Skeletal Radiology
https://doi.org/10.1007/s00256-022-04070-0

SCIENTIFIC ARTICLE

Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning

Daichi Hayashi[1,2] · Andrew J. Kompel[1] · Jeanne Ventre[3] · Alexis Ducarouge[3] · Toan Nguyen[3,4] · Nor-Eddine Regnard[3,5] · Ali Guermazi[1,6]

Received: 16 March 2022 / Revised: 28 April 2022 / Accepted: 28 April 2022

Abstract

Objective We aimed to perform an external validation of an existing commercial AI software program (BoneView™) for the detection of acute appendicular fractures in pediatric patients.

Materials and methods In our retrospective study, anonymized radiographic exams of extremities, with or without fractures, from pediatric patients (aged 2–21) were included. Three hundred exams (150 with fractures and 150 without fractures) were included, comprising 60 exams per body part (hand/wrist, elbow/upper arm, shoulder/clavicle, foot/ankle, leg/knee). The Ground Truth was defined by experienced radiologists. A deep learning algorithm interpreted the radiographs for fracture detection, and its diagnostic performance was compared against the Ground Truth, and receiver operating characteristic analysis was done. Statistical analyses included sensitivity per patient (the proportion of patients for whom all fractures were identified) and sensitivity per fracture (the proportion of fractures identified by the AI among all fractures), specificity per patient, and false-positive rate per patient.

Results There were 167 boys and 133 girls with a mean age of 10.8 years. For all fractures, sensitivity per patient (average [95% confidence interval]) was 91.3% [85.6, 95.3], specificity per patient was 90.0% [84.0,94.3], sensitivity per fracture was 92.5% [87.0, 96.2], and false-positive rate per patient in patients who had no fracture was 0.11. The patient-wise area under the curve was 0.93 for all fractures. AI diagnostic performance was consistently high across all anatomical locations and different types of fractures except for avulsion fractures (sensitivity per fracture 72.7% [39.0, 94.0]).

Conclusion The BoneView™ deep learning algorithm provides high overall diagnostic performance for appendicular fracture detection in pediatric patients.

Keywords Fracture · Pediatric · Adolescent · AI · Emergency · Diagnostic performance

Introduction

Globally speaking, the application of artificial intelligence (AI) based on deep learning in radiology is expanding rapidly. When it comes to its application specifically in Pediatric Radiology, the literature evidence exists since 2009 for bone age assessment based on radiography [1, 2]. In fact, there are only three pediatric-specific AI software programs available in the market today, and they are all targeting bone age assessment [3–7].

When it comes to deep learning tools for fracture detection, there are several software tools available, but none of them are specifically tailored for the pediatric patient population [3]. Literature evidence pertaining to AI-assisted detection of appendicular fractures in pediatric patients remains limited, focusing only on elbow trauma/fractures

✉ Daichi Hayashi
daichi_hayashi@hotmail.com

[1] Department of Radiology, Boston University School of Medicine, 820 Harrison Avenue, FGH Building, 3rd Floor, Boston, MA 02118, USA

[2] Department of Radiology, Stony Brook University Renaissance School of Medicine, HSc Level 4, Room 120, Stony Brook, NY 11794, USA

[3] Gleamer, 117-119 Quai de Valmy, 75010 Paris, France

[4] Service de Radiopédiatrie, Hôpital Armand-Trousseau, AP-HP, Médecine Sorbonne Université, 26 avenue du Docteur Arnold-Netter, 75012 Paris, France

[5] Réseau d'Imagerie Sud Francilien, 2 avenue de Mousseau, 91000 Evry, France

[6] Department of Radiology, VA Boston Healthcare System, 1400 VFW Parkway, Suite 1B105, West Roxbury, MA 02132, USA

Published online: 06 May 2022

[8–10]. These studies reported encouraging results with high sensitivity (ranging 0.91–0.93) and specificity (ranging 0.84–0.92) values, but with relatively low positive predictive values (ranging 0.70–0.87) [8–10]. The current role of deep learning tools seems to be limited to the initial triage of elbow radiographs for pediatric patients presenting with elbow trauma, especially at institutions where dedicated pediatric or musculoskeletal radiologists are not immediately available [3].

In adult patients, however, investigators have reported the use of deep learning tools for the detection of wider varieties of appendicular fractures [11–18]. There has been no publication describing a deep learning model that is capable of detecting all appendicular fractures (i.e., not limited to elbow or wrist fractures) in pediatric populations. Given the fact that fracture patterns and radiographic findings in pediatric patients differ from those in adult patients, it is unknown if a commercial deep learning AI software program (BoneView™, available in Europe, but not yet in the USA) that is created, trained, and validated using adult patients and subsequently trained on pediatric patients can be successfully applied to pediatric patients for automated appendicular fracture detection. We hypothesized that the deep learning algorithm can successfully detect pediatric fractures using the experienced musculoskeletal radiologists' reading as the ground truth. The aim of our study is to determine the diagnostic performance of the AI software based on a deep learning algorithm (BoneView™ which was previously trained using the adult and pediatric patients) for detection of acute appendicular fractures in pediatric patients presenting with a recent history of trauma.

Materials and methods

Dataset

Our current external validation study using the images collected from a USA-based data provider was approved by the WellCare Group (WCG) institutional review board (approval number 20202256). The need for informed consent was waived because our study was retrospective and all images were totally anonymized and stripped of any clinical information.

All anonymized digitalized radiographic images were provided from a single US-based data provider meeting all HIPAA requirements. Greater than 1,000,000 post-traumatic radiographic exams were screened using a Natural Language Processing (NLP) algorithm for relevance based on radiology reports with stratification by body parts and the presence/absence of fractures in pediatric patients (aged 2 to 21 years). Exams with one or more images including a body part not concerned by the intended use of the AI software (pelvis, skull, spine, rib cage) were excluded. This was because pelvic, spine, and rib fractures are rare in children, and we estimated that there were insufficient data in the training dataset to be able to correctly detect these fractures in a pediatric population. The power calculation was based on a previous study [18] that investigated the same AI software for fracture detection in an adult population. Our study narrowed it down from 480 examinations to 300 because we only investigated 5 out of the 8 body parts included in the previous study. We selected the same amount of X-rays per body part and the same prevalence of 50% of positive examinations in each body part. Initially, 380 radiographic exams were provided and presented in a random order to the musculoskeletal radiology readers for interpretation. Of these, 50 exams were excluded due to the fracture being non-acute, 5 exams were excluded due to poor image quality or a lack of minimally required number of views for making a correct radiographic diagnosis, and 25 exams were excluded because the quota for each anatomical location had already reached. In the end, 300 radiographic exams were included in our study, with half of the patients having acute fractures. There is no overlap between these 300 exams and those used to develop the AI algorithm. There were 60 exams per body part (hand/wrist, elbow/upper arm, shoulder/clavicle, foot/ankle, leg/knee). The summary of the study design is shown in Fig. 1. The flowchart of the study sample determination (inclusion and exclusion of cases) is shown in Fig. 2. There were 167 boys (88 with fractures) and 133 girls (62 with fractures) with the mean (± standard deviation) age of 10.8 ± 4.9 years (Table 1), and 173 fractures in total. Details of anatomical locations of the fractures determined by the Ground Truth are summarized in Table 2. Transverse fractures were the most common type of fracture (56 of 173 fractures).

Ground truth

Two experienced board-certified musculoskeletal radiologists (who had 12 years and 8 years of experience in reading pediatric bone X-rays) independently read all radiographs and annotated the presence of fractures on digital radiographs using a proprietary image viewer which was part of the BoneView™ software package (Gleamer, Paris, France). Only acute fractures were considered positive findings in this study. A bounding box was placed to cover the entirety of the fracture, and anatomical location and types of the fracture were recorded [19]. Each fracture was classified as either "non-obvious (subtle nondisplaced fractures that can be missed by nonexpert readers)" or "obvious (clearly displaced, angulated, comminuted, or otherwise readily identifiable)." If the bounding boxes of the two readers overlapped with an Intersection over Union (IoU) of > 25% and the bone specified by the two radiologists was the same,

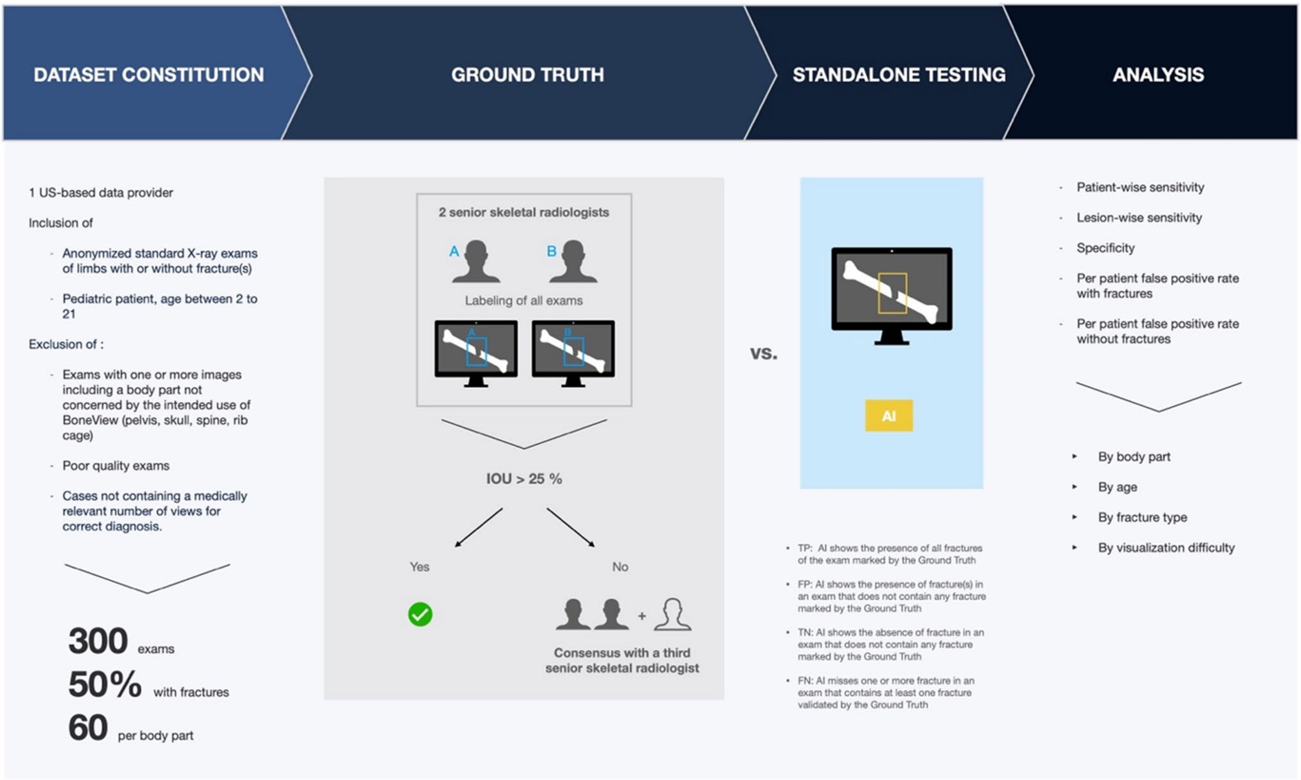
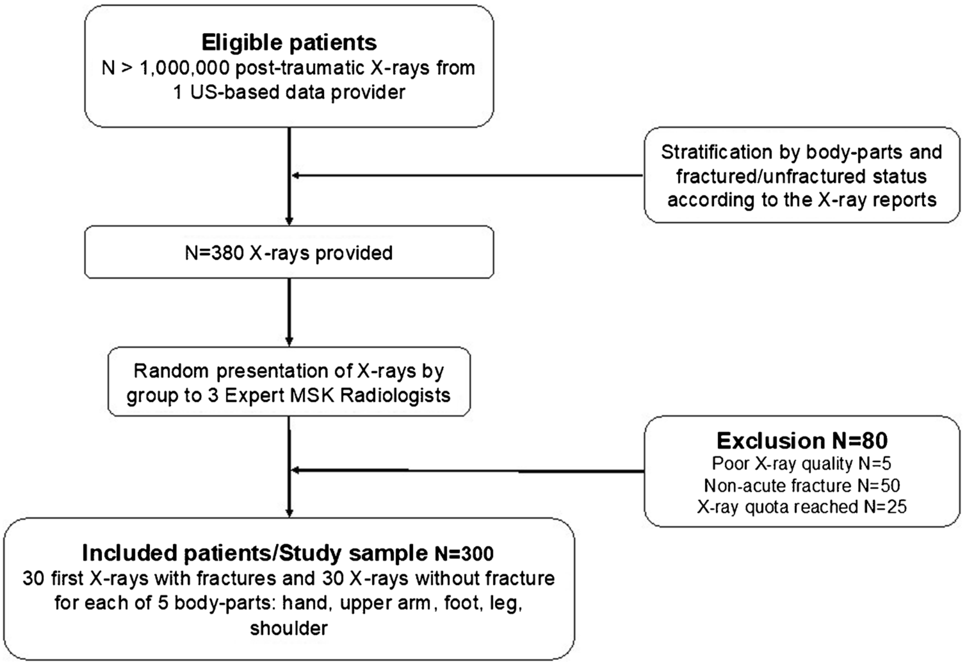

DATASET CONSTITUTION
GROUND TRUTH
STANDALONE TESTING
ANALYSIS
1 US-based data provider
Inclusion of
- Anonymized standard X-ray exams of limbs with or without fracture(s)
- Pediatric patient, age between 2 to 21
Exclusion of :
- Exams with one or more images including a body part not concerned by the intended use of BoneView (pelvis, skull, spine, rib cage)
- Poor quality exams
- Cases not containing a medically relevant number of views for correct diagnosis.
300 exams
50% with fractures
60 per body part
2 senior skeletal radiologists
A
B
Labeling of all exams
IOU > 25 %
Yes
No
Consensus with a third senior skeletal radiologist
vs.
AI
• TP: AI shows the presence of all fractures of the exam marked by the Ground Truth
• FP: AI shows the presence of fracture(s) in an exam that does not contain any fracture marked by the Ground Truth
• TN: AI shows the absence of fracture in an exam that does not contain any fracture marked by the Ground Truth
• FN: AI misses one or more fracture in an exam that contains at least one fracture validated by the Ground Truth
- Patient-wise sensitivity
- Lesion-wise sensitivity
- Specificity
- Per patient false positive rate with fractures
- Per patient false positive rate without fractures
▸ By body part
▸ By age
▸ By fracture type
▸ By visualization difficulty

Fig. 1 Flowchart summarizing the study design

Fig. 2 Flowchart of the study sample determination (inclusion and exclusion of cases)


Eligible patients
N > 1,000,000 post-traumatic X-rays from 1 US-based data provider
Stratification by body-parts and fractured/unfractured status according to the X-ray reports
N=380 X-rays provided
Random presentation of X-rays by group to 3 Expert MSK Radiologists
Exclusion N=80
Poor X-ray quality N=5
Non-acute fracture N=50
X-ray quota reached N=25
Included patients/Study sample N=300
30 first X-rays with fractures and 30 X-rays without fracture for each of 5 body-parts: hand, upper arm, foot, leg, shoulder

then the agreement was reached. The ground truth was then defined as the union between the two boxes. An IoU of 25% seemed reasonable considering that if the fractures pointed by the two radiologists were on different bones, a third expert musculoskeletal radiologist with 26 years of experience was required to reach a consensus by the three readers to adjudicate the discrepancy. Agreement means that two radiologists agreed on the diagnosis without an adjudication

Table 1 Demographics by anatomical location

Body regions	Fracture positive		Fracture negative		Total	
	Boys *n*/*N* (%)	Age Mean ± SD	Boys *n*/*N* (%)	Age Mean ± SD	Boys *n*/*N* (%)	Age Mean ± SD
Foot/ankle	17/30 (56.7%)	11.2 ± 4.9	19/30 (63.3%)	11.1 ± 5.7	36/60 (60.0%)	11.1 ± 5.3
Knee/leg	17/30 (56.7%)	10.3 ± 5.2	16/30 (53.3%)	10.4 ± 4.8	33/60 (55.0%)	10.4 ± 5.0
Hand/wrist	13/30 (43.3%)	10.3 ± 4.2	10/30 (33.3%)	11.3 ± 4.1	23/60 (38.3%)	10.8 ± 4.2
Elbow/arm	24/30 (80.0%)	12.1 ± 5.4	14/30 (46.7%)	10.4 ± 5.0	38/60 (63.3%)	11.3 ± 5.3
Shoulder/clavicle	17/30 (56.7%)	9.8 ± 4.5	20/30 (66.7%)	11.3 ± 4.4	37/60 (61.7%)	10.5 ± 4.5
Total	88/150 (58.7%)	10.7 ± 4.9	79/150 (52.7%)	10.9 ± 4.9	167/300 (55.7%)	10.8 ± 4.9

process. Results of the AI reading were not available to these three radiologists. A fourth reader, a board-certified pediatric radiologist with 4 years of experience, also reviewed all radiographs after determination of Ground Truth and classified all fractures into different types (transverse, oblique, buckle, avulsion, comminuted, greenstick, spiral, longitudinal, hairline, apophyseal avulsion, plastic bending/bowing fracture, and impaction), and ensured known anatomical variants in the pediatric population are correctly identified and not mistakenly diagnosed as a fracture.

Deep learning algorithm

The AI software (BoneView™, Gleamer, Paris, France) was previously engineered, trained, and validated to localize fractures on full-resolution Digital Imaging and Communications in Medicine (DICOM) images. The full details of how the deep learning algorithm was developed have already been described in the existing literature [13, 18]. Thus, only a very brief description of the technical aspects of our deep learning algorithm development is provided herein. To develop the algorithm, we gathered a dataset of 312,602 radiographs of patients from 60 + radiology departments from January 2011 to May 2021 [13, 18]. This dataset was then randomly split into 70% training, 10% validation, and 20% internal test sets. The training dataset included 30% of patients under 21 years old. None of the X-ray images used in the current study was previously used in the development of the BoneView™ software. The deep learning algorithm is based on the Detectron2 framework [20]. Detectron2 is an open-source object detection platform developed by Facebook AI Research and is written in PyTorch (https://pytorch.org/), an open-source machine learning framework containing a library for Python programs that facilitates building deep learning projects. The AI receives as input the full-resolution DICOM image and extracts intermediate feature maps from the X-ray images. The final pipeline of the algorithm can run at different sensitivity and specificity operating points. The software highlights the region of interest using a rectangular box on the radiographic images. Additional editing and refinement of the algorithm were carried out and subsequently integrated into a radiological image interpretation software developed by Gleamer as a tool to support the detection of fractures [13, 18].

The best model was selected based on the free-response receiver operating characteristic (FROC) curves at different operating points (a very specific operating point and a very sensitive operating point) on the internal test set. The final operating points were chosen to obtain a negative predictive value of 99.5% to minimize the false negatives of the algorithm rather than the false positives. All model development was carried out in Python (version 3.6) with all typical libraries including Pydicom used to read DICOM images.

Table 2 Details of anatomical locations of the fractures determined by the Ground Truth

Body region	Fractured bone	Number of fractures: *N* (%)
Total	All	173 (100.0%)
Foot/ankle	Phalanges	10 (5.8%)
	Metatarsus	21 (12.1%)
	Tarsus	3 (1.7%)
	Distal fibula	2 (1.2%)
	Distal tibia	1 (0.6%)
Knee/leg	Femur	8 (4.6%)
	Tibia	18 (10.4%)
	Fibula	5 (2.9%)
	Patella	5 (2.9%)
Hand/wrist	Distal radius	9 (5.2%)
	Distal ulna	5 (2.9%)
	Metacarpus	5 (2.9%)
	Phalanges	18 (10.4%)
Elbow/arm	Humerus	14 (8.1%)
	Radius	14 (8.1%)
	Ulna	5 (2.9%)
Shoulder/clavicle	Proximal humerus	8 (4.6%)
	Clavicle	22 (12.7%)

Statistical analyses

All statistical evaluations were done using Python (version 3.9, libraries SciPy, Scikit-learn, and Pandas). We set the significance threshold at $p < 0.05$ two-sided for all secondary analyses without multiple testing procedures.

To define the true positive, we identified the center of the AI box: if it was inside a Ground Truth box, then the fracture was a true positive; if it was outside the box, it was considered a false positive. The sensitivity (true positive/[true positive + false negative]) per patient was defined as the proportion of patients for whom all actual fractures were identified (each one, on at least one radiographic view), including potentially multiple fractures at more than one region, among patients who had at least one fracture, even if some incorrect marks (false positives) had been detected by the AI. The specificity (true negative/[true negative + false positive]) per patient was defined as the proportion of patients for whom no fracture mark was detected by the AI among patients without any fracture.

The sensitivity per fracture was defined as the proportion of fractures correctly identified by the AI among all fractures, counting multiple fractures per patient where appropriate. The average number of false-positively reported fractures per patient was defined as the mean number of AI boxes positioned outside of a fracture per patient.

We assessed the standalone AI performance using the ROC curve which was derived from the per-patient metrics as described above, using a pre-defined high-sensitivity threshold. This threshold value was manually set in the development set to correspond to a very high negative predictive value (99.5%). Where relevant, 95% confidence intervals (CI) were calculated using Clopper-Pearson exact method [21] to avoid aberrations for proportions equal to 0 or 1.

Results

Overall fracture detection performance

For all fractures, the sensitivity per patient (average [95%CI]) was 91.3% [85.6, 95.3], the specificity per patient was 90.0% [84.0, 94.3], the sensitivity per fracture was 92.5% [87.0, 96.2], the average number of false-positively reported fractures per patient in patients who actually had fractures elsewhere was 0.11 [0.07, 0.18], and the average number of false-positively reported fractures per patient in patients who did not have fractures was 0.11 [0.06, 0.17]. Patient-wise AUC was 0.93 [0.88, 0.97] for all fractures (Table 3). ROC curve for AI diagnostic performance is shown in Fig. 3a.

Performance stratified by demographics and body parts

These excellent diagnostic performances were maintained regardless of patient age group (153 children [aged 2 years or greater and less than 12 years] versus 147 adolescents [aged 12 years or greater and less than or equal to 21 years]; Table 4). There were 100% [88.4, 100] sensitivity per patient and sensitivity per fracture with AUC of 0.99–1.00 [0.87, 1.00] for all upper extremity fractures, while those for lower extremity fractures were lower (sensitivity per patient and sensitivity per fracture were < 85% with AUC of 0.85; Fig. 3b). The average number of false positively reported fractures was highest for the shoulder/clavicle fractures (0.20 [0.08, 0.39] for patients who did not have any fractures, and 0.17 [0.06, 0.35] for patients who had fractures elsewhere).

Table 3 AI performance for fracture detection

	Sensitivity per patient (%) [95% CI]	Specificity per patient (%) [95% CI]	Sensitivity per fracture (%) [95% CI]	Average number of false positively reported fractures per patient in patients without fracture [95% CI]	Average number of false positively reported fractures per patient in patients with fracture [95% CI]	Patient-wise AUC [95% CI]
Foot/ankle	83.3% [65.3, 94.4]	90.0% [73.5, 97.9]	86.5% [69.1, 96.2]	0.13 [0.04, 0.31]	0.13 [0.04, 0.31]	0.85 [0.68, 0.96]
Knee/leg	73.3% [54.1, 87.7]	100.0% [88.4, 100.0]	77.8% [58.9, 90.8]	0.00 [0.00, 0.12]	0.07 [0.01, 0.22]	0.85 [0.67, 0.95]
Hand/wrist	100.0% [88.4, 100]	86.7% [69.3, 96.2]	100.0% [88.4, 100]	0.13 [0.04, 0.31]	0.07 [0.01, 0.22]	0.99 [0.88, 1.0]
Elbow/arm	100.0% [88.4, 100]	90.0% [73.5, 97.9]	100.0% [88.4, 100]	0.10 [0.02, 0.27]	0.10 [0.02, 0.27]	0.99 [0.87, 1.0]
Shoulder/clavicle	100.0% [88.4, 100]	83.3% [65.3, 94.4]	100.0% [88.4, 100]	0.20 [0.08, 0.39]	0.17 [0.06, 0.35]	1.0 [0.88, 1.0]
Total	91.3% [85.6, 95.3]	90.0% [84.0, 94.3]	92.5% [87.0, 96.2]	0.11 [0.07, 0.18]	0.11 [0.06, 0.17]	0.93 [0.88, 0.97]

AUC area under receiver operator curve, *CI* confidence interval

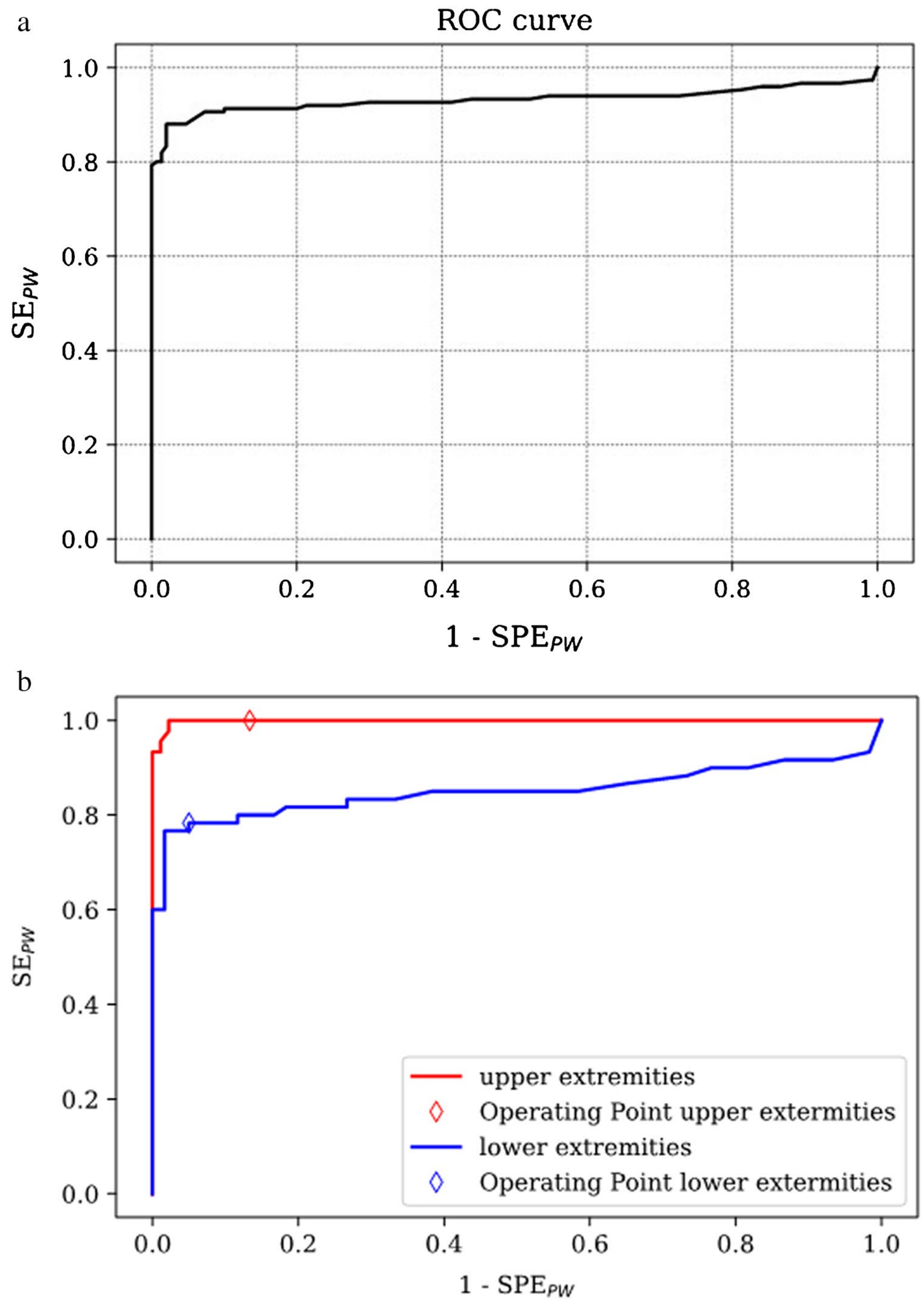


Fig. 3 **a** AI receiver operating characteristic (ROC) for fracture detection in pediatric patients. **b** Separate ROC curves for detection of upper extremity fractures and lower extremity fractures

Performance stratified by types of fractures

When stratifying the data by the types of fractures, AI diagnostic performance was excellent for the buckle, greenstick, hairline, transverse, oblique, spiral, comminuted, and longitudinal fractures with sensitivity per fracture of 90.0 to 100.0%. Except for transverse fractures (91.1% [80.4, 97.0]), these fractures showed relatively wide ranges of 95% CIs due to small sample size, particularly for hairline and longitudinal fractures (both 100.0% [29.2, 100.0] (Table 5). Based on the location of the bone, AI performance was excellent for 32 diaphyseal (sensitivity per fracture of 96.9% [83.8, 99.9]) and 87 metaphyseal fractures (sensitivity per fracture of 93.1% [85.6, 97.4]), but was lower for 24 epiphyseal fractures (sensitivity per fracture of 87.5% [67.6, 97.3]; Table 6). However, AI detected 100% of 21 Salter-Harris type II fractures (95% CI 83.9, 100.0) and 3 type IV fractures (95% CI 29.2, 100.0). AI performance was higher for 125 obvious fractures (sensitivity per fracture of 95.2% [89.9, 98.2]) than 48 non-obvious fractures (85.4% [72.2, 93.9]; Table 7).

Table 4 AI performance for fracture detection depending on age

	Number of patients	Sensitivity per patient (%) [95% CI]	Specificity per patient (%) [95% CI]	Sensitivity per fracture (%) [95% CI]	Average number of false positively reported fractures per patient in patients without fracture [95% CI]	Average number of false positively reported fractures per patient in patients with fracture [95% CI]	Patient-wise AUC [95% CI]
Children (2–12 years old)	153	89.6% [80.6, 95.4]	94.7% [87.1, 98.5]	91.2% [82.5, 96.5]	0.05 [0.02, 0.12]	0.09 [0.04, 0.18]	0.92 [0.84, 0.97]
Children (2–6 years old)	56	93.1% [77.2, 99.2]	96.3% [81.0, 99.9]	93.8% [79.2, 99.2]	0.04 [0.02, 0.13]	0.07 [0.04, 0.13]	0.95 [0.79, 0.99]
Children (6–12 years old)	97	87.5% [74.5, 95.2]	93.9% [83.1, 98.7]	89.8% [79.2, 96.2]	0.10 [0.02, 0.12]	0.06 [0.01, 0.07]	0.91 [0.79, 0.97]
Adolescents (12–21 years old)	147	93.2% [84.7, 97.7]	85.1% [74.9, 92.4]	93.9% [85.7, 98.2]	0.18 [0.10, 0.28]	0.12 [0.06, 0.22]	0.94 [0.86, 0.98]

Children are aged 2 years or greater and less than 12 years. Children are further subdivided into 2 years or greater and less than 6 years, and 6 years or greater and less than 12 years. Adolescents are aged 12 years or greater and less than or equal to 21 years

AUC area under receiver operator curve, *CI* confidence interval

Fractures missed by AI software

Five false-negative cases were noted in the foot/ankle region, including the bowing, buckle, and transverse fractures. Out of 8 false negatives on the leg/knee region, 3 were epiphyseal fractures (for which AI performance was noticeably lower), none were bowing, and 3 were patellar fractures. Figure 4 shows examples of false negative, false positive, and true positive cases.

Discussion

This study showed that our AI software provided high overall diagnostic performance for appendicular fracture detection in pediatric patients, including types of fractures that are unique to skeletally immature patients such as buckle fractures, greenstick fractures, toddler's fractures, and Salter-Harris fractures. Previously, it was shown that AI-assisted adult fracture detection was feasible with high diagnostic performance [13, 18]. Because fractures in pediatric patients are unique and different in terms of types/imaging appearance and typical anatomical locations than those in skeletally mature patients [22, 23], it was necessary to demonstrate that automated fracture detection by AI software was possible in a dedicated pediatric study population. In the current study, overall patient-wise AUC for all fractures was 0.93, which is compatible with previously published adult-based studies in which AUC values of 0.94 [13] and 0.93 [18] were reported.

Our AI software showed excellent diagnostic performance for fractures that are common regardless of patient age (such as transverse and oblique fractures) as well as

Table 5 AI performance for fracture detection depending on the fracture type

	Number of fractures	Sensitivity per fracture (%) [95% CI]
Transverse	56	91.1% [80.4, 97.0]
Oblique	54	100.0% [93.4, 100.0]
Buckle	20	90.0% [68.3, 98.8]
Avulsion	11	72.7% [39.0, 94.0]
Comminuted	8	100.0% [63.1, 100.0]
Greenstick	7	100.0% [59.0, 100.0]
Spiral	7	100.0% [59.0, 100.0]
Longitudinal	3	100.0% [29.2, 100.0]*
Hairline	3	100.0% [29.2, 100.0]*
Apophyseal avulsion	2	50.0% [1.3, 98.7]*
Plastic bending/bowing fracture	1	0.0% [0, 97.5]*
Impaction	1	0.0% [0, 97.5]*

CI confidence interval

*Number of fractures is too small for a statistically meaningful conclusion

Table 6 AI performance for long bone fracture detection depending on fracture location, including humerus, radius, ulna, femur, tibia, fibula, metacarpals, metatarsals, and phalanges

	Number of fractures	Sensitivity per fracture (%) [95% CI]
Diaphysis	32	96.9% [83.8, 99.9]
Metaphysis	87	93.1% [85.6, 97.4]
Epiphysis	24	87.5% [67.6, 97.3]
Physis (Salter Harris II)	21	100.0% [83.9, 100.0]
Physis (Salter Harris IV)	3	100.0% [29.2, 100.0]*

CI confidence interval

*Number of fractures is too small for a statistically meaningful conclusion

those unique to pediatric patients (such as buckle and greenstick fractures). There was only one bowing fracture (or plastic deformations) which AI failed to detect. This type of fracture is typically seen in children commonly affecting the radius, ulna, and fibula [24]. Typical presenting history is a child developing pain and swelling of the forearm after a fall on an outstretched hand. On radiography, these fractures can be visualized as a bowing deformity of the long bone without a visible fracture line or obvious cortical discontinuity. Thus, depending on the view of the radiograph and the direction of bowing, the bone may appear normal [24]. This can make radiographic detection of bowing fractures difficult not only for human interpreters but also for AI software in theory. Unfortunately, we did not have a sufficient number of bowing fractures in our study population and thus cannot derive meaningful scientific evidence. Previously, an attempt was made to use a computer-aided detection (CAD) system to detect bowing fractures in the pediatric forearm [25]. This CAD system showed an AUC of 0.763 to 1.000 for differentiation of radius and ulna with bowing fractures from normal radius and ulna, depending on different threshold values used. However, this was not an AI algorithm and thus the diagnostic performance of their CAD system cannot be directly compared with more recent other AI-related studies. In contrast, our AI showed excellent to perfect sensitivity for detection of greenstick fractures and buckle fractures, which are likely because both these fractures exhibit a visible fracture line and cortical irregularity in addition to concurrent bowing that may be present. Radiographic diagnosis of a bowing fracture can be difficult without relevant clinical history, as other entities such as physiological bowing or bowing secondary to underlying pathology (e.g., neurofibromatosis) may look identical to post-traumatic bowing of a long bone [26].

Table 7 AI performance for fracture detection depending on difficulty

	Number of fractures	Sensitivity per fracture (%) [95% CI]
Obvious	125	95.2% [89.9, 98.2]
Non-obvious	48	85.4% [72.2, 93.9]

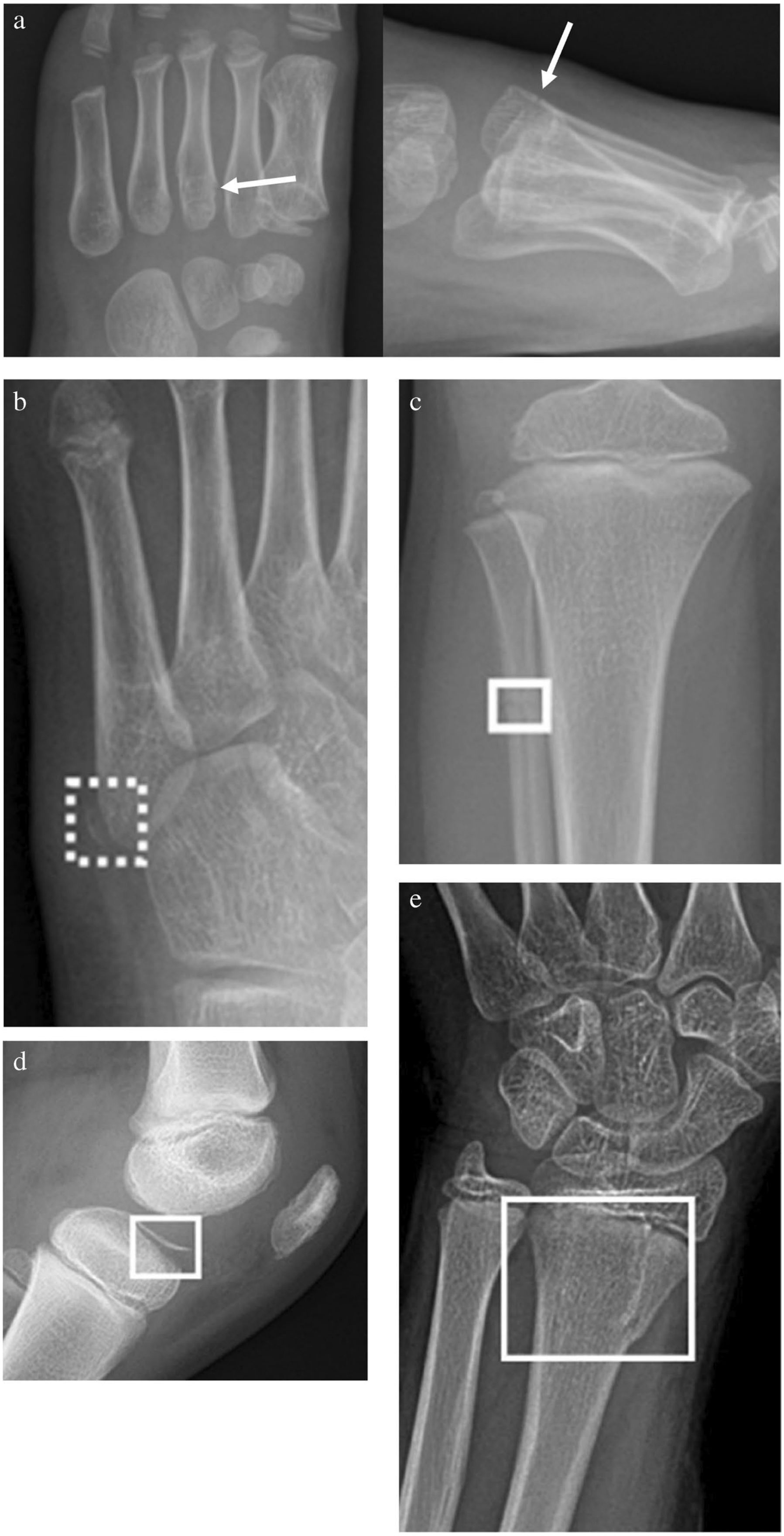
CI confidence interval

Sensitivity for fracture detection per fracture was 96.9% for diaphyseal fractures and 93.1% for metaphyseal fractures. However, detection of epiphyseal fractures was slightly lower (87.5%), even when taking into consideration that the number of epiphyseal fractures was lower than the other two resulting in a much wider 95% CI. Detection of isolated epiphyseal fracture can be difficult even for human readers because of normal-variant irregular ossification of the epiphysis even in the absence of any fracture or other trauma. Similarly, detection of fractures only involving the physis (Salter-Harris Type I) can be difficult because a normal physis can look somewhat irregular in contour as well and the only radiographic abnormality may be a very subtle widening of the physis. However, there was no case of Salter-Harris Type I fracture in our sample; thus, a statistical analysis of AI performance specifically for detection of Salter-Harris Type I fracture cannot be performed. Of note, our AI algorithm detected all 24 Salter-Harris type II and IV fractures in the study sample.

It is interesting to note that our AI software returned perfect fracture detection in all upper extremity fractures (hand/wrist, elbow/arm, shoulder/clavicle) with 100% sensitivity per patient and per fracture. On the other hand, AI diagnostic performance was noticeably lower for lower extremity fractures (< 87% sensitivity per fracture for foot/ankle and < 78% sensitivity per fracture for knee/leg). The overall low diagnostic performance for lower extremity fractures is despite the fact that each anatomical location had the same number of patients (60) and fracture-positive radiographic exams (30). A possible explanation for this is that we had fewer data on lower extremities in the training dataset. Another possible explanation is that there were more non-obvious fractures in the lower extremities (37%) than in the upper extremities (21%). Unsurprisingly, AI detection of obvious fractures showed much higher sensitivity than non-obvious fractures (95.2% versus 85.4%).

There are several reasons that can explain the false positives or false negatives of AI reading. They tend to be more a matter of the data than the AI/computer model. For instance, the algorithm might not be able to recognize a “rare” fracture type, or an unusual view not present in its training dataset. This is a generalization problem from the AI that fails to apply what it has seen in a similar context. Similarly, the image acquisition parameters or the quality of the X-ray may be different from what the AI has seen during its training.

Fig. 4 **a** False negative case, in which AI failed to detect a nondisplaced fracture of the third metatarsal proximal shaft (arrow) which was detected by experienced radiologists. **b** False positive case in which AI erroneously annotated the apophysis of the 5th metatarsal base as a fracture, highlighted by a box with the white dashed line. **c** True positive case showing a nondisplaced transverse fracture of the proximal fibular shaft, highlighted by a white box. **d** True positive case showing an avulsion fracture of the tibial epiphysis at the articular surface, highlighted by a white box. **e** True positive case showing a fracture of the distal radial metaphysis and the physis (Salter-Harris type II fracture), highlighted by a white box



With respect to prior pediatric-dedicated studies of elbow fracture detection using AI, England and colleagues reported an excellent diagnostic performance of their deep convolutional network (DCNN) model with ROC AUC of 0.985 for the validation set and 0.943 for the independent test set [8]. However, radiographically evident fractures were excluded from their study sample and their AI only detected the presence of elbow joint effusion on lateral radiographs (which is a sign of radiographically occult fractures, particularly the supracondylar fractures of the humerus in pediatric patients), and not the fracture itself. Rayan and colleagues used AI to detect various types of fractures in the elbow region, including radiographically occult fractures as indicated by the presence of elbow joint effusion, reporting an overall ROC AUC of 0.9465 [9]. Choi and colleagues applied AI for automated detection of pediatric supracondylar fractures on radiographs, reporting ROC AUC of at least 0.976 for their validation and test sets [10]. Specifically for elbow and arm fractures, our AI showed excellent diagnostic performance with ROC AUC of 0.99 and fracture sensitivity of 100%. Direct comparison of our study with these three studies is difficult, as the study design and outcome measures are notably different. However, our AI identifies fractures and not just elbow effusion, and therefore may be more applicable to other bones and joints as well as different fracture types, making it much more useful and clinically helpful.

There are several limitations to our study. First, we only assessed the standalone performance of the AI software and did not evaluate how the AI can assist human readers. In real-world clinical practice, BoneView™ is used primarily to help human readers (such as but not limited to, radiologists and emergency medicine physicians) and not as a standalone diagnostic tool. Although previous studies have already shown that BoneView™ could help improve human readers' diagnostic performance and also improved diagnostic efficiency [13, 18], those studies did not include pediatric patients. Second, our study was retrospective in nature and AI interpreted the radiographs without any input of clinical information. In real-life scenarios, clinicians typically interpret the radiographs with the knowledge of clinical history and physical examination findings, and thus our study does not reflect the real situation. Third, in our study sample, positive fracture exams are potentially over-represented with the artificially created prevalence of fractures set at 50% in each anatomical location. These include relatively uncommon fracture locations in pediatric patients, and thus it is difficult to generalize our results to real-life scenarios in the pediatric emergency room. However, one cohort study showed actual positive fracture rates (detected in radiographic exams performed in the Emergency Room) of 70.5% in the clavicle, 54.5% in the forearm, 53% in the wrist, and 41.5% in the elbow of pediatric emergency patients, and thus the artificial prevalence of 50% may not be too farfetched for certain upper extremity fractures [27]. Lastly, our study only included acute fractures, although in real life non-acute fractures are also seen in imaging studies. We chose this study design to assess our deep learning algorithm's capability to triage patients who require urgent treatment (i.e., acute fractures) in the emergency room. Since our AI-based tool can immediately notify clinicians which patients have acute fractures as soon as radiographs are acquired, those patients can be given priorities to be treated without delay while waiting for an official radiologist's report. Non-acute fractures were not included in our analysis because they do not require urgent action and clinicians do not need to be alerted immediately for the presence of such fractures. All in all, considering all of the above limitations, our results may not be immediately applicable to real-world clinical practice.

Conclusion

The BoneView™ deep learning algorithm provides high overall diagnostic performance for most types of fracture detection (especially in upper extremities) in pediatric patients.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00256-022-04070-0.

Funding Our research was sponsored by Gleamer, the developer of AI and software (BoneView™). The sponsor was involved with the study design. Manuscript writing was performed by independent authors who were not employees of Gleamer (Daichi Hayashi, Ali Guermazi).

Data availability The trained AI presented in this paper and clinical dataset with its corresponding ground truth, AI readings' results are available upon request from the corresponding author. The development dataset and the AI models, parts of commercial software, are not available to public.

Code availability The code underlying this work can be found online at https://github.com/facebookresearch/detectron2.

Declarations

Ethics approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Our study was approved by the WellCare Group (WCG) institutional review board (approval number 20202256). The need for informed consent was waived because our study was retrospective and all images were totally anonymized and stripped of any clinical information. HIPAA requirements were strictly followed.

Conflict of interest Ali Guermazi is a shareholder of BICL, LLC, and Consultant to Pfizer, Novartis, Regeneron, AstraZeneca, Merck Serono, and TissueGene.
JV, ZZ, AD, TN, AT, EL, AP, Albane Grandjean are employees of Gleamer.
AD is the Chief Technical Officer and co-founder of Gleamer.
NER is the Chief Medical Officer and co-founder of the Gleamer.
All other authors have no other conflict of interest to report.

References

1. Van Rijn RR, Lequin MH, Thodberg HH. Automatic determination of Greulich and Pyle bone age in healthy Dutch children. Pediatr Radiol. 2009;39:591–7.
2. Thodberg HH, Sävendahl L. Validation and reference values of automated bone age determination for four ethnicities. Acad Radiol. 2010;17:1425–32.
3. Offiah AC. Current and emerging artificial intelligence applications for pediatric musculoskeletal radiology. Pediatr Radiol. 2021. https://doi.org/10.1007/s00247-021-05130-8. Online ahead of print.
4. Mutasa C, Chang PD, Ruzal-Shapiro C, Ayyala R. MABAL: a novel deep-learning architecture for machine-assisted bone age labeling. J Digit Imaging. 2018;31:513–9.
5. Tajmir SH, Lee H, Shailam RS, et al. Artificial intelligence assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skelet Radiol. 2019;48:275–83.
6. Pan I, Baird GL, Mutasa S, et al. Rethinking Greulich and Pyle: a deep learning approach to pediatric bone age assessment using pediatric trauma hand radiographs. Radiol Artif Intell. 2020;2:e190198.
7. Reddy NE, Rayan JC, Annapragada AV, et al. Bone age determination using only the index finger: a novel approach using a convolutional neural network compared with human radiologists. Pediatr Radiol. 2020;50:516–23.
8. England JR, Gross JS, White EA, et al. Detection of traumatic pediatric elbow joint effusion using a deep convolutional neural network. AJR Am J Roentgenol. 2018;211:1361–8.
9. Rayan JC, Reddy N, Kan JH, et al. Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiol Artif Intell. 2019;1:e180015.
10. Choi JW, Cho YJ, Lee S, et al. Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. Investig Radiol. 2020;55:101–10.
11. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018;73:439–45.
12. Yu JS, Yu SM, Erdal BS, et al. Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. Clin Radiol. 2020;75:237.e1-237.e9.
13. Duron L, Ducarouge A, Gillibert A, 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. 2021;300:120–9.
14. Tobler P, Cyriac J, Kovacs BK, 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. 2021;31:6816–24.
15. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115:11591–6.
16. Cheng CT, Ho TY, Lee TY, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol. 2019;29:5469–77.
17. Jones RM, Sharma A, Hotchkiss R, et al. Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digit Med. 2020;3:144. https://doi.org/10.1038/s41476-020-00352-w.
18. Guermazi A, Tannoury C, Kompel AJ, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. Radiology. 2022;302:627–36. https://doi.org/10.1148/radiol.210937.
19. Joeris A, Lutz N, Blumenthal A, Slongo T, Audigé L. The AO pediatric comprehensive classification of long bone fractures (PCCF). Acta Orthop. 2017;88:123–8.
20. Wu Y, Kirillov A, Massa F, Lo WY, Girschick R. Detectron2. 2019. https://github.com/facebookresearch/detectron2. Accessed 13th August 2021.
21. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;404–413.
22. Chasm RM, Swencki SA. Pediatric orthopedic emergencies. Emerg Med Clin North Am. 2010;28:907–26.
23. Kim HHR, Menashe SJ, Ngo AV, et al. Uniquely pediatric upper extremity injuries. Clin Imaging. 2021;80:249–61.
24. Crowe JE, Swischuk LE. Acute bowing fractures of the forearm in children: a frequently missed injury. AJR Am J Roentgenol. 1977;128:981–4.
25. Zhou Y, Teomete U, Dandin O, et al. Computer-aided detection (CADx) for plastic deformation fractures in pediatric forearm. Comput Biol Med. 2016;78:120–5.
26. Cheema JI, Grissom LE, Harcke HT. Radiographic characteristics of lower-extremity bowing in children. Radiographics. 2003;23:871–80.
27. Ruffing T, Danko T, Henzler T, Weiss C, Hofmann A, Muhm M. Number of positive radiographic findings in pediatric trauma patients. Emerg Radiol. 2017;24:281–6.