statement: DL matches healthcare experts in diagnosing WFs, which potentially benefits patient diagnosis.

1. Introduction

The most common type of interpretational errors made by physicians on musculoskeletal radiographs in emergency departments (ED) are missed fractures [1–3]. This can result in treatment delays and may lead to malunion or pseudoarthrosis with attendant morbidity [4]. Human and environmental factors can affect the interpretation of the radiograph, such as clinician inexperience, fatigue, distractions, poor viewing conditions, and time pressures. One study concluded that approximately one percent of all fractures in the ED were not correctly diagnosed [5].

Inexperienced physicians or those without specialization in musculoskeletal imaging have limited training in wrist fracture (WF) identification, especially with subtle presentations [6]. Conventional

Abbreviations: AI, Artificial intelligence; CNN, Convolutional neural network; DL, Deep learning; ED, Emergency department; MDCT, Multi detector computed tomography; SROC, Summary receiver operating characteristics curve; WF, Wrist fractures.

* Corresponding author at: Finsensgade 35, 6700 Esbjerg, Denmark.
E-mail address: Simon.Lysdahlgaard@rsyd.dk (S. Lysdahlgaard).

https://doi.org/10.1016/j.ejrad.2024.111399
Received 28 November 2023; Received in revised form 29 January 2024; Accepted 26 February 2024
Available online 27 February 2024

radiography is the first line medical imaging in diagnosing WF, where suboptimal positioning, technique and/or patient cooperation may affect the radiograph [7,8].

Automation of WF diagnosis by deep learning (DL) could potentially augment the work of physicians, and the performance of DL models has significantly progressed in the last decade, where image classification error rates improved significantly [9]. Recent evidence points to the potential usefulness of convolutional neural networks (CNN) in classifying medical radiographs [10–14]. However, small datasets and overfitting are two challenges in applying CNNs, where the performance of an algorithm should be generalizable to unfamiliar data [15]. A systematic review and meta-analysis by Liu et al. [16] compared healthcare professionals to CNN performance in a spectrum of clinical domains, including breast cancer, skin cancer, and hepatology and found a summary sensitivity of 87 % (95 % CI: 83 % to 90 %) and a summary specificity of 92.5 % (95 % CI: 85.1 % to 96.4 %). Diagnoses aided by CNNs in detection and diagnosis can help physicians identify and classify radiographs accurately, improve or maintain patient outcomes, and reduce interpretation times.

Evidence of AI algorithms detecting WFs is still limited, and should be investigated further for a deeper understanding. Therefore, this systematic review and meta-analysis aimed to investigate the evidence of WF detection by CNNs on a per-patient level, considering issues of reporting and study design. No previous meta-analysis has assessed if CNNs can reliably diagnose WF on radiographs with healthcare experts consensus as the reference standard.

2. Methods/materials

The meta-analysis was registered at PROSPERO ID: CRD42023431398. Embase, Medline, PubMed, Scopus and Web of Science were searched in the period from 1 Jan 2012 to 9 Mar 2023 by V.H. and S.L. Search strings are documented (Suppl 1), and the Preferred Reporting Items for Systematic reviews and Meta-analysis (PRISMA) were followed [17]. The Rayyan QCRI online platform was used for article screening [18]. Duplicates were removed, and non-relevant studies excluded during screening. Full-text assessments were performed independently by V.H. and S.L. according to pre-defined inclusion criteria described below. Any conflicts were resolved by consensus.

We included studies of patients over the age of 18 undergoing radiographs for the detection of WFs, using CNN algorithms and with healthcare experts' annotation as the reference standard (cf. below). The inclusion of articles was limited to articles published after 1 Jan 2012, based on the recognized change in the development of DL performance in the ImageNet classification challenge [9].

Acceptable reference standards were: Manual, semi-automated or automated image labelling extracted from reports or electronic health records using natural language processing or recurrent neural networks. Also acceptable was labelling by independent readers when the number of human annotators and their qualifications were specified, including a detailed description of annotation flow process [19,20].

The following article types were excluded: Conference papers, editorials, commentaries, reviews, guidelines, book chapters, technical papers, papers in other languages than English and Danish, and papers with insufficient reference standards.

Risk of bias and applicability were assessed independently by M.W. K., S.L. and J.J. using a modified Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool [21] (Suppl 2) and any conflicts resolved by consensus. Diagnostic measures on a per-patient basis for WF on radiographs were included in the meta-analysis. Data were calculated from available information, or authors were contacted when diagnostic measure extraction was impossible. In case of several analyses within the same study, one set of data was chosen per study conservatively (e.g. in favor of more challenging external test set than internal test set).

We applied a bivariate random-effects model for meta-analysis of diagnostic test accuracy data and derived forest plots and a summary receiver operating characteristics (SROC) curve to compare CNN to the reference standards. The SROC curve provides a simultaneous estimate of summary sensitivity and specificity of the included studies with 95 % confidence intervals (95 % CI) [22]. Further, we assessed the heterogeneity between the studies with bivariate I-squared statistics. Significance level was 5 %. We conducted all statistical analyses with the package *metadta* in STATA/MP 17.0 (StataCorp, College Station, Texas 77,845 USA). Validation was performed with *metandi* in Stata. The hierarchical summary receiver operating characteristic model (HSROC) derived with *metandi* and the SROC derived from our bivariate random-effects meta-analysis model with *metadata* are equivalent when when no covariates are included [22] as was the case here.

3. Results

A total of 15,735 records were initially screened, as illustrated in Fig. 1, leading to the inclusion of ten studies with a cumulative count of 185,221 radiographs used for algorithm training [23–32]. These studies are comprehensively detailed in Table 1, with additional data presented in Suppl 3 and Suppl 4 for the PRISMA checklist and data extraction chart, respectively.

The 10 included studies demonstrate a broad spectrum in dataset sizes, from several hundred to over a hundred thousand images, covering a variety of radiograph types like PA, LAT, AP, and frontal views. This diversity enables an extensive evaluation of CNN models under varied conditions, showcasing the versatility of CNNs through the use of diverse architectures and training methods. Employing both pretrained models (such as VGG16, Inception v3, DenseNet, ResNet, EfficientNet) and different training strategies (including full image and ROI-based training), they underscore the adaptability and effectiveness of CNNs in medical image analysis. Excluding Tobler et al (2021), Lindsey et al (2018), Kim et al (2018), and Ürethen et al (2022) reduces the range of dataset sizes and types, yet the remaining studies maintain significant diversity in dataset composition and fracture type representation. Despite the diminished variety in CNN architectures and training strategies due to their omission, the CNN models and methodologies in the remaining research continues to illustrate the diversity and innovation in applying CNNs for WF detection.

The 6 included studies for quantitative synthesis, varied between using radiology residents, orthopaedic surgeons, and radiologists with varying levels of experience as reference standards. For instance, Blüthgen et al. [23] and Kim et al. [31] employed radiology residents and a registrar, respectively, while Thian et al. [26] and Raisuddin et al. [27] utilized annotations by experienced radiologists.

A significant limitation of the study not included in the meta-analysis by Üreten et al. [32] is the inadequate characterization of the reference standard, as it only mentions the re-evaluation of hand fracture radiographs by an orthopedist and a radiologist, each with over five years of experience, without providing detailed criteria for their assessment proficiency or methodological consistency.

For the index test, different CNN models was employed across the studies. For example, Blüthgen et al. [23] investigated two different transfer CNNs, while Oka et al. [29] used a modified VGG16 CNN, and Suzuki et al. [30] applied a CNN with EfficientNet for fine-tuning.

Several studies incorporated external datasets to validate their models. For instance, Blüthgen et al. [23] used the MURA dataset for external testing, and the external test sets varied in size, from the 200 x-rays in Blüthgen et al. [23] to 9090 wrist radiographs in the study by Kim et al. [31]. The studies predominantly focused on wrist radiographs, particularly distal radius and ulnae fractures. Dataset compositions varied, with Kim et al. [31] annotating 9,984 wrist radiographs and Blüthgen et al. [23] retrospectively including 824 wrist radiographs.

Risk of bias is presented in Table 2 and Fig. 2. Four studies were excluded from meta-analysis because of insufficient information on diagnostic measures and a high risk of bias, respectively [24,25,28,32].

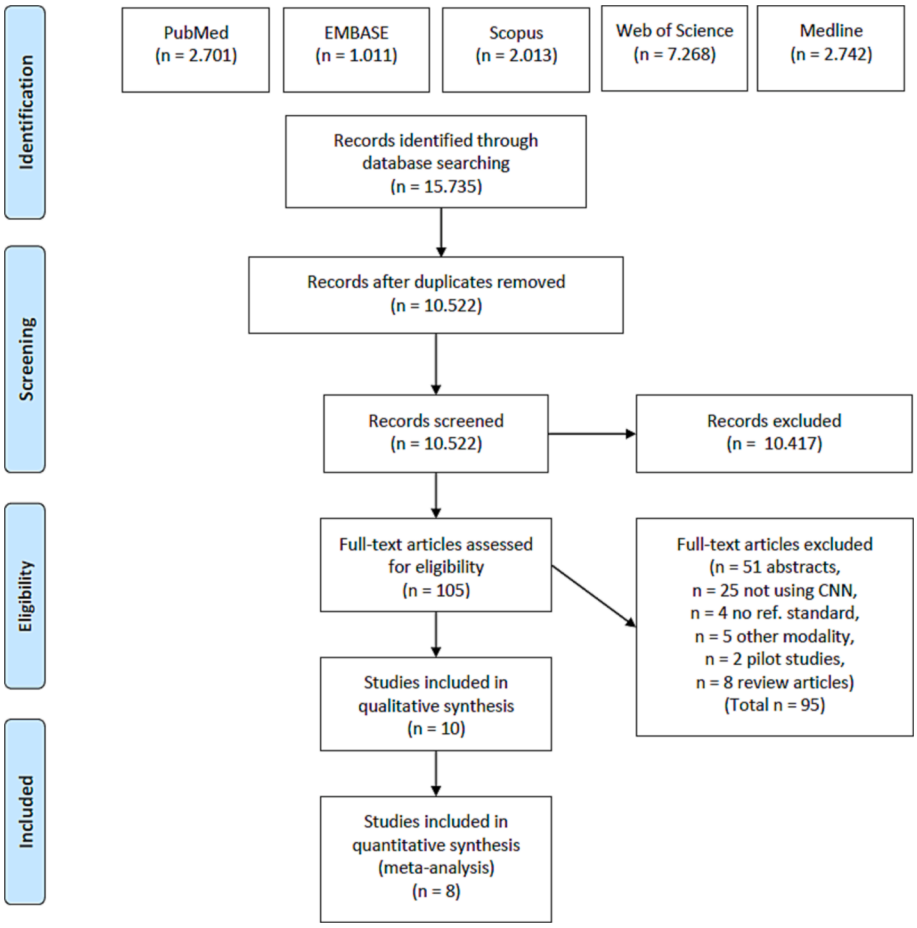

Fig. 1. Flowchart.

Figs. 3 and 4 present summary sensitivity and specificity by means of forest plots and an SROC-curve. Comparing the CNN performance to reference standard found a summary sensitivity of 92 % (95 % CI: 80 % to 97 %) and a summary specificity of 93 % (95 % CI: 76 % to 98 %). It should be noted that the bivariate I-squared analysis showed considerable heterogeneity between the studies (81.90 %). The I-squared statistics for sensitivity and specificity were 80.94 and 85.48, respectively. Source data and results from the statistical analysis with Stata as well as validation can be found in Suppl 5 and Suppl 6.

4. Discussion

4.1. Statement of principal findings

The total amount of test data used in the meta-analysis of the six included studies was 33,026 plain wrist radiographs with a summary sensitivity of 92 % (95 % CI: 80 % to 97 %) and a summary specificity of 93 % (95 % CI: 76 % to 98 %), in detecting WF with CNNs on radiographs.

4.2. Strengths and weaknesses

To our knowledge, our study is the first meta-analysis to quantify the performance of CNNs in detecting WF. The stringent choice of reference standards for inclusion strengthens the validity of our findings, compared to, e.g. Kuo et al., who did not define any a priori in-/exclusion criteria in this domain [33].

Despite the comprehensive search strategy employed, relevant studies may have been excluded, limiting the systematic review and meta-analysis 's comprehensiveness. Our reliance on only published data may have introduced publication bias. The guidelines (CLAIM) used for non-randomized studies, though specifically designed for AI studies, is novel and not yet well-established. The bivariate I-squared analysis results suggested high heterogeneity among studies, implying possible unsuitability for data pooling, and the varying experience levels of the authors may have introduced a level of subjective bias into the assessment. Another limitation is the potential variability and error margins inherent in human readers' interpretations used as the reference standard, which may affect the accuracy of the results.

4.3. Strengths and weaknesses in relation to other studies

The six studies included in the quantitative synthesis employed various DL algorithms for image recognition, classification, and object detection tasks. These algorithms can be grouped into four categories: Transfer learning CNNs [23,24,29], deep CNNs [25,28,30], region-based CNNs [26], and modified CNNs [27].

Both studies by Blüthgen et al. [23] and Oka et al [29] utilized transfer learning, with dataset sizes of 1,626, and 1,474, respectively. However, Blüthgen et al. [23] included 100 radiographs in their test set and 200 radiographs in the external test set, with a fracture distribution of 42 % fractures and 50 % fractures, respectively, and Oka et al [29] creating two additional datasets of 120 and 50 wrist radiographs with a split ratio of 80 %/20 % and 40 %/60 % fractures/without fractures, respectively. Using small test sets for model evaluation may lead to issues related to representation, diversity, and balance. Limited sample sizes may not adequately represent the real-world fracture distribution, which could impact the reliability of the model's performance and its generalizability to unseen data. Moreover, the fracture distribution in the test sets may not mirror real-world rates, which could lead to model

Table 1
Study characteristics of studies included **for meta-**analysis.

Author, year	Study design	Dataset size in total	Target condition	Reference standard	CNN model type	Training set size	Validation set size	External test set size	Sensitivity (95 % CI)	Specificity (95 % CI)	AUC (95 % CI)
Raisuddin et al (2021)	R (Test set #1)	2.258 wrist studies (4.497 PA and LAT)	Distal radius fractures	CT imaging – existing radiology report was manually labeled as normal or fracture by a medical student who received basic training in diagnostic radiology.	Transfer learning CNN in ROI cropped image	1.946 wrist studies (3.873 PA and LAT)	NA	207			
	PA								0.97 (0.94–1.00)	0.88 (0.80–0.94)	0.98 (0.97–0.99)
	LAT								0.97 (0.94–1.00)	0.91 (0.84–0.96)	0.98 (0.97–0.99)
	Ensemble								0.97 (0.94–1.00)	0.87 (0.79–0.93)	0.99 (0.98–0.99)
	R (Test set #2)							105			
	PA								0.50 (0.30–0.70)	0.89 (0.82–0.95)	0.81 (0.69–0.91)
	LAT								0.50 (0.30–0.70)	0.94 (0.88–0.98)	0.83 (0.70–0.93)
	Ensemble								0.60 (0.40–0.80)	0.92 (0.87–0.97)	0.84 (0.72–0.93)
Tobler et al (2021)	R (Test set #1)	15.775 frontal and lateral radiographs	Distal radius fractures	Two musculoskeletal radiologists (Test set A)	Transfer learning CNN		NA				
	Frontal					4.06		291			
	LAT			Three radiology residents (Test set B)		3.937		291			
	R (Test set #2)										
	Frontal					4.06		163			
	LAT					3.937		163			
Thian et al (2019)	R	7.356 wrist studies (7.295 frontal and 7.319 lateral)	Radius and Ulna fractures	Training set annotated by three radiologists	Object detection transfer CNN	13.052	1.562				
	Frontal			Two radiologists		6.515	780	365	0.98 (0.94–0.99)	0.84 (0.79–0.89)	0.93 (0.90–0.95)
	LAT					6.537	782	365	0.99 (0.95–1.00)	0.86 (0.81–0.91)	0.94 (0.92–0.96)
Oka et al (2021)	P - Dataset #1		Distal radius fracture	Clinical diagnosis results by specialized orthopedic surgeons	Transfer learning CNN (VGG16)						
	AP										

(*continued on next page*)

Table 1 (*continued*)

Author, year	Study design	Dataset size in total	Target condition	Reference standard	CNN model type	Training set size	Validation set size	External test set size	Sensitivity (95 % CI)	Specificity (95 % CI)	AUC (95 % CI)
	LAT	498				390	48	60	0.95 (0.92–0.98)	0.97 (0.95–0.99)	
		485				353	72	60	0.99 (0.97–1.00)	0.97 (0.93–1.00)	
	P - Dataset #2		Ulnar styloid fracture								
	AP										
		491				391	50	50	0.92 (0.87–0.98)	0.90 (0.87–0.94)	
Lindsey et al (2018)	R - Random subset	135.409 radiographs	Wrist fractures	18 senior sub-specialized orthopedic surgeons	CNN for fracture detection and localization	28.341	3.149	3.5			0.97 (0.96–0.97)
	P - Separate dataset							1.4			0.98 (0.97—0.98)
Kim et al (2018)	R -	11.112	Distal radius and ulna fractures	Radiological report confirmed by a radiology registrar with 3 years of radiology experience	Transfer CNN (Inception v3)	1.389 80/10/10					
	R - Test set							139			0.95
Suzuki et al (2021)	R - Dataset	1.633 radiographs	Distal radius fractures	Diagnoses were confirmed by two board certified orthopedic surgeons.	Ensemble model based on EfficientNet B2 and EfficientNet B4	1.333		300	0.99 (0.93–1.00)	1.00 (0.95–1.00)	0.99
Blüthgen et al (2020)	P - Internal test set	824 AP and LAT radiographs	Distal radius fractures	Internal test set: 2 radiology residents with 3 and 5 years of experience with electronic health record and available CT scans	Two CNN models with optimal parameters created in a generic image analysis software	524		100			
	Model 1 AP LAT								0.86 (0.64–0.97)	0.86 (0.68–0.96)	0.93 (0.82–0.98)
	Combined								0.86 (0.64–0.97)	1.00 (0.88–1.00)	0.94 (0.84–0.99)
									0.81 (0.58–0.95)	1.00 (0.88–1.00)	0.95 (0.85–0.99)
	Model 2 AP			External test set: Evaluated by 2 attending radiologists with 16 and 7							

(*continued on next page*)

Table 1 (*continued*)

Author, year	Study design	Dataset size in total	Target condition	Reference standard	CNN model type	Training set size	Validation set size	External test set size	Sensitivity (95 % CI)	Specificity (95 % CI)	AUC (95 % CI)
				years of experience + one 2nd year radiology resident							
	LAT								0.86 (0.64–0.97)	0.97 (0.82–1.00)	0.95 (0.85–0.99)
	Combined								0.90 (0.70–0.99)	0.90 (0.73–0.98)	0.94 (0.83–0.99)
									0.90 (0.70–0.99)	0.97 (0.82–1.00)	0.96 (0.87–1.00)
	P - External test set							200			
	Model 1										
	AP										
	LAT										
	Combined								0.74 (0.60–0.85)	0.64 (0.49–0.77)	0.80 (0.71–0.88)
									0.80 (0.66–0.90)	0.66 (0.51–0.79)	0.83 (0.74–0.90)
	Model 2								0.80 (0.66–0.90)	0.86 (0.73–0.94)	0.87 (0.79–0.93)
	AP										
	LAT										
	Combined								0.64 (0.49–0.77)	0.90 (0.78–0.97)	0.82 (0.73–0.89)
									0.92 (0.81–0.98)	0.60 (0.45–0.74)	0.84 (0.76–0.91)
									0.82 (0.69–0.91)	0.78 (0.64–0.88)	0.89 (0.81–0.94)
Kim et al (2021)	P - Model 1	9.984 wrist radiographs	Distal radius and ulnar fractures	Dual radiological reporting	DenseNet-161	8.994		990	0.90 (0.89–0.92)	0.90 (0.89–0.92)	0.96
	P - Model 2				ResNet-152				0.89 (0.88–0.90)	0.88 (0.87–0.89)	0.95
Ürethen et al (2022)	P - Model 1	545 hand and wrist radiographs	Wrist fractures	Orthopedist and radiologist, both with over 5 years experience	VGG-16	697	123	135	0.97	0.9	
	P - Model 2				ResNet-50				0.95	0.84	
	P - Model 3				GoogLeNet				0.91	0.86	

AUC: Area under the curve.
NA: Not applicable (Anything that is not reported in the study).
CNN: Convolutional Neural Network.
R: Retrospective.
P: Prospective.

Table 2
QUADAS-2 evaluation of risk of bias and applicability.

Study	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Raisuddin et al (2021)	Yes Unclear Yes Low	Yes Yes Yes Low	Yes Yes Yes Low	Yes Yes No Unclear	Low	Low	Low
Tobler et al (2021)	Unclear Yes Yes Unclear	Yes No Yes Low	Yes Yes Yes Low	Yes Yes Unclear Low	Unclear	Low	Low
Thian et al (2019)	Yes Yes Yes Low	Yes Unclear Yes Low	Unclear Yes Unclear Unclear	Yes Yes Yes Low	Low	Low	Unclear
Oka et al (2021)	Unclear Yes Unclear Unclear	Yes Unclear Yes Low	Unclear No No High	Yes No Unclear Unclear	Unclear	Low	High
Lindsey et al (2018)	Yes Yes Yes Low	Yes Yes Yes Low	Unclear No No High	Yes Unclear Yes Unclear	Low	Low	High
Kim et al (2018)	Yes Yes Yes Low	Yes Yes Yes Low	Yes Yes No Unclear	Yes Yes Unclear Unclear	Low	Low	Unclear
Blüthgen et al (2020)	Yes Yes Yes Low	Yes No Yes Unclear	Yes Yes Unclear Unclear	Yes Yes Yes Low	Low	Unclear	Unclear
Kim et al (2021)	Yes No Yes Unclear	Yes Yes Yes Low	Yes Yes No Unclear	Yes Yes No Unclear	Unclear	Low	Unclear
Suzuki (2022)	Yes Yes Yes Low	Yes Yes Yes Low	Unclear Unclear Unclear High	Yes UnclearYes Unclear	Low	Low	High
Üreten et al (2022)	Yes Yes Unclear Unclear	Yes Yes Yes Low	Unclear Unclear Unclear High	Unclear No Yes Unclear	Unclear	Low	High

Low, high and unclear risk.

bias towards the more prevalent class and affect the model's performance on datasets with different fracture distributions [19,34].

The studies by Blüthgen et al. [23] and Suzuki et al. [30] investigated deep CNNs to analyze large datasets. The large dataset sizes, ranging from 824 to 1633 radiographs, provide the models with a decent amount of data to learn from. However, these larger datasets require more computational resources and longer training times. In terms of limitations, while deep CNNs can deliver impressive results, they require large amounts of labeled data to train effectively. Labeled medical images can be difficult and time-consuming to produce, as they require the expertise of trained clinicians [20]. The method by Suzuki et al. [30] is comprehensive, involving diagnosis confirmation by two board-certified

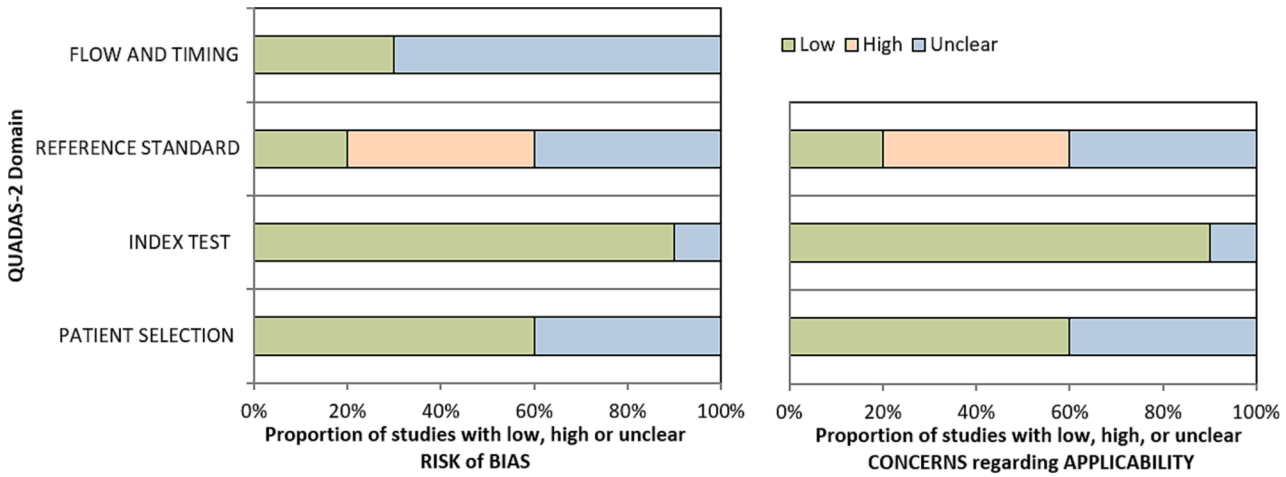

Fig. 2. QUADAS-2 overview.

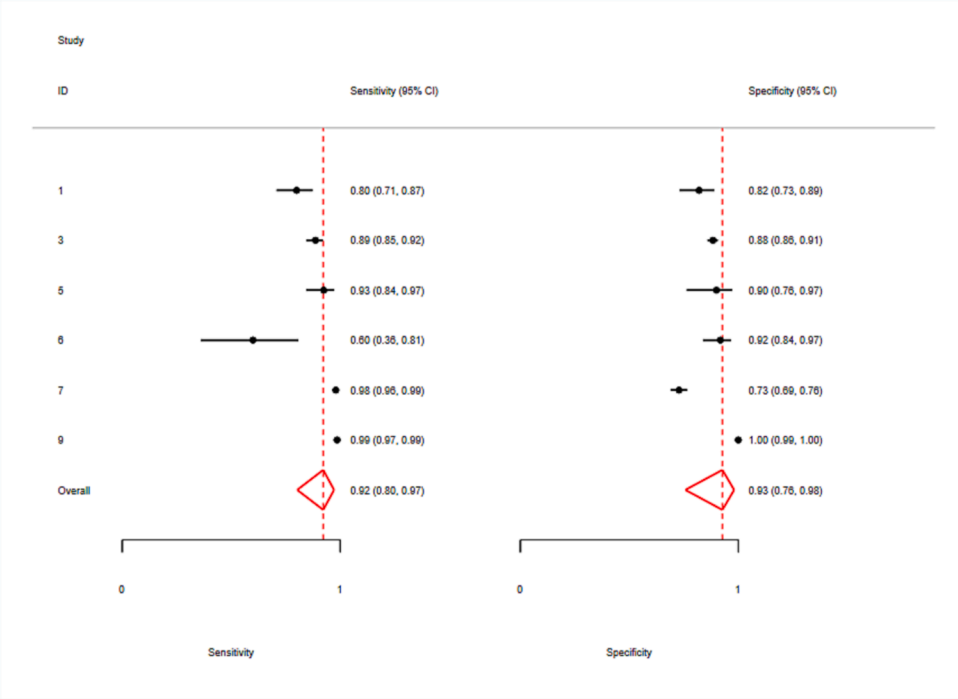

Fig. 3. Forest plot for sensitivity and specificity.

orthopedic surgeons, and uses multiple modalities - radiographs and computed tomography - along with clinical information. From the qualitative synthesis the method from Lindsey et al. [25] comes out on top, due to its rigorous approach involving subspecialized orthopedic surgeons using an annotation software tool. Tobler et al. [28] presents a more automated methods, relying on key phrases to label and classify fractures. While this could be highly efficient, its accuracy is contingent on the quality of the reports. Finally, Kim et al. [24] involves classifying images into fracture and non-fracture groups based on dual radiological reporting, is considered the least detailed. It may not provide the same level of detail or accuracy as the other methods, despite its potential efficiency.

An increasing number of studies are investigating the performance of AI compared to healthcare professionals. In our quantitative synthesis, we found two studies comparing AI performance with expert clinicians [23,27]. Four studies compared their trained DL algorithm with clinicians showing comparable performance of the DL algorithms when the clinicians had varying degrees of experience. However, Blüthgen et al. [23] found that the less experienced radiologist (2 years) was less sensitive and specific than the DL models. They found their algorithm comparable to the orthopedists and outperformed the radiologists. Meanwhile, Raisuddin et al. [27] compared their algorithm to two radiologists and two primary care physicians. They found high disagreement between the primary care physicians and high agreement between the two radiologists on their test set with trivial cases. However, all raters had a low agreement on the test set labelled as complex cases, with CT as the ground truth. The performance of their algorithm was lower than radiologists and higher than the primary care physicians on the test set with trivial cases but higher than all raters on the test set with complicated cases, highlighting the clinical value of a fracture algorithm in cases with subtle findings that may be initially overlooked by healthcare professionals. In the study from the qualitative synthesis by Tobler et al. [28], the AI performance was compared to radiology residents showing similar abilities in detecting fractures and classifying

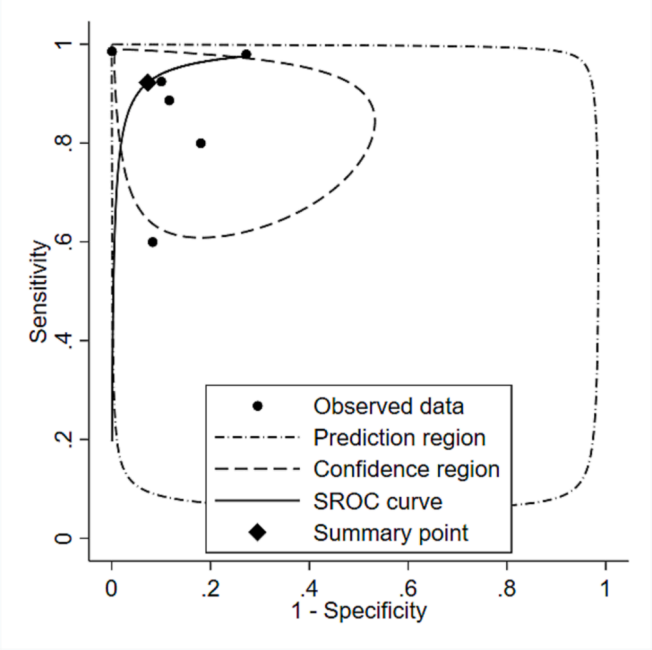

Fig. 4. Summary receiver operating charcateristics curve.

multiple fragments. Still, it differed significantly in classifying fragment displacement and joint involvement, with worse AI performance. One similar study found comparable diagnostic accuracy between AI and clinicians with a summary sensitivity of 91 % (95 % CI: 84 % to 85 %) and specificity of 91 % (95 % CI: 81 % to 95 %), including studies focusing on both upper and lower limb fractures [33]. In the qualitative synthesis in the study by Lindsey et al. [25] investigated fracture detection of 16 physician assistants and 24 medical doctors with and without the assistance of AI and found their model to improve the sensitivity and specificity for both groups. A similar study found improved clinical performance with aided AI assistance, improving the sensitivity and specificity of physicians by 8.7 % (95 % CI: 3.1 % to 14.2 %; P = 0.003) and 4.1 % (95 % CI: 0.5 % to 7.7 %; P < 0.001), respectively [33,35]. Another prospective study reviewed a commercially available AI algorithm and found increased sensitivity for two humans with AI assistance from 84.74 % (95 % CI: 84.34 % to 85.14 %) to 91.28 % (95 % CI: 91.25 % to 91.31 %), with almost similar specificity of humans without assistance of 97.11 % (95 % CI: 97.10 % to 97.12 %) and with assistance of 97.36 % (95 % CI: 97.35 % to 97.37 %) [36]. These results suggest that the AI algorithm used can enhance human sensitivity in this particular context, without significantly affecting specificity.

4.4. Implications for practice and future research

The potential benefits of AI in healthcare must be balanced against the challenges and potential risks. These include ethical issues related to data privacy and responsibility for decision-making, as well as practical considerations such as cost and the need for ongoing training and algorithm validation.

An important area for further research is the evaluation of CNN performance in the presence of multiple findings. Flagging an area suspect for fracture will, in our analysis, count as a positive finding. Hovewer more discrete findings, such as ligamental injuries or subluxation of carpal bones may present concurrent with fractures, some of which require treatment. By marking a fracture, there is a real risk that secondary findings may be overlooked due to satisfaction of search errors, especially by less experienced clinicians [6]. At present such algorithms are separate entities [37], but will need to be integrated with WF technologies for comprehensive evaluation. Thus the ultimate metric should be the ability of CNNs to allocate patients for correct treatment.

In light of the study's findings that AI may compensate for less experienced clinicians, research could focus on how AI can be used to augment the skills of newer or less specialized practitioners. Given that AI has shown promise in improving or maintaining diagnostic accuracy, further research could focus on how these improvements translate into improved patient outcomes. This could include studying the impact of AI on patient care in terms of reduced misdiagnosis, shorter hospital stays, or improved treatment plans.

5. Conclusion

Our meta-analysis found high performance of CNN algorithms detecting WF on plain radiographs, but the conlusion is limited by the small number of available studies. Studies with external dataset testing and evaluation with uniformity of methods and robust reference standard by independent experts in unselected patient cohorts are needed. For clinicians, AI could potentially be used to enhance diagnostic confidence, especially in fields of radiology. AI algorithms may be particularly useful for less experienced clinicians.

Key points

- Convolutional neural network algorithms has high diagnostic performance in finding WF in plain wrist radiographs.
- There is a dire need for studies with a robust reference standard, external data-set validation with investigation of diagnostic performance.
- Future studies on convolutional neural network algorithms should be evaluated with patient outcomes as the reference.

CRediT authorship contribution statement

V. Hansen: Writing – review & editing, Methodology, Investigation, Conceptualization. **J. Jensen:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **M.W. Kusk:** Writing – review & editing, Validation, Methodology, Supervision. **O. Gerke:** Writing – review & editing, Validation, Supervision, Resources, Methodology. **H.B. Tromborg:** Writing – review & editing, Supervision. **S. Lysdahlgaard:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2024.111399.

References

[1] H.R. Guly, Diagnostic errors in an accident and emergency department, Emerg. Med. J. 18 (2001) 263–269.
[2] L. Berlin, Defending the "Missed" radiographic diagnosis, Am. J. Roentgenol. 176 (2001) 317–322.
[3] J.J. Donald, S.A. Barnard, Common patterns in 558 diagnostic radiology errors, J. Med. Imaging Radiat. Oncol. 56 (2012) 173–178.
[4] J.S. Whang, S.R. Baker, R. Patel, L. Luk, A. Castro, The causes of medical malpractice suits against radiologists in the United States, Radiology 266 (2013) 548–554.

[5] P. Hallas, T. Ellingsen, Errors in fracture diagnoses in the emergency department – characteristics of patients and diurnal variation, BMC Emerg. Med. 6 (2006) 4.
[6] K. Hames, M.N. Patlas V.M. Mellnick, D.S. Katz, Errors in Emergency and Trauma Radiology: General Principles. In: Patlas MN, Katz DS, Scaglione M, editors. Errors in Emergency and Trauma Radiology. Springer International Publishing, Cham, 2019, Available via https://doi.org/10.1007/978-3-030-05548-6_1 (accessed 12 Sep 2023).
[7] R. Kaewlai, L.L. Avery, A.V. Asrani, H.H. Abujudeh, R. Sacknoff, R.A. Novelline, Multidetector CT of carpal injuries: anatomy, fractures, and fracture-dislocations, Radiographics 28 (2008) 1771–1784, https://doi.org/10.1148/rg.286085511.
[8] M. Geijer, G.Y. El-Khoury, MDCT in the evaluation of skeletal trauma: principles, protocols, and clinical applications, Emerg. Radiol. 13 (2006) 7–18, https://doi.org/10.1007/s10140-006-0509-5.
[9] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inform. Processing Syst. (2012) 1097–1105.
[10] W. Gale, L. Oakden-Rayner, G. Carneiro, A.P. Bradley, L.J. Palmer, Detecting hip fractures with radiologist-level performance using deep neural networks, 2017, Available via http://arxiv.org/abs/1711.06504 (accessed 8 Mar 2021).
[11] M. Adams, W. Chen, D. Holcdorf, M.W. McCusker, P.D. Howe, F. Gaillard, Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures, J. Med. Imaging Radiat. Oncol. 63 (2019) 27–32.
[12] M.A. Badgeley, J.R. Zech, L. Oakden-Rayner, B.S. Glicksberg, M. Liu, W. Gale, et al., Deep learning predicts hip fracture using confounding patient and healthcare variables, Npj Digit Med 2 (2019) 1–10.
[13] U. Raghavendra, N.S. Bhat, A. Gudigar, U.R. Acharya, Automated system for the detection of thoracolumbar fractures using a CNN architecture, Future Gener Comput Syst 85 (2018) 184–189.
[14] S.W. Chung, S.S. Han, J.W. Lee, et al., Automated detection and classification of the proximal humerus fracture by using deep learning algorithm, Acta Orthop. 89 (2018) 468–473.
[15] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W. M. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, https://doi.org/10.1016/j.media.2017.07.005.
[16] X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J.R. Ledsam, M.K. Schmid, K. Balaskas, E. J. Topol, L.M. Bachmann, P.A. Keane, A.K. Denniston, A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digit Health 1 (2019) e271–e297, https://doi.org/10.1016/S2589-7500(19)30123-2.
[17] J. Salameh, P.M. Bossuyt, T.A. McGrath, B.D. Thombs, C.J. Hyde, P. Macaskill, et al., Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist, BMJ 370 (2020) m2632.
[18] M. Ouzzani, H. Hammady, Z. Fedorowicz, A. Elmagarmid, Rayyan-a web and mobile app for systematic reviews, Syst. Rev. 5 (2016) 210.
[19] J. Mongan, L. Moy, C.E. Kahn, Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers, Radiol. Artif. Intell. (2020).
[20] M.J. Willemink, W.A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, et al., Preparing medical imaging data for machine learning, Radiology 295 (2020) 4–15.
[21] P.F. Whiting, A.W.S.S. Rutjes, M.E. Westwood, S. Mallett, J.J. Deeks, J.B. Reitsma, et al., Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies, Ann. Int. Med. 529 (2011).
[22] V.N. Nyaga, M. Arbyn, Metadta: a Stata command for meta-analysis and meta-regression of diagnostic test accuracy data - a tutorial, Arch. Public Health 80 (1) (2022) 95, https://doi.org/10.1186/s13690-021-00747-5.
[23] C. Blüthgen, A.S. Becker, I. Vittoria de Martini, A. Meier, K. Martini, T. Frauenfelder, Detection and localization of distal radius fractures: deep learning system versus radiologists, Eur. J. Radiol. (2020).
[24] D.H. Kim, D.H. Kim, T. MacKinnon, D.H. Kim, Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks, Clin. Radiol. 73 (2018) 439–445.
[25] R. Lindsey, A. Daluiski, S. Chopra, et al., Deep neural network improves fracture detection by clinicians, PNAS 115 (2018) 11591–11596, https://doi.org/10.1073/pnas.1806905115.
[26] Y.L. Thian, Y. Li, P. Jagmohan, et al., Convolutional neural networks for automated fracture detection and localization on wrist radiographs, Radiol. Artif. Intell. 1 (1) (2019) e180001.
[27] A.M. Raisuddin, E. Vaattovaara, M. Nevalainen, et al., Critical evaluation of deep neural networks for wrist fracture detection, Sci. Rep. 11 (2021) 6006.
[28] P. Tobler, J. Cyriac, B.K. Kovacs, et al., AI-based detection and classification of distal radius fractures using low-effort data labeling: evaluation of applicability and effect of training set size, Eur. Radiol. 31 (9) (2021) 6816–6824.
[29] K. Oka, R. Shiode, Y. Yoshii, et al., Artificial intelligence to diagnosis distal radius fracture using biplane plain X-rays, J. Orthop. Surg. 16 (1) (2021) 694.
[30] T. Suzuki, S. Maki, T. Yamazaki, et al., Detecting distal radial fractures from wrist radiographs using a deep convolutional neural network with an accuracy comparable to hand orthopedic surgeons, J. Digit. Imaging 35 (1) (2022) 39–46.
[31] M.W. Kim, J. Jung, S.J. Park, et al., Application of convolutional neural networks for distal radio-ulnar fracture detection on plain radiographs in the emergency room, Clin Exp Emerg Med 8 (2) (2021) 120–127.
[32] K. Üreten, H.F. Sevinç, U. İğdeli, et al., Use of deep learning methods for hand fracture detection from plain hand radiographs, Ulus Travma Ve Acil Cerrahi Derg Turk J. Trauma Emerg. Surg. TJTES 28 (2) (2022) 196–201.
[33] R.Y.L. Kuo, C. Harrison, T.A. Curran, et al., Artificial intelligence in fracture detection: a systematic review and meta-analysis, Radiology 304 (1) (2022) 50–62.
[34] A.C. Yu, J. Eng, One algorithm may not fit all: how selection bias affects machine learning performance, Radiographics 40 (7) (2020) 1932–1937.
[35] L. Duron, A. Ducarouge, A. Gillibert, et al., Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study, Radiology 300 (1) (2021) 120–129.
[36] J. Oppenheimer, S. Lüken, B. Hamm, et al., A prospective approach to integration of AI fracture detection software in radiographs into clinical workflow, Life 13 (1) (2023) 223.
[37] B. Pridgen, L. von Rabenau, A. Luan, et al., Automatic detection of perilunate and lunate dislocations on wrist radiographs using deep learning, Plast Reconstr. Surg. Epub (2023 Jul 15), https://doi.org/10.1097/PRS.0000000000010928.