Radiology: Artificial Intelligence
ORIGINAL RESEARCH

Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs

Yee Liang Thian, MBBS, FRCR • Yiting Li, BEng • Pooja Jagmohan, MBBS, FRCR • David Sia, MBBS, FRCR • Vincent Ern Yao Chan, MB, BCh, BAO • Robby T. Tan, PhD

From the Department of Diagnostic Imaging (Y.L.T., P.J., D.S., V.E.Y.C.) and Department of Electrical and Computer Engineering (Y.L., R.T.T.), National University of Singapore, 5 Lower Kent Ridge Rd, Singapore 119074; and Science Division, Yale-NUS College, Singapore (R.T.T.). Received June 27, 2018; revision requested September 5; revision received October 1; accepted December 10. **Address correspondence to** Y.L.T. (e-mail: *yee_liang_thian@nuhs.edu.sg*).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2019; 1(1):e180001 • https://doi.org/10.1148/ryai.2019180001 • Content codes: MK IN

Purpose: To demonstrate the feasibility and performance of an object detection convolutional neural network (CNN) for fracture detection and localization on wrist radiographs.

Materials and Methods: Institutional review board approval was obtained with waiver of consent for this retrospective study. A total of 7356 wrist radiographic studies were extracted from a hospital picture archiving and communication system. Radiologists annotated all radius and ulna fractures with bounding boxes. The dataset was split into training (90%) and validation (10%) sets and used to train fracture localization models for frontal and lateral images. Inception-ResNet Faster R-CNN architecture was implemented as a deep learning model. The models were tested on an unseen test set of 524 consecutive emergency department wrist radiographic studies with two radiologists in consensus as the reference standard. Per-fracture, per-image (ie, per-view), and per-study sensitivity and specificity were determined. Area under the receiver operating characteristic curve (AUC) analysis was performed.

Results: The model detected and correctly localized 310 (91.2%) of 340 and 236 (96.3%) of 245 of all radius and ulna fractures on the frontal and lateral views, respectively. The per-image sensitivity, specificity, and AUC were 95.7% (95% confidence interval [CI]: 92.4%, 97.8%), 82.5% (95% CI: 77.4%, 86.8%), and 0.918 (95% CI: 0.894, 0.941), respectively, for the frontal view and 96.7% (95% CI: 93.6%, 98.6%), 86.4% (95% CI: 81.9%, 90.2%), and 0.933 (95% CI: 0.912, 0.954), respectively, for the lateral view. The per-study sensitivity, specificity, and AUC were 98.1% (95% CI: 95.6%, 99.4%), 72.9% (95% CI: 67.1%, 78.2%), and 0.895 (95% CI: 0.870, 0.920), respectively.

Conclusion: The ability of an object detection CNN to detect and localize radius and ulna fractures on wrist radiographs with high sensitivity and specificity was demonstrated.

Missed fractures on emergency department radiographs are one of the common causes of diagnostic errors and litigation (1,2). Interpretation errors on radiographs are contributed by human and environmental factors, such as clinician inexperience, fatigue, distractions, poor viewing conditions, and time pressures. Automated analysis of radiographs by computers, which are consistent and indefatigable, would be invaluable to augment the work of emergency physicians and radiologists.

In recent years, a machine deep learning technique known as convolutional neural networks (CNNs) has gained rapid traction in the field of computer vision. CNNs "learn" discriminating features from the pixel information of large image datasets to fit the diagnostic problem. Continuous improvements of CNN architectures coupled with a geometric increase in hardware computational power have enabled deep learning CNNs to achieve human level performance in lay tasks, such as facial recognition, handwriting recognition, and natural-world image classification (3,4). Early work applying deep learning CNNs to medical diagnostic image analysis has shown promise in areas such as mammographic mass classification, pulmonary tuberculosis classification on chest radiographs, bone age assessment, and diabetic retinopathy classification (5–8).

A few early studies have shown the feasibility of CNNs in fracture detection on radiographs (9–11). Olczak et al (9) achieved an accuracy of 83% for fracture detection using a network trained on a heterogeneous group of hand, wrist, and ankle radiographs. Kim and MacKinnon (10) were able to attain an area under the receiver operating characteristic curve (AUC) of 0.954 with a model trained on 1389 lateral wrist radiographs. However, these studies were based on binary classification of entire radiographs into fracture or nonfracture categories, and these deep learning networks could not localize the actual region of abnormality. It is difficult for clinicians to trust broad classification labels of such "black-box" models, as it is not transparent how the network arrived at its conclusion. Location information of the abnormality is important to support the classification result by providing visual evidence that is verifiable by the clinician.

Object detection CNNs are extensions of image classification models that not only recognize and classify objects on images, but also localize the position of each object by

Abbreviations

AUC = area under the receiver operating characteristic curve, CI = confidence interval, CNN = convolutional neural network, DICOM = Digital Imaging and Communications in Medicine

Summary

Deep learning object detection networks can be trained to accurately detect and localize fractures on wrist radiographs.

Key Points

- A deep learning object detection network detected and localized radius and ulna fractures on wrist radiographs with high sensitivity at a per-fracture (frontal 91.2%, lateral 96.3%), per-image (frontal 95.7%, lateral 96.7%), and per-study (98.1%) level, even with a relatively modest training dataset size of 7356 radiographic studies.
- There was no significant difference in the performance of the trained network between pediatric and adult wrist radiographs and between radiographs with or without casts present.
- The trained network had significantly lower sensitivity for minimally or undisplaced fractures compared with displaced fractures on both frontal and lateral views.

drawing an appropriate bounding box around it (12). Our hypothesis is that an object detection CNN could be used to detect and localize fractures on wrist radiographs, by treating a fracture as an object. The aim of our study was to determine the feasibility and performance of an object detection CNN to both detect and localize fractures on wrist radiographs.

Materials and Methods

Dataset and Study Population

Our institutional review board approved this single-center retrospective study with a waiver of informed consent. We extracted all wrist radiographs obtained in our institution emergency department between January 1, 2015, and December 31, 2017. A total of 7356 studies were extracted from the picture archiving and communication system repository and were deidentified. Wrist studies where dedicated scaphoid views were obtained were not extracted. The Digital Imaging and Communications in Medicine (DICOM) images were converted to lossless 24-bit grayscale Joint Photographic Experts Group format, maintaining their original resolution and default window and level settings as stored in the DICOM metadata. Images were separated into anteroposterior and lateral projections, and automated cropping was performed to remove noninformative parts of the image outside the collimated field of view. This resulted in a total of 14 614 images (7295 frontal and 7319 lateral), with a mean patient age of 43.5 years ± 23.9 (standard deviation). A total of 4296 (58.9%) of 7295 of the frontal images and 4314 (58.9%) of 7319 of the lateral images were in male patients.

Neural Network Architecture

We used the Faster R-CNN architecture (13), which is a state-of-the-art object detection framework. Faster R-CNN is based on a CNN with additional components for detecting, localizing, and classifying objects on an image. Briefly, an input image is first processed by a base CNN, which consists of convolutional and max-pooling layers, to produce feature maps. After the last convolutional layer, a regional proposal network is trained to propose candidate fracture regions on the image on the basis of a fixed set of anchors on each position of the feature maps. The location and size of each anchor are fine-tuned with bound-box regression. Candidate fracture regions are filtered by performing nonmaximum suppression, which removes redundant overlapping candidate regions of the same fracture. Region proposals are sent to a region-of-interest pooling layer, which resamples the feature maps inside each proposal and are fed to another branch of the network that predicts the confidence scores. The base CNN used in our model was Inception-ResNet version 2, which was pretrained on a large-scale object detection, segmentation, and captioning dataset (common objects in context, or COCO) (14). The model outputs a bounding box for each detected fracture and a score that reflects the confidence of identifying the fracture.

Training Details

We applied horizontal flipping to augment the training dataset. Training set annotations were performed by three radiologists (Y.L.T., P.J., and D.S., with 13, 12, and 10 years of experience, respectively). Each image was reviewed and annotated by one of the three radiologists who drew bounding boxes of any acute radius and ulna fracture locations. Radiologists were informed to draw a rectangular bounding box encompassing the entirety of each fracture, including some adjacent soft tissue and/or normal bone. In instances in which the radiologist thought the fracture was radiologically occult on one view, the radiologist did not mark an annotation on that view (eg, if a fracture was visible on the lateral image but the radiologist could not see it on the frontal image, the radiologist would only annotate the fracture on the lateral image). On lateral images where there were two overlapping fracture locations (eg, overlapping radius and ulna fractures), only one bounding box encompassing both fractures was drawn. The first 500 annotations of each radiologist were checked by the first author for consistency. If there was uncertainty about how to annotate any particular image, the image would be reviewed in consensus with the first author of the study, and a consensus decision made on annotation.

The dataset was split into training (90%) and validation (10%) sets and used to train the fracture localization model. Details of the breakdown of the training and validation sets for the frontal and lateral CNN models are shown in Table 1. We implemented Faster R-CNN by using TensorFlow (*https://www.tensorflow.org/*). One network was trained on frontal images, and another network was trained on lateral images (Fig 1). The intersection over union threshold for foreground objects in the regional proposal network was 0.7, the intersection over union threshold of the final nonmaximum suppression was set to 0.6, and the model trained for 60 000 iterations on two graphics cards (GTX 1080Ti; NVIDIA, Santa Clara, Calif). Gradient descent optimization

Table 1: Details of Training and Validation Sets for Frontal and Lateral CNN Models

Variable	No. of Images	No. of Images with No Fracture Annotation	No. of Images with One Fracture Annotation	No. of Images with Two Fracture Annotations
Frontal images				
Training set	6515	3778 (58.0)	1815 (27.9)	922 (14.2)
Pediatric	1362	755 (55.4)	438 (32.2)	169 (12.4)
Adult	5153	3023 (58.7)	1377 (26.7)	753 (14.6)
With cast	1102	121 (11.0)	704 (63.9)	277 (25.1)
Without cast	5413	3657 (67.6)	1111 (20.5)	645 (11.9)
Validation set	780	439 (56.3)	227 (29.1)	114 (14.6)
Pediatric	147	69 (46.9)	57 (38.8)	21 (14.3)
Adult	633	370 (58.5)	170 (26.9)	93 (14.7)
With cast	146	15 (10.3)	87 (59.6)	44 (30.1)
Without cast	634	424 (66.9)	140 (22.1)	70 (11.0)
Lateral images				
Training set	6537	3878 (59.3)	2659 (40.7)	0
Pediatric	1387	758 (54.7)	629 (45.3)	0
Adult	5150	3120 (60.6)	2030 (39.4)	0
With cast	1118	117 (10.5)	1001 (89.5)	0
Without cast	5419	3761 (69.4)	1658 (30.6)	0
Validation set	782	444 (56.8)	338 (43.2)	0
Pediatric	148	68 (45.9)	80 (54.1)	0
Adult	634	376 (59.3)	258 (40.7)	0
With cast	146	10 (6.8)	136 (93.2)	0
Without cast	636	434 (68.2)	202 (31.8)	0

Note.—Data in parentheses indicate the percentage of images with respect to the total number of images in the corresponding row. CNN = convolutional neural network.

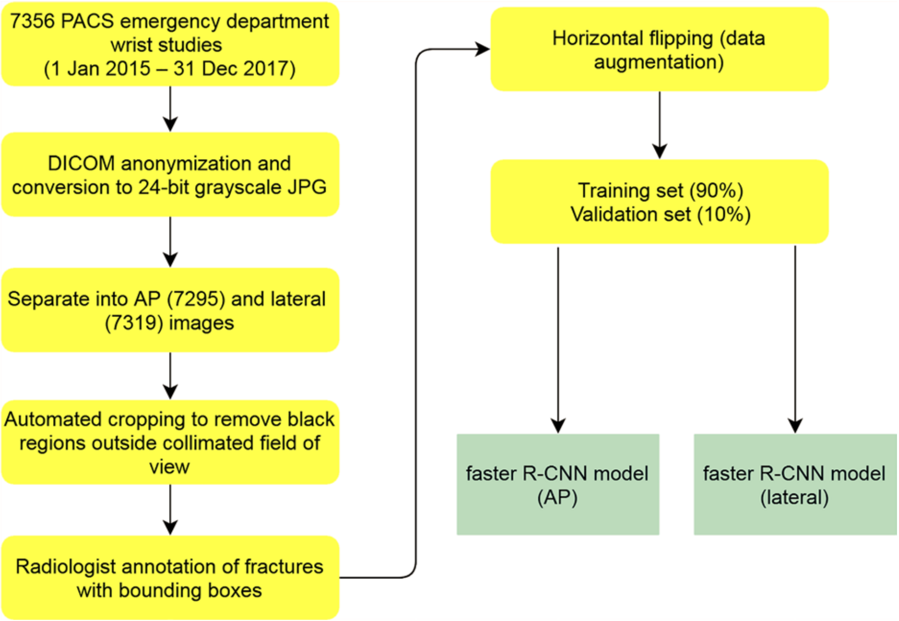

7356 PACS emergency department wrist studies (1 Jan 2015 – 31 Dec 2017)
DICOM anonymization and conversion to 24-bit grayscale JPG
Separate into AP (7295) and lateral (7319) images
Automated cropping to remove black regions outside collimated field of view
Radiologist annotation of fractures with bounding boxes
Horizontal flipping (data augmentation)
Training set (90%) Validation set (10%)
faster R-CNN model (AP)
faster R-CNN model (lateral)

Figure 1: Data flow diagram from raw Digital Imaging and Communications in Medicine (DICOM) images to model training. AP = anteroposterior, CNN = convolutional neural network, JPG = Joint Photographic Experts Group, PACS = picture archiving and communication system.

was used with a batch size of two and an initial learning rate of 0.0003.

Testing Details

The trained neural network was tested on a previously unseen test set of 524 consecutive emergency department wrist radiographs (total 1048 images) obtained between January 1, 2018, and March 31, 2018. Two radiologists (Y.L.T. and D.S., 13 years and 10 years of experience, respectively) looked at the full resolution test images in DICOM format on medical-grade monitors and had access to the final reports, prior as well as follow-up studies, and clinical details in cases with indeterminate findings. All fracture locations were marked with bounding boxes by the two reviewers in consensus, and these constituted the reference standard. Instances where a fracture was present

Table 2: Performance of Fracture Predictions by Deep Learning Object Detection Network

Image Projection	No. of Ground-Truth Fracture Marks	No. of CNN Fracture Marks	No. of True-Positive CNN Marks	Per-Mark Sensitivity (%)	Per-Image Sensitivity (%)	Per-Image Specificity (%)	AUC (%)
Frontal	340	370	310	91.2 (87.6, 94.0)	95.7 (92.4, 97.8)	82.5 (77.4, 86.8)	0.918 (0.894, 0.941)
Lateral	245	276	236	96.3 (93.1, 98.3)	96.7 (93.6, 98.6)	86.4 (81.9, 90.2)	0.933 (0.912, 0.954)

Note.—Data in parentheses are 95% confidence intervals. AUC = area under the receiver operating characteristic curve, CNN = convolutional neural network.

on one view but deemed radiologically occult on the orthogonal view were only marked on the view where the fracture was visible. True-positive marks by the network were scored by reviewers when a region of interest localized by the CNN as fracture with greater than 95% probability overlapped with a reference mark. All other annotations by the CNN were considered false-positive. Per-image (per-view) true-positive determination required at least one true-positive fracture mark on the image. Per-study true-positive determination required at least one true-positive fracture mark on either or both frontal and lateral view of that study. Subgroup analysis was performed for pediatric (defined as younger than 21 years) patients, images with casts, and minimally displaced fractures (defined as angulation < 15°, tilt < 20°, articular step-off < 2 mm, axial radial shortening < 5 mm) (15) to determine if these variables significantly affected model performance.

Statistical and Data Analysis

All statistical analysis was performed with software (Stata 12, Stata Statistical Software: Release 12 [2011]; Stata, College Station, Tex). We determined per-image (ie, per-view) and per-study sensitivity and specificity. AUC analysis was performed on a per-image (per-view) and per-study basis, and comparison of AUCs was made with the nonparametric approach of DeLong et al (16). Evaluation of statistically significant sensitivity differences of the network between subgroups was performed by using χ^2 analysis. A P value less than .05 was considered to indicate a significant difference.

Results

Test Cohort

The test cohort consisted of 524 consecutive emergency department wrist radiographs (524 frontal and 524 lateral images) in 447 patients. The mean age of the test cohort was 39.7 years ± 24.5 (standard deviation), and 285 (54.4%) of 524 were male patients. The test cohort included 159 (30.3%) of 524 radiographs in pediatric patients (defined as age < 21 years). A total of 112 (21.4%) of 524 of the radiographs were obtained with a cast in situ. There were 585 radius and ulna fracture locations in total marked by the two expert readers on all 1048 images. There were 256 frontal images with fractures, consisting of 172 images with one fracture mark and 84 images with two fracture marks, accounting for 340 fracture locations in total. There were 244 lateral images with fractures, consisting of 243 images with one fracture mark and one image with two fracture marks, accounting for 245 fracture locations in total. Of these, 188 fractures on the frontal view and 97 fractures on the lateral view (48.7% of all fracture marks) were classified as minimally displaced. Other fractures (eg, metacarpals, carpal bone) were not considered in the analysis. A total of 548 images (268 frontal, 280 lateral) had no marks and for purposes of this study were considered normal.

Deep Learning Model Performance

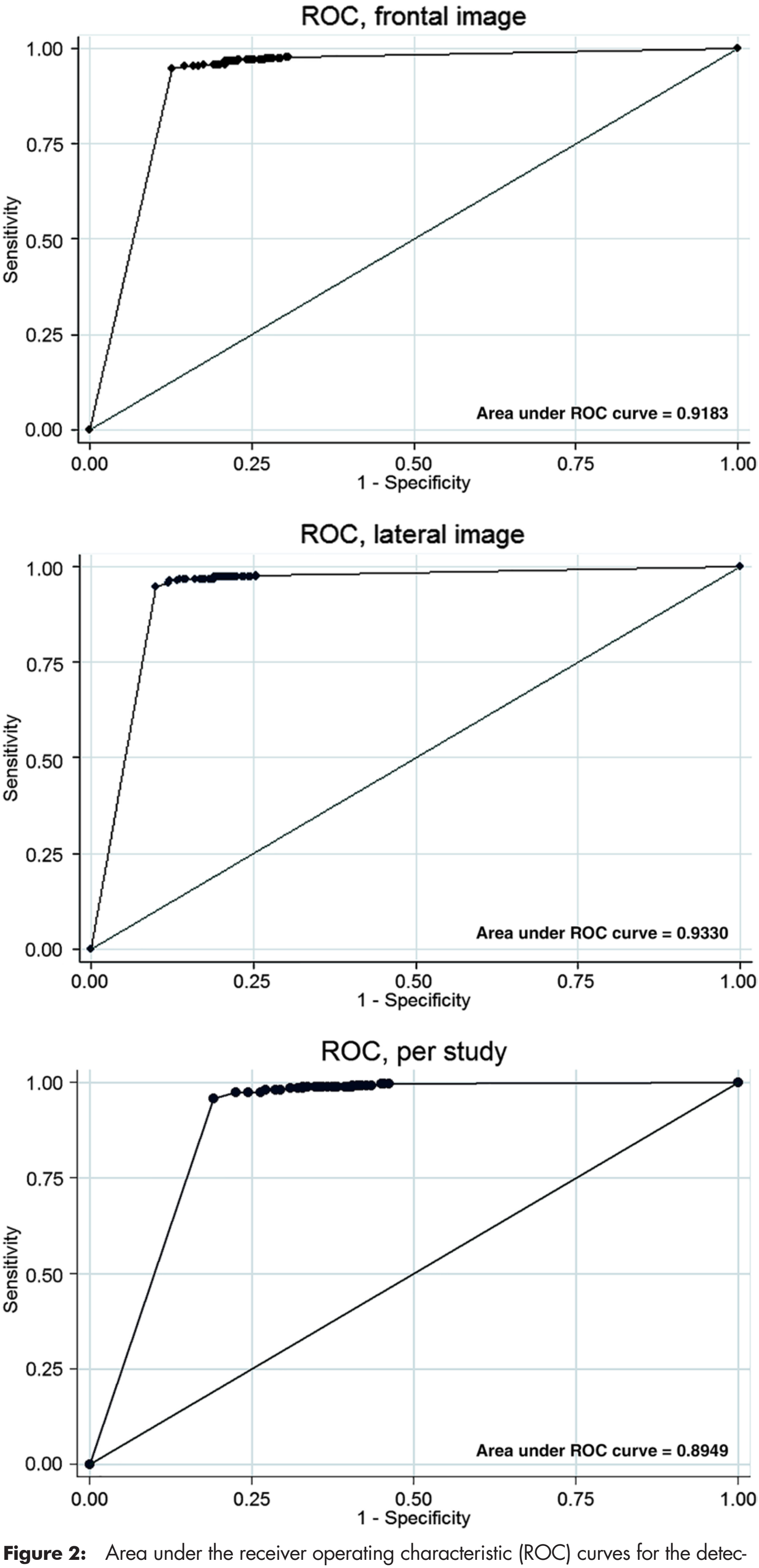
Total training time of the deep learning model was 14.3 hours. Mean processing time per test image was 0.18 second. With a threshold of 95% probability as positive for fracture by the deep learning model, there were 646 marks (370 frontal, 276 lateral) in total made by the model on the test set, of which 546 of 646 marks (310 of 370 frontal, 236 of 276 lateral) were true-positive (overlapped with marks made by expert reviewers) and 100 of 646 (60 frontal, 40 lateral) were false-positive. The per-image sensitivity, specificity, and AUC were 95.7% (95% confidence interval [CI]: 92.4%, 97.8%), 82.5% (95% CI: 77.4%, 86.8%), and 0.918 (95% CI: 0.894, 0.941), respectively, for the frontal view and 96.7% (95% CI: 93.6%, 98.6%), 86.4% (95% CI: 81.9%, 90.2%), and 0.933 (95% CI: 0.912, 0.954), respectively, for the lateral view (Table 2).

The per-study sensitivity, specificity, and AUC were 98.1% (95% CI: 95.6%, 99.4%), 72.9% (95% CI: 67.1%, 78.2%), and 0.895 (95% CI: 0.870, 0.920), respectively (Fig 2).

Performance examples are shown in Figures 3-5.

Subgroup Analysis

Subgroup analysis was performed to determine if any variables influenced network performance for fracture detection. Sensitivity, specificity, and AUC for the pediatric subgroup, subgroup of images with casts, and subgroup of minimally displaced fractures are shown in Table 3. There were no significant differences in AUCs between pediatric and adult images (P = .23 for frontal and P = .42 for lateral) or between images with and without casts (P = .78 for frontal and P = .07 for lateral). The neural networks were significantly more sensitive for displaced fractures compared with minimally or undisplaced fractures on both frontal and lateral views (P = .005 for frontal and P = .01 for lateral).



ROC, frontal image

ROC, lateral image

ROC, per study

Figure 2: Area under the receiver operating characteristic (ROC) curves for the detection of fractures on a per-image and per-study basis. Data points represent empirical operating points based on a cutoff value of the convolutional neural network confidence score.

Discussion

In this study, we approached fracture detection as an object detection problem in computer vision and used a deep learning object detection network for image analysis. The task of object detection involves two fundamental questions about an image: what object is in it, and where it is within the image. This is in contrast to prior studies involving deep learning that approached fracture detection as an image classification problem, which describes what is in the image, but not where it is. We found that a deep learning object detection network was able to detect and localize radius and ulna fractures on wrist radiographs with high sensitivity at a per-fracture (frontal 91.2%, lateral 96.3%), per-image (frontal 95.7%, lateral 96.7%), and per-study (98.1%) level, even with a relatively modest training dataset size of 7356 radiographic studies.

Our deep learning network performance for fracture detection exceeds that of prior published work using deep learning on orthopedic radiographs. Olczak et al trained a deep learning network with 256000 wrist, hand, and ankle radiographs to a final accuracy of 83% for fracture classification (9) compared with our overall accuracy of 88.9% for frontal radiographs and 91.2% for lateral radiographs. Our AUC for lateral wrist radiographs (0.933) is also comparable to Kim and MacKinnon's results (10) for classifying lateral wrist radiographs (0.954). There are differences in our methodology, which we believe are key to our improved fracture detection rate. First, we used a more homogeneous set of wrist radiographs only for training, but Olczak et al included wrist, hand, and ankle radiographs to train their network. A diversity of image types (ie, different parts of the body) may make it more difficult for a single network to converge on an accurate solution to accommodate all anatomic regions (17). Second, all our training data were manually checked and annotated by radiologists. This is more time-consuming but provides more accurate data labeling. In contrast, Olczak et al used automated annotation by using language extraction software applied to radiologists' reports. Automatic annotation is subject to errors, which introduces noisy labels in the training dataset, which could adversely affect the classification accuracy (18). Third, our use of bounding boxes to indicate the location of abnormality helps to refine training of the network to detect image features that are pertinent to the problem at hand (19). Last, we used a state-of-the-art object detection network

(Inception-ResNet version 2 with Faster R-CNN), which may be more efficient at extracting relevant features from the training data compared with the older VGGNet and Inception networks used by Olczak et al and Kim and MacKinnon, respectively (20).

We further improved on Kim and MacKinnon's study on wrist radiographs as our training and test datasets better reflect the diversity of wrist radiographs seen in clinical practice. We analyzed frontal wrist images, pediatric radiographs, and radiographs with casts, which were all excluded from their study. Pediatric radiographs of the wrist have multiple growth plates that can mimic the appearance of a fracture. Radiographs with casts have a diffuse artifact (Fig 3), which obscures underlying bone detail, again compounding the difficulty of fracture detection. Our results show few false-positive marks due to growth plates on pediatric radiographs, indicating the deep learning CNN was able to learn features that distinguish between a growth plate and a fracture. There was no significant difference in performance in the network between fractures in casts and fractures without casts, although we noted a trend toward lower specificity for radiographs obtained with a cast on both frontal and lateral images. This may be because of superimposed linear artifacts from a cast, which can mimic the appearance of fractures, causing false-positive results. We also observed a significantly lower sensitivity for undisplaced or minimally displaced fractures compared with displaced fractures on both frontal and lateral views (Fig 5). This reflects the inherent difficulty of the task of detection of minimally displaced fractures, because their detection depends on a small proportion of pixels of the entire image.

The ability of an object detection network to accurately localize the area of suspected abnormality

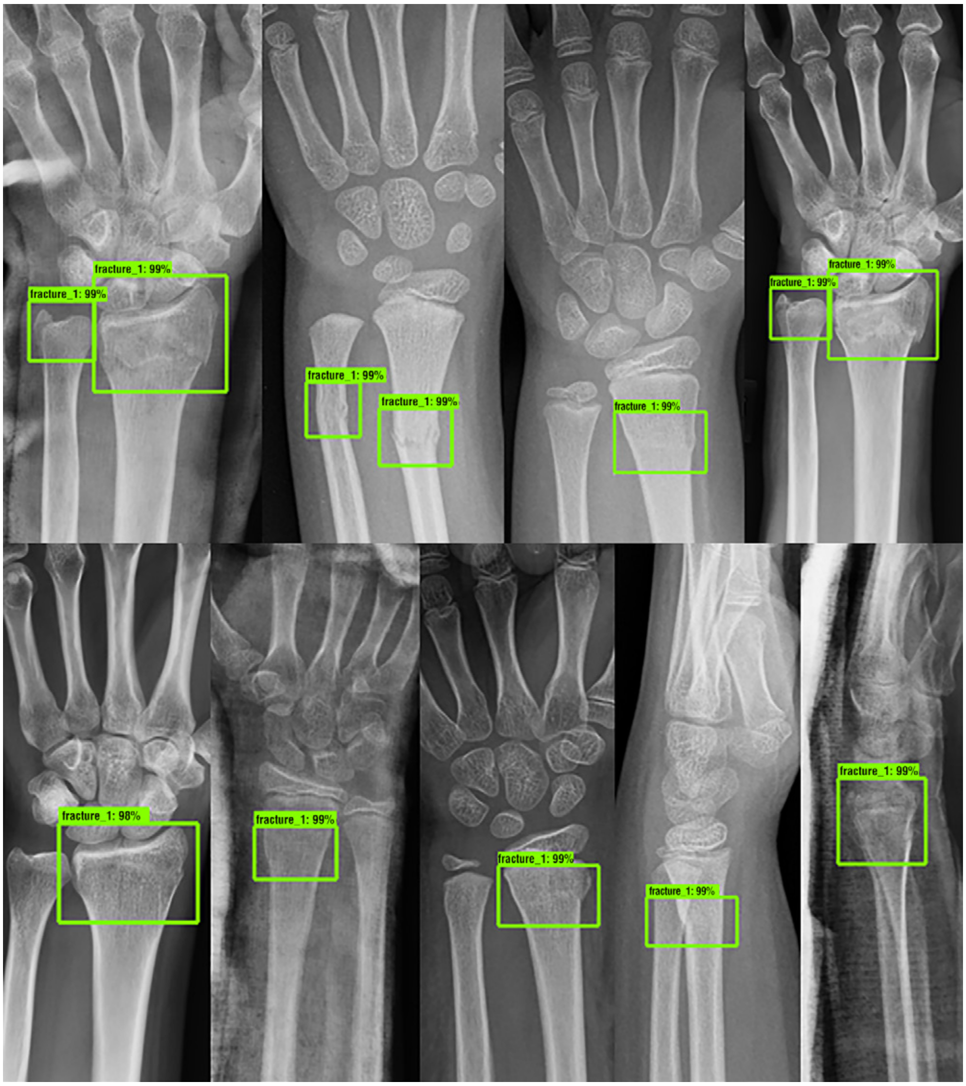


Figure 3: Radiographs show selected true-positive examples of radius and ulna fractures. Green boxes are marks made by the Faster R-convolutional neural network deep learning network trained to detect and localize fractures. Percentages given for each mark reflect the confidence score by the network of a fracture located within the marked box.

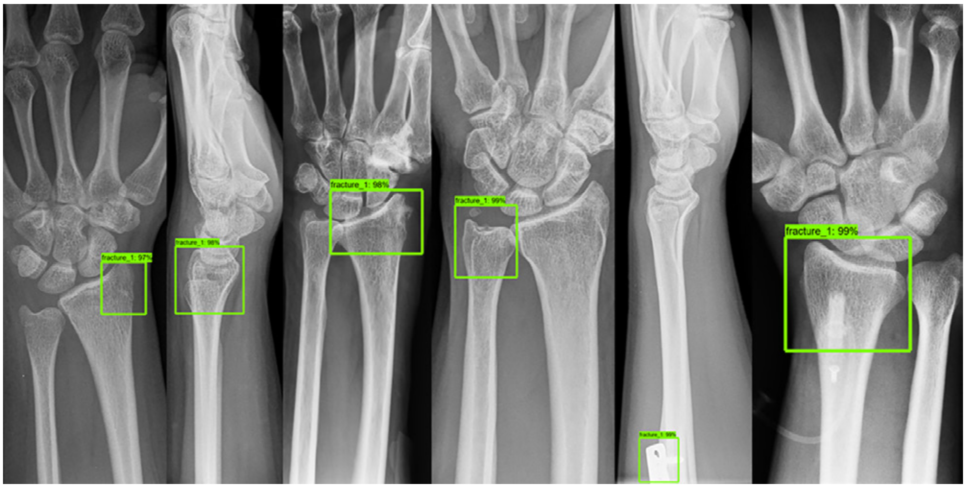


Figure 4: Radiographs show selected false-positive examples. Old fractures and artifacts on the image were a common cause of false-positive marks (green boxes) made by the trained networks. Percentages given for each mark reflect the confidence score by the network of a fracture located within the marked box.

Table 3: Subgroup Analysis of Variables Influencing Model Performance for Fracture Detection on a Per-Image Level

	No. of Images		Sensitivity (%)		Specificity (%)		AUC (%)		*P* Value*	
Segment	Frontal	Lateral	Frontal	Lateral	Frontal	Lateral	Frontal	Lateral	Frontal	Lateral
Age group										
Pediatric	159	159	92.7 (85.6, 97.0)	93.5 (86.5, 97.6)	76.2 (63.8, 86.0)	86.4 (75.7, 93.6)	0.893 (0.841, 0.945)	0.921 (0.877, 0.966)	.23	.42
Adult	365	365	97.5 (93.7, 99.3)	98.7 (95.3, 99.8)	84.4 (78.7, 89.1)	86.4 (81.1, 90.7)	0.928 (0.902, 0.954)	0.942 (0.920, 0.964)		
Presence of cast										
With cast	112	112	98.1 (93.4, 99.8)	95.2 (89.1, 98.4)	40.0 (5.27, 85.3)	62.5 (24.5, 91.5)	0.886 (0.689, 1.00)	0.779 (0.603, 0.956)	.78	.07
Without cast	412	412	94.0 (88.8, 97.2)	97.9 (93.9, 99.6)	83.3 (78.2, 87.6)	87.1 (82.6, 90.9)	0.914 (0.888, 0.941)	0.942 (0.921, 0.964)		
Fracture displacement										
Minimal or undisplaced	188	97	85.6 (79.8, 90.3)	92.7 (85.6, 97.0)		…	…		.005	.01
Displaced	152	148	95.3 (90.7, 98.1)	99.3 (96.3, 100)		…	…			

Note.—Data in parentheses are 95% confidence intervals.

* *P* values shown refer to area under the receiver operating characteristic curve (AUC) comparisons for each variable, except for fracture displacement where *P* value refers to sensitivity comparisons.

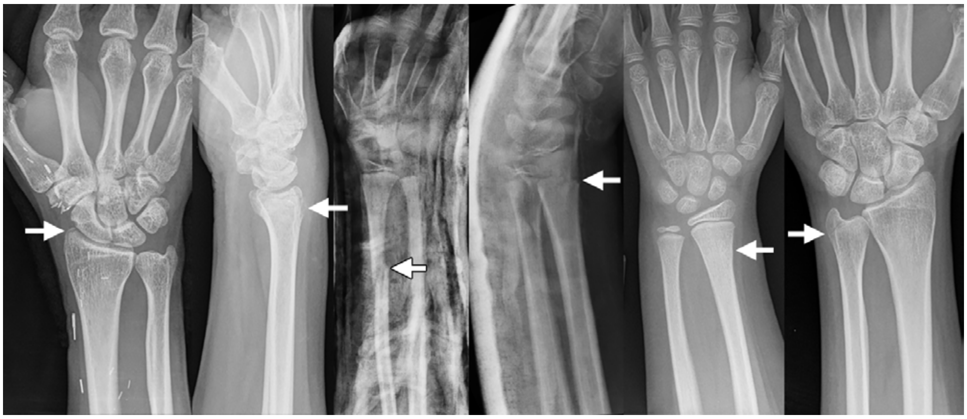


Figure 5: Radiographs show selected false-negative examples with no marks made by the trained networks. The networks had lower sensitivity for undisplaced or minimally displaced fractures (arrows).

(in our case, fractures) overcomes some of the shortcomings of using deep learning classification networks for image analysis. It has been previously argued that deep learning classification networks constitute black boxes in image analysis (21), as they do not provide any rationale for their final classification result. It is thus difficult for a clinician to trust or verify the validity of the network result (22). The object detection network used in our study provides classification as well as spatial localization information, which is more informative than a single classification label and easily verifiable by the clinician. Such location information would be useful in developing deep learning clinical algorithms to aid radiologists in reporting.

We reviewed the misclassifications and made the following observations. The network frequently made false-positive labels on old fractures or deformities (Fig 4), suggesting there was considerable overlap in these findings with the learned features of acute fractures. Our trained network also does not have access to clinical information and old radiographs that a radiologist can use to differentiate the acute nature of an imaging finding. The network also made false-positive labels on a number of normal-appearing ulnar styloid processes, especially those obtained with the ulnar styloid process in some degree of rotation or with an unusually large-appearing ulnar styloid. This highlights the fact that "learned" features of the network using deep learning are not perfect, and unexpected results and misclassifications may occur. Another observation is that despite the overall high sensitivity of the CNN, we noted an example of a relatively obvious Salter-Harris type I fracture on the test set that was not detected by the network. This false-negative case is likely because of the rare occurrence of such fractures, even in our large training set. This illustrates a pitfall of existing deep learning models—rare or unusual-appearing abnormality may be missed because of insufficient training examples of such abnormality in the training set used for developing the model (23).

There were several limitations of our study. First, we only included radius and ulna fractures and did not evaluate all potential fractures on a wrist radiograph, such as carpal or metacarpal fractures. This is because radius and ulna fractures are much more prevalent, and thus obtaining sufficient training examples for deep learning was feasible. We are uncertain if the model would be able to perform adequately if there are limited training examples of certain classes (eg, rare carpal fractures). Deep learning with only limited training examples of a specific class is a computer vision problem that is currently an area of active research (24). Second, we only tested our model on emergency department wrist radiographs, and our results may not generalize to other settings such as orthopedic outpatient radiographs. We excluded training and testing with orthopedic outpatient radiographs because of the large proportion of metallic implants in routine orthopedic outpatient radiographs of the wrist. Including such radiographs may unintentionally teach the CNN to associate the presence of metallic implant with presence of a fracture, rather than discriminative features of the fracture per se. Furthermore, the problem of fracture detection and localization is better represented in the emergency department setting, rather than in the outpatient orthopedic setting where fractures are frequently known, and follow-up of healing and/or alignment are the main concerns. Third, this was a retrospective study with the training and test sets from a single institution, and the ability of the model to generalize to radiographs obtained at external institutions with other machines and processing techniques is unknown. Fourth, the ultimate clinical verification of a predictive artificial intelligence tool requires demonstration of value through effect on patient outcomes, beyond the performance metrics that are evaluated in this study.

In conclusion, our study demonstrated the feasibility of a deep learning object detection network for accurate detection and spatial localization of radius and ulna fractures on wrist radiographs. The ability to predict location information of abnormality with deep neural networks is an important step toward developing clinically useful artificial intelligence tools to augment radiologist reporting.

Author contributions: Guarantor of integrity of entire study, Y.L.T.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, Y.L.T., Y.L., P.J., D.S., R.T.T.; clinical studies, Y.L.T., Y.L., P.J., D.S., V.E.Y.C., R.T.T.; statistical analysis, Y.L.T., V.E.Y.C., R.T.T.; and manuscript editing, Y.L.T., P.J., D.S., R.T.T.

Disclosures of Conflicts of Interest: **Y.L.T.** disclosed no relevant relationships. **Y.L.** disclosed no relevant relationships. **P.J.** disclosed no relevant relationships. **D.S.** disclosed no relevant relationships. **V.E.Y.C.** disclosed no relevant relationships. **R.T.T.** disclosed no relevant relationships.

References

1. Petinaux B, Bhat R, Boniface K, Aristizabal J. Accuracy of radiographic readings in the emergency department. Am J Emerg Med 2011;29(1):18–25.
2. Pinto A, Reginelli A, Pinto F, et al. Errors in imaging patients in the emergency setting. Br J Radiol 2016;89(1061):20150914.
3. Goodfellow IJ, Bulatov Y, Ibarz J, Arnoud S, Shet V. Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv [preprint]. https://arxiv.org/abs/1312.6082. Posted December 20, 2013. Updated April 14, 2014. Accessed June 22, 2018.
4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2016: 770–778.
5. Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J Digit Imaging 2017;30(4):499–505.
6. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284(2):574–582.
7. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402–2410.
8. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287(1):313–322.
9. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017;88(6):581–586.
10. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73(5):439–445.
11. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018;89(4):468–473.
12. Batchelor O, Green R. The role of focus in object instance recognition. In: 2016 International conference on image and vision computing New Zealand (IVCNZ), November 21, 2016: 1–5.
13. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems 2015: 91–99.
14. Lin TY, Maire M, Belongie S, et al. Microsoft coco: common objects in context. In European conference on computer vision, September 6, 2014: 740–755.
15. Bentohami A, de Korte N, Sosef N, Goslings JC, Bijlsma T, Schep N. Study protocol: non-displaced distal radial fractures in adult patients: three weeks vs. five weeks of cast immobilization: a randomized trial. BMC Musculoskelet Disord 2014;15(1):24.
16. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
17. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv [preprint]. https://arxiv.org/abs/1511.06348. Posted November 19, 2015. Updated January 7, 2016. Accessed June 22, 2018.
18. Xiao T, Xia T, Yang Y, Huang C, Wang X. Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2015: 2691–2699.
19. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition 2014: 580–587.
20. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI 2017 Feb 4 Vol. 4 p. 12. https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14806/14311. Accessed June 25, 2018.
21. Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol 2018;15(3,3 Pt B):504–508.
22. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. AJR Am J Roentgenol 2017;208(4):754–760.
23. Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 2006;28(4):594–611.
24. Long L, Wang W, Wen J, Zhang M, Lin Q, Ooi BC. Object-level representation learning for few-shot image classification. arXiv [preprint]. https://arxiv.org/abs/1805.10777 Posted May 28, 2018. Accessed June 22, 2018.