Pediatric Radiology (2023) 53:1125–1134
https://doi.org/10.1007/s00247-023-05588-8

ORIGINAL ARTICLE

Detecting pediatric wrist fractures using deep-learning-based object detection

John R. Zech[1] · Giuseppe Carotenuto[2] · Zenas Igbinoba[1] · Clement Vinh Tran[1] · Elena Insley[3] · Alyssa Baccarella[4] · Tony T. Wong[1]

Received: 4 October 2022 / Revised: 9 December 2022 / Accepted: 30 December 2022 / Published online: 18 January 2023

Abstract

Background Missed fractures are the leading cause of diagnostic error in the emergency department, and fractures of pediatric bones, particularly subtle wrist fractures, can be misidentified because of their varying characteristics and responses to injury.

Objective This study evaluated the utility of an object detection deep learning framework for classifying pediatric wrist fractures as positive or negative for fracture, including subtle buckle fractures of the distal radius, and evaluated the performance of this algorithm as augmentation to trainee radiograph interpretation.

Materials and methods We obtained 395 posteroanterior wrist radiographs from unique pediatric patients (65% positive for fracture, 30% positive for distal radial buckle fracture) and divided them into train ($n = 229$), tune ($n = 41$) and test ($n = 125$) sets. We trained a Faster R-CNN (region-based convolutional neural network) deep learning object-detection model. Two pediatric and two radiology residents evaluated radiographs initially without the artificial intelligence (AI) assistance, and then subsequently with access to the bounding box generated by the Faster R-CNN model.

Results The Faster R-CNN model demonstrated an area under the curve (AUC) of 0.92 (95% confidence interval [CI] 0.87–0.97), accuracy of 88% ($n = 110/125$; 95% CI 81–93%), sensitivity of 88% ($n = 70/80$; 95% CI 78–94%) and specificity of 89% ($n = 40/45$, 95% CI 76–96%) in identifying any fracture and identified 90% of buckle fractures ($n = 35/39$, 95% CI 76–97%). Access to Faster R-CNN model predictions significantly improved average resident accuracy from 80 to 93% in detecting any fracture ($P < 0.001$) and from 69 to 92% in detecting buckle fracture ($P < 0.001$). After accessing AI predictions, residents significantly outperformed AI in cases of disagreement (73% resident correct vs. 27% AI, $P = 0.002$).

Conclusion An object-detection-based deep learning approach trained with only a few hundred examples identified radiographs containing pediatric wrist fractures with high accuracy. Access to model predictions significantly improved resident accuracy in diagnosing these fractures.

Keywords Artificial intelligence · Bone · Buckle fracture · Children · Convolutional neural network · Deep learning · Radiography · Wrist

✉ John R. Zech
jrz2111@columbia.edu

[1] Department of Radiology, Columbia University Irving Medical Center/New York Presbyterian Hospital, 622 W 168th St., New York, NY 10032, USA

[2] Department of Radiology, University of California San Diego, San Diego, CA, USA

[3] Department of Pediatrics, Columbia University Irving Medical Center/Morgan Stanley Children's Hospital of New York, New York, NY, USA

[4] Division of Gastroenterology, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Introduction

Missed fractures are the leading cause of diagnostic error in the emergency department (ED), and prior work has estimated these account for 80% of all diagnostic errors in the ED [1, 2]. Fractures of pediatric bones can be misidentified because of their varying characteristics and responses to injury [3]. Pediatric wrist fractures, in particular subtle buckle fractures, often go unrecognized [3–5]. Deep learning has demonstrated strong performance in identifying fractures for both adults [6, 7] and children [8]. While deep learning has been shown to perform strongly in identifying fractures on adult wrist radiographs [9–13], few studies have specifically evaluated its performance on pediatric wrist radiographs.

Deep learning object detection models are trained to identify the specific part of an image containing a given finding. While originally used to identify objects such as bicycles and cars in general image datasets, this approach has increasingly been used to identify pathology in radiologic imaging, and has demonstrated strong utility for detecting fractures in adults [6, 14, 15]. We investigated how an object detection approach would perform in identifying pediatric wrist fractures and evaluated whether access to its predictions could improve physicians' ability to detect these fractures.

Materials and methods

Data collection

This retrospective study was approved by the institutional review board, which waived the requirement for informed consent, and the study complied with the Health Insurance Portability and Accountability Act.

Our sample consisted of 395 posteroanterior (PA) radiographs from 395 children younger than 18 years who had wrist radiographs performed between Jan. 9, 2015, and Nov. 15, 2019 (Fig. 1). We chose this sample size because prior work demonstrated that effective deep learning models had been developed for fracture detection in other contexts with datasets of similar size (e.g., scaphoid fracture, $n = 300$ radiographs [16]; wrist fracture in a primarily adult population, $n = 542$ radiographs [17]). Mean age was 10.1 years, minimum 0.8, maximum 17.8; interquartile range was 4.8 years (25th–75th percentile age 7.9–12.7 years). In our sample, 37% of the radiographs were from female patients ($n = 148/395$) and 49% were of the right wrist ($n = 192/395$). Of these radiographs, 65% were positive for any fracture ($n = 256/395$) and 30% were positive for buckle fracture ($n = 118/395$), an incomplete fracture distinct to pediatric patients characterized by cortical bulging rather than cortical break [18].

Radiographs were rescaled (average height 1,177 pixels with a standard deviation [SD] of 241 pixels, average width

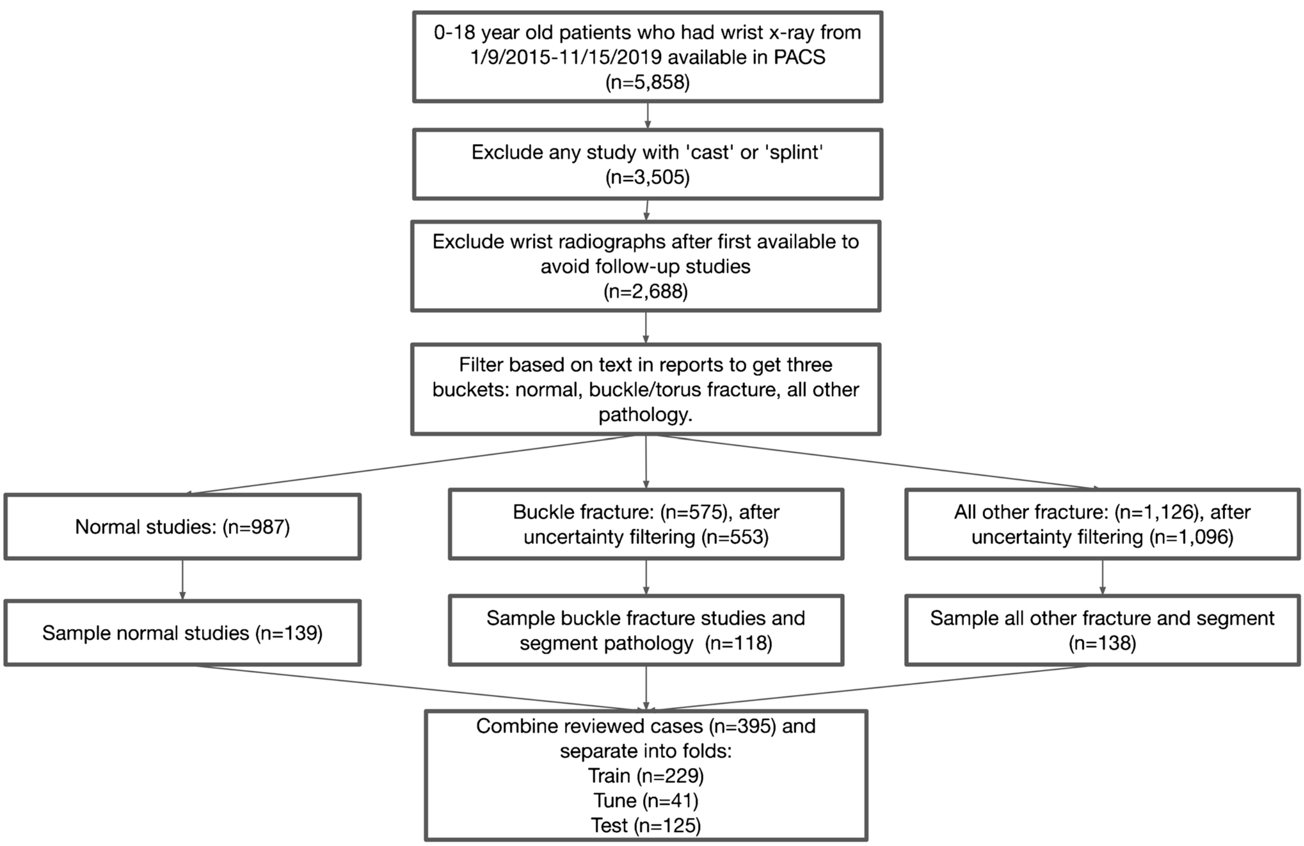

Fig. 1 Inclusion criteria. "Normal" studies are those whose impressions contained "no acute fracture," "no evidence of fracture" or "unremarkable radiographic examination." "Buckle fracture" studies contained the words "buckle" or "torus." "All other pathology" consisted of studies not in these two groups. "Uncertainty filtering" consisted of excluding studies containing the words "possible" or "uncertain" that were not clearly designated as normal. *PACS* picture archiving and communication system

880 pixels with SD 222 pixels). Only examinations that served as an initial evaluation for fracture were used. Any follow-up radiographs after initial fracture diagnosis were excluded from consideration. Any radiographs in which the child had been casted or splinted were excluded. Figure 1 provides a flowchart describing inclusion criteria. Online Supplementary Material 1 provides details of the fractures included in train and test data. All examinations categorized as positive for fracture were confirmed to have a fracture demonstrated on PA projection.

Data processing

Radiographs were randomly divided into train ($n=229$), tune ($n=41$) and test ($n=125$) examinations. There was no patient overlap between groups. We chose partition sizes to preserve a sufficient number of test cases to draw meaningful inferences about model performance, while dividing the remaining data into approximately 85% train and 15% tune, a train/tune split similar to that in other work [19]. We manually reviewed images to confirm they contained no identifiable information.

The original radiology report served as ground truth and we coded it based upon whether the interpreting radiologist had (1) identified any fracture within the wrist radiograph and (2) identified a buckle fracture of the distal radius. The report created by the pediatric fellowship-trained attending radiologist at the time of clinical interpretation was used to establish ground truth given that it reflected the standard of care at our institution. Informed by this radiology report, bounding boxes containing imaging findings indicative of fracture were manually segmented by a postgraduate year (PGY)-4 radiology resident. This resident did not participate in subsequent physician image review. Any questions that arose regarding identification of the fracture on imaging was reviewed by the senior author (13 years radiology experience).

Images were reviewed and bounding boxes annotated in a custom JupyterLab Notebook [20]. Boxes were drawn as tightly as possible to encompass a given imaging finding, and multiple boxes could be drawn on the same image to annotate different findings. When images were rotated, boxes were annotated on the original rotated radiograph.

Model training

We used a Faster R-CNN (region-based convolutional neural network) pretrained to the benchmark MS COCO (Microsoft Common Objects in Context) object detection dataset [21]. This was fine-tuned on train data in PyTorch 1.7.0 using the freely available Detectron2 library contributed by Facebook Artificial Intelligence (AI) Research (batch size 10, learning rate 0.001) [21–23]. Test data ($n=125$) were not used during the model training process and were reserved for final evaluation of the trained model. The model with lowest tune loss was retained for further analysis. Image preprocessing consisted of resizing images to maximum length of 833 pixels while preserving aspect ratio and randomly flipping images horizontally with probability of 0.5.

Faster R-CNN is a convolutional neural network–based model that can be trained to predict bounding boxes for specific objects in images, facilitating both identification and localization [21]. We chose this model because of its established use in medical imaging AI research and strong support within the freely available and widely used Detectron2 object detection and segmentation library (Detectron2 Faster R-CNN R50-FPN) [22]. Given that object detection models employ more complicated architectures than traditional classification-oriented convolutional neural networks, we think it is important to use thoroughly tested libraries when training such models; we note that Detectron2 has been used to train a leading commercial AI product for adult fracture detection [15, 22, 24]. Detectron2 can be downloaded within a Docker image to allow for seamless deployment [25]. As shorthand, we refer to the Faster R-CNN model trained in this fashion as the AI algorithm in this paper.

Model evaluation and resident comparison

The model predicted the absence or presence of a fracture on each test radiograph. We chose a classification threshold setting of 80% empirically to consider a region positive for pathology.

Each PA test radiograph ($n=125$) was blindly and independently reviewed by a PGY-2 and PGY-4 pediatrics resident/fellow as well as a PGY-2 and PGY-4 radiology resident. These trainee physicians were provided with a blank spreadsheet and asked to briefly describe any relevant pathology they identified in the PA radiographs without the assistance of AI. They then performed a second review of these images after regions suspicious for fractures had been highlighted with a bounding box proposed by the AI algorithm. These were submitted 3–12 weeks after initial review, and the resident was not provided with access to the original interpretations during re-interpretation. In the initial review, each resident was presented with the unannotated radiograph at full acquired resolution. In the second review, each resident was presented with both the unannotated radiograph at full acquired resolution and with a rescaled version of the radiograph with any AI-predicted bounding boxes overlaid (maximum image dimension 833 pixels, preserved aspect ratio). The residents reviewed the radiographs in a research interface separate from the clinical picture archiving and communication system (PACS) and were allowed to adjust the reading environment to their preference. We compared these evaluations to those of the deep learning model using the original interpretation as ground truth.

Statistics

We report area under the curve (AUC), accuracy, sensitivity and specificity of the Faster R-CNN model, as well as mean intersection over union for the bounding boxes proposed by this model. We report accuracy, sensitivity and specificity for each individual resident physician in identifying any fracture both without and with AI support. We report resident accuracy in identifying buckle fracture without and with AI assistance.

We used SciPy 1.7.1 and scikit-learn 1.0.1 for all statistical analysis except for estimating DeLong AUC confidence intervals, for which we used the pROC package in R. We use $\chi 2$ tests to compare (1) the accuracy of AI on younger versus older children, (2) the accuracy of residents overall on younger versus older children, (3) the accuracy of residents without AI versus the AI alone, (4) the accuracy of residents with access to AI predictions versus without AI and (5) the accuracy of residents with AI versus AI alone. Chi-squared ($\chi 2$) tests were performed as 2×2 contingency tables without a Yates correction. Binomial tests with expected probability 0.5 evaluated the significance of differences in accuracy in cases of disagreement between (1) AI versus residents without AI and (2) AI versus residents with access to AI predictions.

A checklist for artificial intelligence in medical imaging (CLAIM) is included as Online Supplementary Material 2 [26].

Results

Artificial intelligence model performance

The Faster R-CNN model demonstrated an AUC of 0.92 (95% confidence interval [CI] 0.87–0.97), accuracy of 88% ($n = 110/125$; 95% CI 81–93%), sensitivity of 88% ($n = 70/80$; 95% CI 78–94%) and specificity of 89% ($n = 40/45$; 95% CI 76–96%) in identifying any fracture (Table 1). The model identified 90% of buckle fractures ($n = 35/39$; 95% CI 76–97%). Mean intersection over union between model-proposed bounding boxes and ground truth boxes in cases containing at least one box was 0.44 ($n = 85$). AI model accuracy was not significantly different for children younger than the median age of 10.5 years (accuracy 90%, $n = 56/62$, 95% CI 80–96%) compared to those at or older than the median age (accuracy 86%, $n = 54/63$, 95% CI 75–93%, $\chi 2 = 0.63$, P-value = 0.43).

The test cases misclassified by AI were manually reviewed and causes of errors were identified (Figs. 2, 3

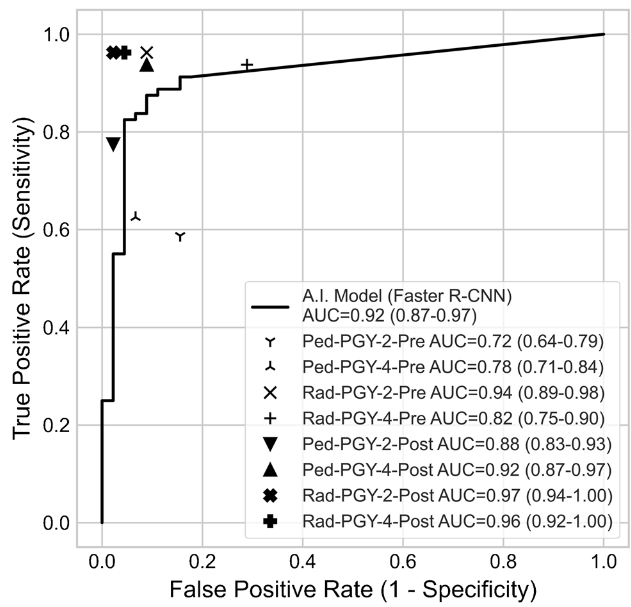

Fig. 2 Comparative AUC for pediatric wrist fracture detection of the Faster R-CNN (“A.I. model”) and individual residents without A.I. assistance (“Pre”, linear symbols) and after A.I. assistance (“Post”, solid symbols). 95% confidence intervals provided in parentheses

Table 1 Performance of artificial intelligence (AI) model and residents alone in identifying all fractures

All fractures	Sensitivity[a]	Specificity[a]	Accuracy[a]	AUC[a]
Faster R-CNN	88% (78–94%, 70/80)	89% (76–96%, 40/45)	88% (81–93%, 110/125)	0.92 (0.87–0.97)
PGY-2 pediatric resident	59% (47–70%, 47/80)	84% (71–94%, 38/45)	68% (59–76%, 85/125)	0.72 (0.64–0.79)
PGY-4 pediatric fellow	63% (51–73%, 50/80)	93% (82–99%, 42/45)	74% (65–81%, 92/125)	0.78 (0.71–0.84)
PGY-2 radiology resident	96% (89–99%, 77/80)	91% (79–98%, 41/45)	94% (89–98%, 118/125)	0.94 (0.89–0.98)
PGY-4 radiology resident	94% (86–98%, 75/80)	71% (56–84%, 32/45)	86% (78–91%, 107/125)	0.82 (0.75–0.90)
All residents	78% (73–82%, 249/320)	85% (79–90%, 153/180)	80% (77–84%, 402/500)	N/A

AUC area under the curve, *N/A* not available

[a] 95% confidence intervals and proportions included in parentheses

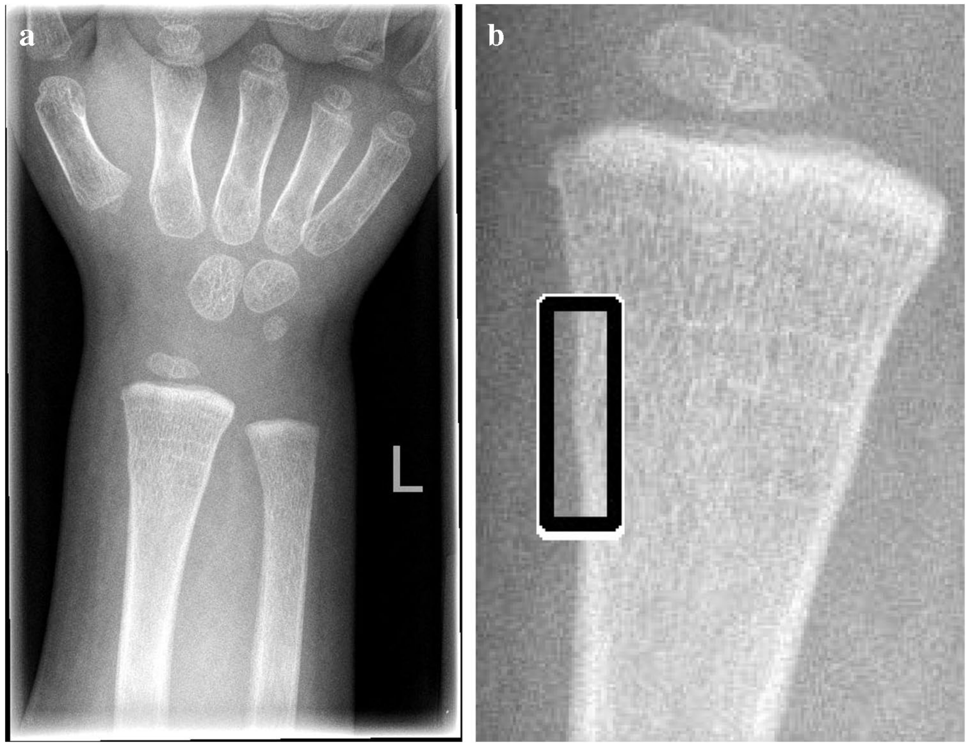

Fig. 3 **a** Buckle fracture of distal radius. Posteroanterior wrist radiograph from 4 year old male. **b** A.I. prediction (*white box*) was concordant with ground truth (*black box*). 0% (0/4) of the residents correctly diagnosed as fracture without A.I. 75% (3/4) of the residents correctly diagnosed the fracture after seeing A.I. predictions

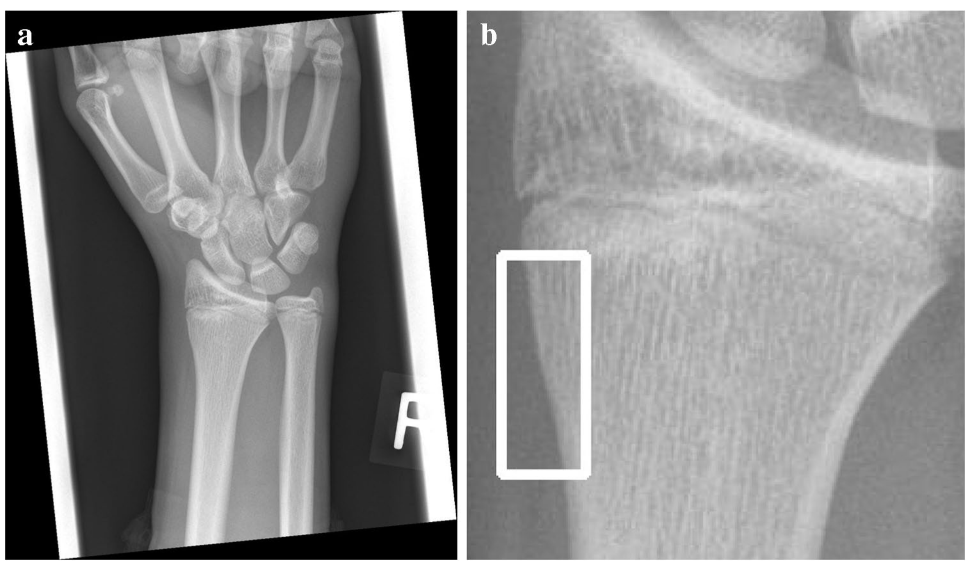

Fig. 4 **a** No fracture. Posteroanterior wrist radiograph from 13 year old male. **b** A.I. prediction (*white box*) was discordant with ground truth, as it erroneously called buckle fracture of distal radius. 100% (4/4) of the residents initially correctly diagnosed as no fracture without A.I. 50% (2/4) of the residents changed their response and incorrectly diagnosed this as a fracture after seeing A.I. predictions

and 4). Of the five true-negative cases that AI misclassified as positive, a small buckle fracture was incorrectly identified in one case, while distal radial physes were incorrectly identified as fracture in four cases (Table 2). Of the 10 true-positive cases that AI misclassified as negative, 3 cases contained distal radial buckle fractures; 2 cases contained mildly displaced Salter–Harris 2 distal radial fractures; and 1 case each contained the following findings: buckle fracture of both the radius and ulna, nondisplaced scaphoid wrist fracture, nondisplaced transverse distal radius fracture, minimally displaced third metacarpal fracture and nondisplaced ulnar styloid fracture, and mildly angulated greenstick fractures of radius and ulna (Fig. 5).

Table 2 Artificial intelligence (AI) confusion matrix

AI confusion matrix	True positive	True negative
Predicted positive	70	5
Predicted negative	10	40

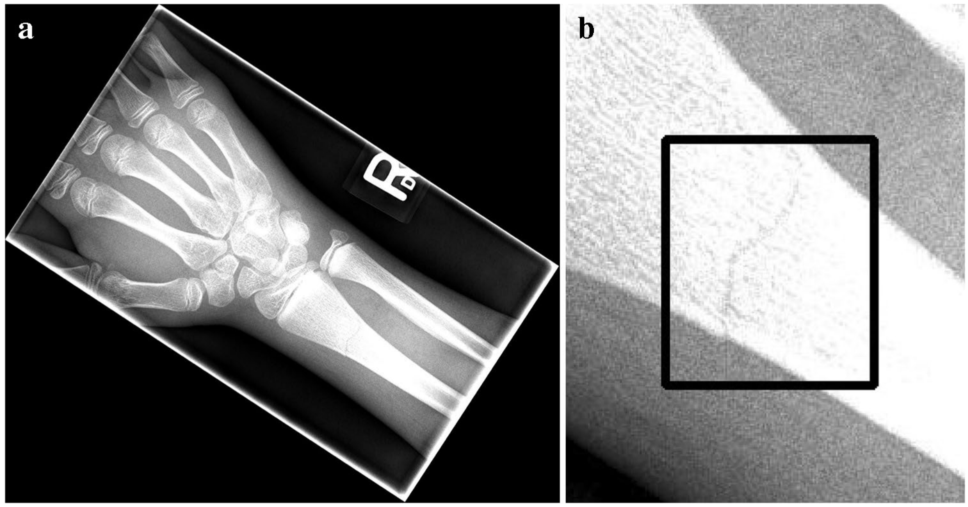

Fig. 5 **a** Transverse fracture distal radius. Posteroanterior wrist radiograph from 15 year old male. **b** A.I. prediction was discordant with ground truth (*black box*), as it did not identify a fracture, and thus no bounding box was offered. 100% (4/4) of the residents correctly diagnosed the fracture without A.I. 100% (4/4) of the residents correctly diagnosed the fracture after seeing A.I. predictions

Table 3 Aggregate resident without artificial intelligence (AI) confusion matrix

Resident confusion matrix	True positive	True negative
Predicted positive	249	27
Predicted negative	71	153

Resident performance with and without artificial intelligence

Accuracy of residents without access to AI was significantly worse than accuracy of AI alone (80% vs. 88%, 95% CI 77–84% vs. 81–93%, $\chi2=3.9$, *P*-value = 0.05) (Table 3). There was no significant difference in aggregate resident accuracy in children younger than the median age of 10.5 years (accuracy 82%, $n=203/248$, 95% CI 76–86%) compared to those at or older than the median age (accuracy 79%, $n=199/252$, 95% CI 73–84%, $\chi2=0.66$, *P*-value = 0.4). Access to AI predictions significantly improved overall resident accuracy from 80 to 93% in detecting all fractures (95% CI 77–84% vs. 90–95%, $\chi2=31.9$, $P<0.001$) (Fig. 4, Table 4) and from 69 to 92% in detecting buckle fractures (95% CI 61–76% vs. 86–95%, $\chi2=26.1$, $P<0.001$) (Table 5).

The difference between the average accuracy of residents with access to AI predictions compared to AI alone did not reach statistical significance (93% vs. 88%, 95% CI 90–95% vs. 81–93%, $\chi2=2.8$, $P=0.10$) (Table 4).

Comparison of artificial intelligence and residents in cases of disagreement

Pooled comparison of resident performance with and without AI is shown in Table 6. When residents did not have access to AI predictions and disagreed with AI, they were significantly more likely to be wrong (33% resident correct [$n=37/112$; 95% CI 24–43%] vs. 67% AI correct [$n=75/112$; 95% CI 57–76%]; binomial test *P*-value < 0.001).

When residents had access to AI predictions and disagreed with AI, they were significantly more likely to be

Table 4 Performance of residents with access to artificial intelligence (AI) in identifying fractures

All fractures	Sensitivity with AI[a]	Specificity with AI[a]	Accuracy with AI[a]	Accuracy improvement with AI	*P*-value of improvement in accuracy ($\chi2$)[b]
PGY-2 pediatric resident	78% (67–86%, 62/80)	98% (88–100%, 44/45)	85% (77–91%, 106/125)	17%	**0.002**
PGY-4 pediatric fellow	94% (86–97%, 75/80)	91% (79–98%, 41/45)	93% (87–97%, 116/125)	19%	**< 0.001**
PGY-2 radiology resident	96% (89–99%, 77/80)	98% (88–100%, 44/45)	97% (92–99%, 121/125)	2%	0.35
PGY-4 radiology resident	96% (89–99%, 77/80)	96% (85–99%, 43/45)	96% (91–99%, 120/125)	10%	**0.004**
All residents	91% (87–94%, 291/320)	96% (91–98%, 172/180)	93% (90–95%, 463/500)	12%	**< 0.001**

[a] 95% confidence intervals and proportions included in parentheses

[b] Statistical significance $P \le 0.05$ (bold)

Table 5 Performance of residents in identifying buckle fractures with and without artificial intelligence (AI) assistance

Buckle fractures	Accuracy pre AI[a]	Accuracy with AI[a]	Accuracy improvement with AI	*P*-value of improvement in accuracy ($\chi 2$)[b]
PGY-2 pediatric resident	41% (26–58%, 16/39)	69% (52–83%, 27/39)	28%	**0.01**
PGY-4 pediatric fellow	44% (28–60%, 17/39)	97% (87–100%, 38/39)	54%	**< 0.001**
PGY-2 radiology resident	95% (83–99%, 37/39)	100% (91–100%, 39/39)	5%	0.15
PGY-4 radiology resident	95% (83–99%, 37/39)	100% (91–100%, 39/39)	5%	0.15
All residents	69% (61–76%, 107/156)	92% (86–95%, 143/156)	23%	**< 0.001**

[a] 95% confidence intervals and proportions included in parentheses

[b] Statistical significance $P \leq 0.05$ (bold)

Table 6 Pooled comparison of resident performance with and without artificial intelligence (AI)

	AI and resident both correct[a]	AI and resident both incorrect[a]	AI correct and resident incorrect[a]	AI incorrect and resident correct[a]
Resident assessments *without* A.I. predictions	73% (69–77%, 365/500)	5% (3–7%, 23/500)	15% (12–18%, 75/500)	7% (5–10%, 37/500)
Resident assessments *with* A.I. predictions	85% (82–88%, 426/500)	5% (3–7%, 23/500)	3% (2–5%, 14/500)	7% (5–10%, 37/500)

[a] 95% confidence intervals and proportions included in parentheses

right (73% resident correct [$n=37/51$; 95% CI 58–84%] vs. 27% AI correct [$n=14/51$; 95% CI 16–42%]; binomial test $P=0.002$).

Resident accuracy improved in some cases when there were correct AI predictions (Fig. 6). Examples of incorrect AI predictions are also shown (Figs. 2, 3 and 4). Some of these incorrect predictions were seen with no change in resident accuracy and others were found with a decrease in accuracy.

Discussion

An object-detection-based Faster R-CNN deep learning approach classified radiographs containing pediatric wrist fractures with high accuracy and demonstrated promising performance both overall and specifically on subtle buckle fractures of the distal radius. Access to AI predictions significantly improved overall average pediatric- and radiology-trained resident accuracy in diagnosing any fracture from 80 to 93% ($P<0.001$) and in diagnosing buckle fracture of the distal radius from 69 to 92% ($P<0.001$).

To our knowledge, machine-learning-based approaches to identifying pediatric fractures of the wrist have not been studied extensively in prior work. Rayan et al. [8] identified elbow fractures using an unsupervised approach on a large dataset of 20,350 training cases and reported an AUC of 0.95 on test data. We note that we used a training set nearly two orders of magnitude smaller and used a single projection and achieved a comparable AUC of 0.92 on a different fracture detection task. Researchers have given more attention to the detection of wrist fracture in adult radiographs, with object-detection-based approaches reporting test AUCs of 0.90 (total dataset $n=14{,}614$ radiographs) [11] and 0.99 (total dataset $n=715{,}343$ radiographs) [6]. More recent work trained and evaluated an ensemble of different object detection models using 542 radiographs from 275 patients, a group that included 21 children younger than 12 years, and reported promising average precision at 50% intersection over union (AP50) of 0.86 [17].

We emphasize that while Faster R-CNN is fundamentally an object detection model, its utility for detecting an object can serve as the basis for classification, and we have evaluated our trained model's classification performance in this work. An object detection approach conveniently learns to generate bounding boxes for findings, allowing model predictions to be shared in a straightforward way with radiologists to maximize effectiveness of human–computer collaboration.

We think the relative ease of interpreting the AI's predictions enabled residents both to incorporate information from it to accurately identify fractures they might otherwise have

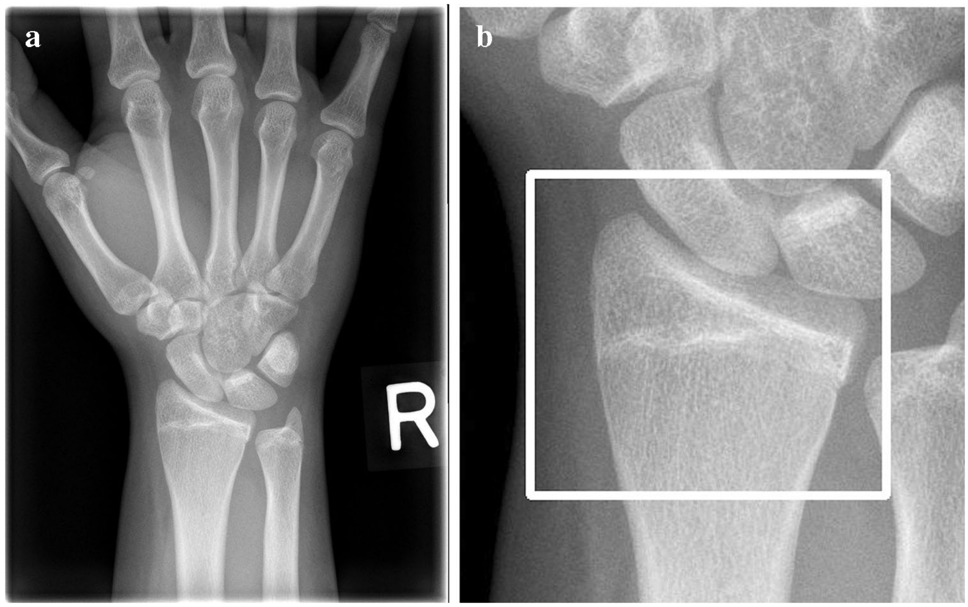

Fig. 6 **a** No fracture. Posteroanterior wrist radiograph from 13 year old female. **b** A.I. prediction (*white box*) was discordant with ground truth, as it erroneously considered a nearly-fused physis a fracture. 100% (4/4) of the residents correctly diagnosed this as no fracture without A.I. 100% (4/4) of the residents correctly diagnosed this as no fracture after seeing A.I. predictions

missed and to critically evaluate its predictions and overrule them when appropriate. AI significantly outperformed residents in cases of disagreement when residents did not have access to its predictions (33% resident correct vs. 67% AI, $P < 0.001$), but the situation reversed when residents could access the AI predictions and still disagreed with the AI (73% resident correct vs. 27% AI, $P = 0.002$). This highlights the complementary nature of human and machine intelligence and demonstrates the potential value of combining them to achieve highest performance.

At our institution, pediatric and radiology residents are responsible for the preliminary interpretation of pediatric ED radiographs for the majority of the day (5 p.m.–7:30 a.m.) and are often the only interpreters before a patient is discharged. While trainees are not prevalent everywhere, this lack of subspecialist review is a model of service that mirrors the situation in the wider medical community [27, 28].

The AI predictions in our study did not benefit everyone equally. They were more helpful for pediatric trainees as compared to radiology trainees, which is intuitive given the increased experience radiology residents have with radiograph interpretation and similar to what has been found by prior investigators in other contexts [29, 30]. While experience could diminish the value of AI assistance, this relationship might not be linear in actual practice. Other variables affect the ability to accurately interpret a radiograph, including the complexity of pathology, time pressure, mental fatigue and the presence of any distractions. It is therefore conceivable that under certain real-life circumstances, AI would benefit experienced readers more than in a controlled study environment. We think similar situations arise in detection of most pathology, where a small group of subspecialists concentrated at academic centers has specialized expertise that might be usefully shared with the wider radiologic community via AI algorithms.

The failures of models such as ours should always be critically assessed. We shared several figure examples of incorrect AI predictions (Figs. 2, 3 and 4). Specific to our dataset, we think that factors such as rotation of images in terms of how they were displayed, degradation of native image resolution, and anatomical variations such as a closing growth plate contributed to inaccurate AI predictions. The effect these incorrect predictions might have had on human interpretation also warrants discussion. In some cases, residents were still able to provide correct responses when the AI prediction was incorrect; however, there were examples when resident accuracy decreased in this setting (e.g., Fig. 2). While we cannot conclude that AI predictions directly led to a decrease in accuracy, this certainly needs to be considered when potential clinical adaptation of such tools is discussed. The false-positive diagnosis of a fracture might not result in a significant clinical consequence, assuming operative intervention is not taken. However, in other potential disease applications, such as identifying malignancy, a false-positive diagnosis can initiate undesired workup and treatment with more harmful consequences [31].

We note several limitations of this study. First, our dataset was limited in size. While certain fractures like buckle fractures were well represented in the data, others were rare, such as scaphoid fracture, which appeared only once each in train and test sets (Online Supplementary Material 1). With very few examples of specific pathology, it is highly uncertain how reliably this model would be able to identify them. Nevertheless, it is remarkable that the model demonstrated strong ability to identify fractures overall despite being trained on a small dataset of only a few hundred examples containing a variety of fracture types. We think this is because many fractures display similar imaging features, and so the trained model develops some ability to generalize to less commonly seen fractures. How much such generalization can be relied on remains highly uncertain, and it would be preferable to have a larger dataset with ample representation of all fractures of concern; a model trained in a similar fashion to a larger dataset would be expected to demonstrate superior performance.

A second limitation is that we used a Faster R-CNN architecture for object detection. While this model has been demonstrated to be effective in medical imaging [32–34], new object detection models are being continuously developed and some have demonstrated superior performance to Faster R-CNN in head-to-head technical comparisons in other contexts [35, 36]. Experimentation with these models, additional data augmentation and additional hyperparameter optimization might offer promising avenues for further improving model performance. Third, our ground-truth bounding boxes were contributed by a single radiology resident guided by the text report created by one of the multiple pediatric attending radiologists at our institution at the time of clinical interpretation. A stronger dataset would contain multiple sets of annotations for each image provided by different radiologists based on imaging findings and establish consensus ground truth between them. Fourth, we acknowledge that detection of pediatric fractures, particularly buckle fractures, might have limited clinical impact in terms of patient outcome. We still think there is value in an accurate diagnosis to help children and parents understand the source of a child's pain and to set expectations for recovery. Fifth, we considered only a single PA view. The standard of practice is for radiologists to have 2–3 views available to them in evaluating fractures of the wrist — typically posteroanterior, lateral and oblique — and a stronger approach would incorporate all of these. Finally, our study was performed at a single site. The performance of medical imaging deep learning models can degrade when applied to different subsets of patients or different sites, and careful real-world performance assessment is critical [37, 38].

While this approach demonstrates promising retrospective performance on a small dataset, further work is clearly needed to translate this technology into real-world deployment. The most important next steps include training models on larger datasets, incorporating all available radiographic views into a single prediction, and rigorously evaluating generalization performance of the model across external sites.

Conclusion

An object-detection-based deep learning approach trained with only a few hundred examples identified radiographs containing pediatric wrist fractures with high accuracy. Access to model predictions significantly improved resident accuracy in diagnosing these fractures.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00247-023-05588-8.

Declarations

Conflicts of interest None

References

1. Hallas P, Ellingsen T (2006) Errors in fracture diagnoses in the emergency department —characteristics of patients and diurnal variation. BMC Emerg Med 6:4
2. Guly HR (2001) Diagnostic errors in an accident and emergency department. Emerg Med J 18:263–269
3. George MP, Bixby S (2019) Frequently missed fractures in pediatric trauma: a pictorial review of plain film radiography. Radiol Clin North Am 57:843–855
4. Jadhav SP, Swischuk LE (2008) Commonly missed subtle skeletal injuries in children: a pictorial review. Emerg Radiol 15:391–398
5. Halsted MJ, Kumar H, Paquin JJ et al (2004) Diagnostic errors by radiology residents in interpreting pediatric radiographs in an emergency setting. Pediatr Radiol 34:331–336
6. Jones RM, Sharma A, Hotchkiss R et al (2020) Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digit Med 3:144
7. Kalmet PHS, Sanduleanu S, Primakov S et al (2020) Deep learning in fracture detection: a narrative review. Acta Orthop 91:215–220
8. Rayan JC, Reddy N, Kan JH et al (2019) Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiol Artif Intell 1:e180015
9. Lindsey R, Daluiski A, Chopra S et al (2018) Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 115:11591–11596
10. Ebsim R, Naqvi J, Cootes TF (2019) Automatic detection of wrist fractures from posteroanterior and lateral radiographs: a deep learning-based approach. In: Glocker B, Yao J, Vrtovec T et al (eds) Computational methods and clinical applications in musculoskeletal imaging. Springer International Publishing, Cham, pp 114–125
11. Thian YL, Li Y, Jagmohan P et al (2019) Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 1:e180001
12. Blüthgen C, Becker AS, Vittoria de Martini I et al (2020) Detection and localization of distal radius fractures: deep learning system versus radiologists. Eur J Radiol 126:108925
13. Ren M, Yi PH (2022) Deep learning detection of subtle fractures using staged algorithms to mimic radiologist search pattern. Skeletal Radiol 51:345–353
14. Lin T-Y, Maire M, Belongie S et al (2014) Microsoft COCO: common objects in context. arXiv [cs.CV]
15. Guermazi A, Tannoury C, Kompel AJ et al (2021) Improving radiographic fracture recognition performance and efficiency using artificial intelligence. Radiology 302:627–636
16. Langerhuizen DWG, Bulstra AEJ, Janssen SJ et al (2020) Is deep learning on par with human observers for detection of radiographically visible and occult fractures of the scaphoid? Clin Orthop Relat Res 478:2653–2659

17. Hardalaç F, Uysal F, Peker O et al (2022) Fracture detection in wrist X-ray images using deep learning-based object detection models. Sensors 22:1285
18. Hernandez JA, Swischuk LE, Yngve DA, Carmichael KD (2003) The angled buckle fracture in pediatrics: a frequently missed fracture. Emerg Radiol 10:71–75
19. Wang X, Peng Y, Lu L et al (2017) ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv [cs.CV]
20. Kluyver T, Ragan-Kelley B, Pérez F et al (2016) Jupyter Notebooks — a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B (eds) Positioning and power in academic publishing: players, agents and agendas. IOS Press, pp 87–90
21. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149
22. Wu Y, Kirillov A, Massa F et al (2019) Detectron2. https://github.com/facebookresearch/detectron2. Accessed 22 Dec 2022
23. Paszke A, Gross S, Chintala S et al (2017) Automatic differentiation in PyTorch
24. Zhao Z-Q, Zheng P, Xu S-T, Wu X (2019) Object Detection With Deep Learning: A Review. IEEE Trans Neural Netw Learn Syst 30:3212–3232
25. Merkel D (n.d.) Docker: lightweight linux containers for consistent development and deployment. https://www.seltzer.com/margo/teaching/CS508.19/papers/merkel14.pdf. Accessed 26 Nov 2022
26. Mongan J, Moy L, Kahn CE (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2:e200029
27. Merewitz L, Sunshine JH (2006) A portrait of pediatric radiologists in the United States. AJR Am J Roentgenol 186:12–22
28. Rosenkrantz AB, Wang W, Hughes DR, Duszak JRR (2018) Generalist versus subspecialist characteristics of the U.S. radiologist workforce. Radiology 286:929–937
29. Eng DK, Khandwala NB, Long J et al (2021) Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. Radiology 301:692–699
30. Wu N, Phang J, Park J et al (2020) Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 39:1184–1194
31. Fenton JJ, Xing G, Elmore JG et al (2013) Short-term outcomes of screening mammography using computer-aided detection: a population-based study of Medicare enrollees. Ann Intern Med 158:580–587
32. Liu B, Luo J, Huang H (2020) Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN. Int J Comput Assist Radiol Surg 15:457–466
33. Su Y, Li D, Chen X (2021) Lung nodule detection based on Faster R-CNN framework. Comput Methods Programs Biomed 200:105866
34. Lu Y, Yu Q, Gao Y et al (2018) Identification of metastatic lymph nodes in MR imaging with faster region-based convolutional neural networks. Cancer Res 78:5135–5143
35. Tan M, Pang R, Le QV (2019) EfficientDet: scalable and efficient object detection. arXiv [cs.CV]
36. Wang C-Y, Bochkovskiy A, Liao H-YM (2022) YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv [cs.CV]
37. Raisuddin AM, Vaattovaara E, Nevalainen M et al (2021) Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep 11:6006
38. Mårtensson G, Ferreira D, Granberg T et al (2020) The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. Med Image Anal 66:101714