Deep neural network improves fracture detection by clinicians

Robert Lindsey[a,b,1], Aaron Daluiski[a,c], Sumit Chopra[a], Alexander Lachapelle[a,d], Michael Mozer[a,b], Serge Sicular[a,e], Douglas Hanel[a,f], Michael Gardner[a,g], Anurag Gupta[a,h], Robert Hotchkiss[a,c], and Hollis Potter[a,i]

[a]Imagen Technologies, New York, NY 10012; [b]Department of Computer Science, University of Colorado, Boulder, CO 80309; [c]Department of Orthopaedic Surgery, Hospital for Special Surgery, New York, NY 10021; [d]Faculty of Medicine, McGill University, Montreal, QC, Canada, H3A 2R7; [e]Department of Radiology, Mount Sinai Health System, New York, NY 10029; [f]Department of Orthopaedics and Sports Medicine, Harborview Medical Center, University of Washington, Seattle, WA 98104; [g]Department of Orthopaedic Surgery, Stanford University School of Medicine, Stanford, CA 94305; [h]Department of Emergency Medicine, Northwell Health, New Hyde Park, NY 11040; and [i]Department of Radiology and Imaging, Hospital for Special Surgery, New York, NY 10021

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved September 14, 2018 (received for review April 25, 2018)

Suspected fractures are among the most common reasons for patients to visit emergency departments (EDs), and X-ray imaging is the primary diagnostic tool used by clinicians to assess patients for fractures. Missing a fracture in a radiograph often has severe consequences for patients, resulting in delayed treatment and poor recovery of function. Nevertheless, radiographs in emergency settings are often read out of necessity by emergency medicine clinicians who lack subspecialized expertise in orthopedics, and misdiagnosed fractures account for upward of four of every five reported diagnostic errors in certain EDs. In this work, we developed a deep neural network to detect and localize fractures in radiographs. We trained it to accurately emulate the expertise of 18 senior subspecialized orthopedic surgeons by having them annotate 135,409 radiographs. We then ran a controlled experiment with emergency medicine clinicians to evaluate their ability to detect fractures in wrist radiographs with and without the assistance of the deep learning model. The average clinician's sensitivity was 80.8% (95% CI, 76.7–84.1%) unaided and 91.5% (95% CI, 89.3–92.9%) aided, and specificity was 87.5% (95 CI, 85.3–89.5%) unaided and 93.9% (95% CI, 92.9–94.9%) aided. The average clinician experienced a relative reduction in misinterpretation rate of 47.0% (95% CI, 37.4–53.9%). The significant improvements in diagnostic accuracy that we observed in this study show that deep learning methods are a mechanism by which senior medical specialists can deliver their expertise to generalists on the front lines of medicine, thereby providing substantial improvements to patient care.

deep learning | radiology | CAD | fractures | X-ray

Clinicians lack the subspecialized expertise and experience necessary to accurately identify fractures on radiographs, particularly in busy clinical settings where experienced radiologists or other practitioners may be unavailable. Clinicians may also be subject to excessive workloads, which cause fatigue and susceptibility to interpretational errors (1, 2). Radiographic interpretation often takes place in environments without qualified colleagues available for second opinions (2). Circumstances like those increase the risk of inaccurate identification of fractures on radiographs and often negatively impact patient care (3–6), especially in emergency departments, where missed fractures account for between 41 and 80% of reported diagnostic errors (5, 7, 8). These errors can have a devastating impact on subsequent function, resulting in malunion, osteonecrosis, and arthritis, all with attendant morbidity.

Computer-assisted detection (CAD) systems are a potential solution to this problem if they can quickly provide clinicians with a reliable second opinion, identifying regions of radiographs highly likely to contain pathology. However, the clinical use of CAD in medical imaging has produced mixed results. For instance, despite the use of CAD for the majority of mammography readings in the United States (9), several large prospective studies have indicated that mammography CAD actually decreases the specificity of radiologists without improving their sensitivity, resulting in an increased incidence of unnecessary diagnostic tests and biopsies with no improvement in cancer detection rates (9–12). A contributing factor to the ineffectiveness of early CAD systems is their underlying technology. Many early CAD algorithms functioned by identifying regions of an input image containing predefined texture patterns or geometric shapes, with the expectation that alerting clinicians to these visual features would be useful. Because of the limitations in the image analysis algorithms on which early CAD systems were based, they often would reliably mark pathological regions of images at the expense of overzealously identifying many nonpathological regions.

Recent advances in deep learning, a subfield of artificial intelligence, have allowed for the creation of computer models that can accurately solve many visual tasks involving object detection, localization, and classification (13). Within medical imaging, deep learning has shown immense initial promise at tasks, such as predicting the severity of diabetic retinopathy from retinal fundus images (14), classifying skin lesions (15), and analyzing histopathology (16, 17). Deep learning models differ from the technology used by early CAD systems in that they do not

Significance

Historically, computer-assisted detection (CAD) in radiology has failed to achieve improvements in diagnostic accuracy, decreasing clinician sensitivity and leading to unnecessary further diagnostic tests. With the advent of deep learning approaches to CAD, there is great excitement about its application to medicine, yet there is little evidence demonstrating improved diagnostic accuracy in clinically-relevant applications. We trained a deep learning model to detect fractures on radiographs with a diagnostic accuracy similar to that of senior subspecialized orthopedic surgeons. We demonstrate that when emergency medicine clinicians are provided with the assistance of the trained model, their ability to accurately detect fractures significantly improves.

Author contributions: R.L., A.D., A.L., M.M., and H.P. designed research; R.L., A.L., S.S., D.H., M.G., A.G., and R.H. performed research; R.L. and M.M. analyzed data; and R.L., A.D., S.C., M.M., and H.P. wrote the paper.

Conflict of interest statement: The authors are affiliated with Imagen Technologies, a startup company, the eventual products and services of which will be related to the subject matter of the article. The research was funded by Imagen Technologies. The authors own stock options in the company.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. Email: rob@imagen.ai.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1806905115/-/DCSupplemental.

Published online October 22, 2018.

COMPUTER SCIENCES

MEDICAL SCIENCES

rely on predefined representations of low-level visual features within images. Instead, they can learn to discover task-specific visual features that support making accurate clinical interpretations. Because the models learn by example, subspecialized experts can train models to detect fractures by carefully labeling them in large datasets of radiographs. This is a unique approach, centered on improving the diagnostic skills of clinicians and radiologists rather than replacing them by the use of an algorithm. With a sufficient supply of expertly labeled examples, an appropriately designed model can learn to emulate the judgments of those expert clinicians who provided the labels. In this work, we hypothesized that a deep learning model trained on a large dataset of high-quality labels would produce an automated fracture detector capable of emulating the diagnostic acumen of a team of experienced orthopedic surgeons. We further hypothesized that, when the model's output is provided to less experienced emergency medicine clinicians, their fracture detection sensitivity and specificity would be significantly improved.

Methods

Overview. For the purpose of model development, we retrospectively obtained a collection of radiographs from a specialty hospital in the United States. A group of senior subspecialized orthopedic surgeons provided clinical interpretations for each radiograph in the collection. The interpretations were provided through a web-based tool that allowed the surgeons to draw bounding boxes around every fracture that they could identify in the radiographs. We designed a deep learning model to detect and localize fractures in wrist radiographs and then trained it on a subset of the labeled dataset. We then clinically tested the trained model's performance on two test datasets: (*i*) a random subset of the development dataset's wrist radiographs that had been withheld from the model during training and validation and (*ii*) a separate dataset of all wrist radiographs obtained from the same hospital over a 3-mo period (which followed and did not overlap with the period in which the model's training data were acquired). To determine whether the trained model can help emergency medicine clinicians improve at fracture detection, we then ran a controlled experiment with emergency medicine clinicians, in which we evaluated each clinician's ability to detect fractures in wrist radiographs both with and without the availability of the model's output while making their interpretations.

Datasets. Radiographs acquired between September 2000 and March 2016 at the Hospital for Special Surgery (HSS) were used for this study. The radiographs were deidentified according to the Health Insurance Portability and Accountability Act Safe Harbor before being provided to the investigators. The dataset consisted of 135,845 radiographs of a variety of body parts. Of these, 34,990 radiographs were posterior–anterior or lateral wrist views. The remaining 100,855 radiographs belonged to 11 other body parts: foot, elbow, shoulder, knee, spine, femur, ankle, humerus, pelvis, hip, and tibia. The nonwrist body part with the maximum number of radiographs was shoulder with 26,042 images, and spine had the least number of radiographs with only 885 images.

Two datasets were used for clinical tests of the model. The first dataset (hereafter, "Test Set 1") consisted of 3,500 wrist radiographs, which were randomly withheld from the above wrist dataset of 34,990 radiographs. The second dataset (hereafter, "Test Set 2") consisted of 1,400 deidentified posterior–anterior- and lateral-view wrist radiographs from the HSS. These radiographs were consecutively sampled over a 3-mo period in 2016 (July to September) to ensure that the dataset was representative of a real-world clinical environment.

In total, 132,345 radiographs were used for model development (i.e., model training and validation), which consisted of all the original 135,845 radiographs provided by HSS except for those withheld for Test Set 1. Of these, the 100,855 radiographs corresponding to all of the body parts other than wrist (hereafter referred to as "Pretraining Set") were used for bootstrapping the model training process. The 31,490 radiographs of the wrist (hereafter referred to as "Wrist Training Set") were used during the training of the model.

Reference Standard. Ground truth labels were assigned to every radiograph in the datasets for the purpose of training the model, evaluating its accuracy, and then evaluating the accuracy of the emergency medicine clinicians with and without the model's assistance. Each ground truth assignment was made by one or more subspecialized orthopedic surgeons using an annotation software tool. The tool allowed the surgeons to label the presence and location of any fractures visible within each radiograph.

Development and Training of the Model. We used a deep convolutional neural network (DCNN) approach to fracture detection and localization. DCNNs are a type of nonlinear regression model: they are composite functions that transform their input variables (radiographs) into one or more outputs (pathology identified within radiographs). The equations specifying the input–output relationship have free parameters. Training of the model involves fitting these free parameters to a dataset of example input–output pairings (called the "training set"). The model that we developed for fracture detection poses the task as a simultaneous binary classification and conditional semantic segmentation problem, meaning that one of the outputs is a single probability value for a yes or no decision and another output is a dense conditional probability map, which we refer to as a heat map. The single probability value represents the model's confidence that the input

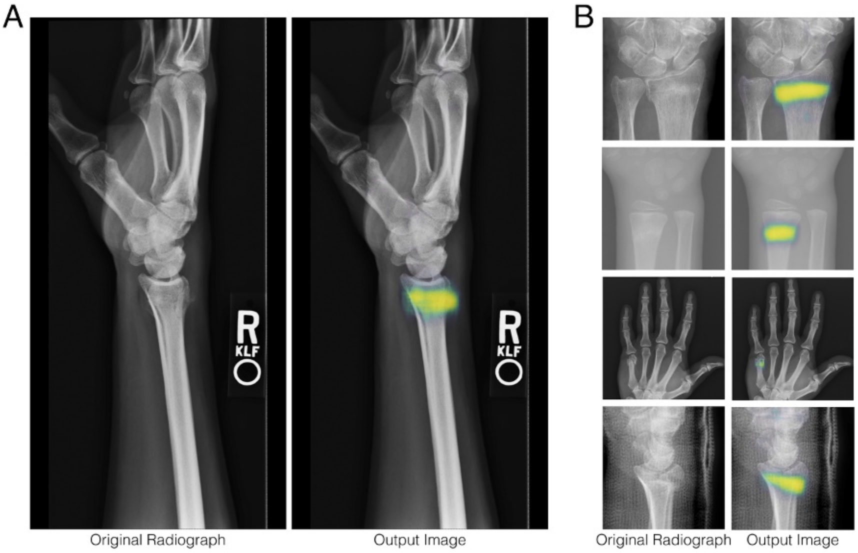

Fig. 1. *A*, *Left* shows a typical radiograph, which is provided as an input to the model. *A*, *Right* depicts a heat map overlaid on the radiograph. When the model determines that a fracture is present, the heat map represents the model's confidence that a particular location is part of the fracture, with yellow and blue being more and less confident, respectively. (*B*) Close-up views of four additional example inputs and heat map overlays.

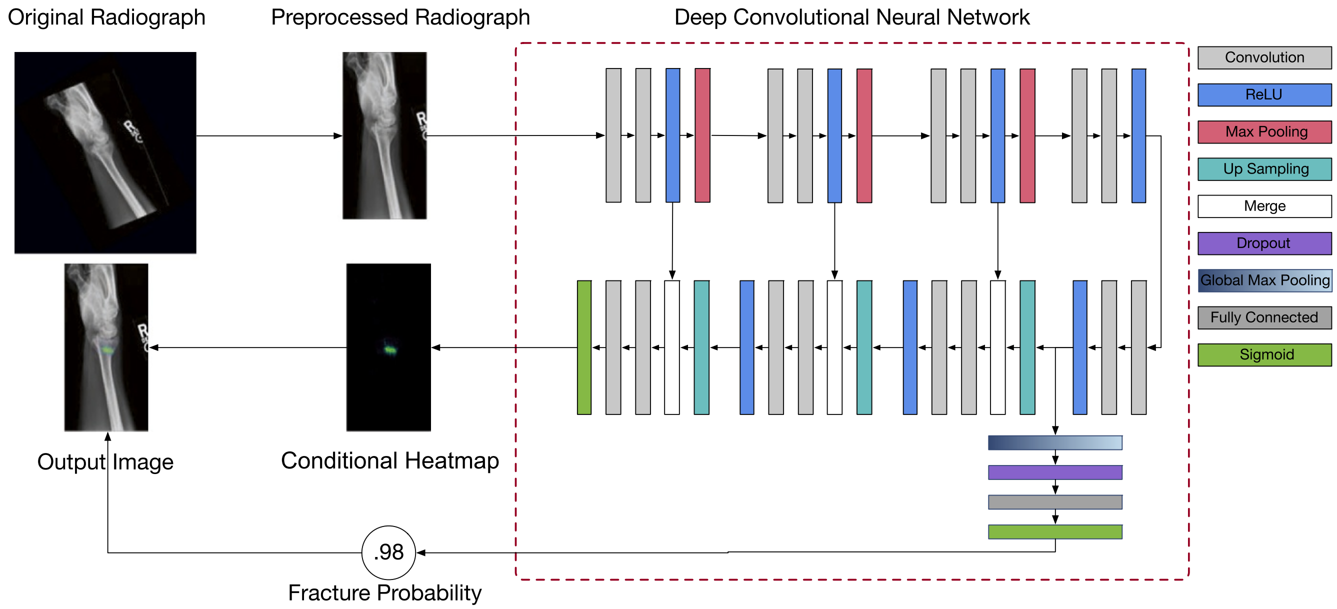

Fig. 2. A schematic of how radiographs are processed to detect and localize fractures. An input radiograph is first preprocessed by rotating, cropping, and applying an aspect ratio preserving rescaling operation to yield a fixed resolution of 1,024 × 512. The resulting image is then fed to a DCNN. The architecture of this DCNN is an extension of the U-Net architecture (18). The DCNN has two outputs: (*i*) the probability that the radiograph has a visible fracture any place in the image and (*ii*) conditioned on the presence of a fracture, a heat map indicating for each location in the image the probability that the fracture spans that location. When the probability of a fracture is high enough to render a clinical decision in favor of a fracture being present, the CAD system shows users the heat map overlaid on the preprocessed image. More information about the model design and training process can be found in *SI Appendix*.

radiograph contains a visible fracture. In the heat map, each pixel location represents the confidence that the corresponding pixel location in the input radiograph is part of a given fracture. This factored approach of having dual outputs allows us to disentangle the model's diagnostic decision-making capabilities (i.e., how often the model will make the correct diagnostic recommendation) and its localization ability (i.e., how precisely it can identify the location and extent of a fracture). Fig. 1*A* shows a typical input radiograph and the corresponding heat map overlaid on the radiograph. When the model produces a "yes" decision for the presence of a visible fracture, the overlay is shown to clinicians (Fig. 2).

The training of the model was done in two stages. The first stage involved the bootstrapping stage, in which the parameters of the model were randomly initialized and the model was trained on all of the 100,855 radiographs in the Pretraining Set (ankles, knees, spines, etc.). The goal of the bootstrapping stage was to better initialize the parameters of the model before starting to train it for wrist fracture detection, since better initialization of the parameters results in faster training and reduces overfitting. In the second stage, we took the model obtained from the first stage and used 31,490 wrist radiographs (Wrist Training Set) to specialize it to the task of detecting and localizing wrist fractures. The training of the model's parameters was accomplished by a variant of the standard stochastic gradient descent algorithm called Adam (19). Because the model had a large number of parameters and a relatively small number of labeled wrist radiographs to train with, we used a number of techniques to prevent overfitting, including early stopping and data augmentation. To perform early stopping, we split the Wrist Training Set into two disjoint subsets: 90% of the radiographs (28,341) were used to optimize the model parameters, and the remaining 10% (3,149) were used as an internal validation set. We stopped training the model parameters after the performance of the model on the validation set had not improved for five epochs. Data augmentation was performed and involved simulating having a larger labeled dataset by synthetically generating randomly altered versions of the radiographs on the fly during training. The alterations included random rotations, cropping, horizontal mirroring, and lighting and contrast adjustments. Performing data augmentation in this manner is intended to make the resulting model more robust to irrelevant sources of variability, including suboptimal positioning of patients within the radiograph and suboptimal exposure settings.

Evaluation of the Model. The trained model's ability to detect the presence of fractures in wrist radiographs was evaluated on the test datasets by calculating receiver operating characteristic (ROC) curves and measuring the area under the curve (AUC). AUC is a standard way to summarize an ROC curve, where an AUC of 1.0 would indicate that the CAD system perfectly predicts the reference standard and an AUC of 0.5 would indicate that the CAD system is no better than chance. An operating point for the model was fixed by choosing the threshold estimated to yield 95% sensitivity on the Wrist Training Set, and the resulting sensitivities and specificities on Test Sets 1 and 2 are also reported; 95% CIs for each statistic were estimated using bootstrap sampling via a bias-corrected and accelerated percentile method (20).

Evaluation of the Clinicians. We conducted an experiment to evaluate the utility of our trained model by measuring its effect on the diagnostic accuracy of a group of emergency medicine clinicians. The experiment followed a within-subjects design to evaluate the performance of a number of practicing emergency medicine clinicians on a sequence of 300 radiographs randomly chosen from Test Set 2, where the independent variable was whether or not the clinician could view the model's predictions when interpreting the radiograph. All the clinicians were shown the same set of 300 radiographs, although the order of the radiographs was randomized per clinician; 266 of 300 radiographs had no disagreements among the three clinicians used to define the reference standard about the presence or absence of a fracture. We recruited 40 practicing emergency medicine clinicians, of whom 16 were physician assistants (PAs) and 24 were medical doctors (MDs). Any clinician who had an across-condition sensitivity index (d' score) of 0 ± 0.05 was dropped from the analysis. This resulted in one PA being dropped for having an across-condition d' of 0.005. Ethics review and institutional review board exemption were obtained using the New England Institutional Review Board, informed consent was obtained from clinicians, and clinicians were deidentified during analyses.

For each radiograph shown, the clinicians were asked whether or not a fracture was present. After a clinician made a response, the model's semantic segmentation prediction was shown overlaid on the radiograph; the model's clinical determination was shown as text (either the statement "CAD Estimate: Fracture Present" or "CAD Estimate: Fracture Not Present"), and the clinician was asked the same question again. The clinical determination was produced from the model's probability estimate by thresholding at the predetermined operating point. An experimental advantage of this design is that every clinician interpreted every radiograph without and with the model's assistance (back to back), and therefore, the expertise of the clinicians and the diagnostic difficulties of the radiographs were balanced across conditions. A disadvantage of this design is that it did not counterbalance for condition order, meaning that the aided presentation of a radiograph always followed the unaided. This sequential reading methodology is standard for evaluating many CAD systems.

We report the sensitivity and specificity of clinicians for 266 radiographs on which there was no uncertainty about the reference standard. We also report the across-clinician average sensitivities and specificities stratified by the clinician's training (MD vs. PA), and we compare them against the model on the same imagery. Finally, we report an analysis of diagnostic accuracy as a function of the time that it took the clinicians to read the radiographs.

Results

ROC curves for the trained model on the two test sets are shown in Fig. 3. On Test Set 1, the model achieved an AUC of 0.967 (n = 3,500; 95% CI, 0.960–0.973). On Test Set 2, the model achieved an AUC of 0.975 (n = 1,400; 95% CI, 0.965–0.982). On the subset of images in Test Set 2 where there is no uncertainty about the reference standard (no interexpert disagreement), the model achieved an AUC of 0.994 (n = 1,243; 95% CI, 0.989–0.996). This indicates a very high level of agreement between the model's assessment of each radiograph and the senior subspecialized orthopedic hand surgeons who created the reference standard. Examples of the model's localizations are shown in Fig. 1. Qualitatively, the model is generally able to precisely identify the presence and location of visible fractures.

The sensitivity and specificity of the emergency medicine MDs were significantly improved with the assistance of the deep learning model (one-sided, two-sample Wilcoxon signed rank test for sensitivity: $P < 10^{-4}$, $d = 1.17$; specificity: $P < 10^{-5}$, $d = 1.24$) as was the sensitivity and specificity of the emergency medicine PAs (sensitivity: $P < 10^{-4}$, $d = 1.24$; specificity: $P < 10^{-4}$, $d = 1.19$). The average emergency medicine MD's sensitivities were 82.7% (95% CI, 78.1–86.6%) unaided and 92.5% (95% CI, 89.8–94.0%) aided, and specificities were 87.4% (95% CI, 84.5–89.9%) unaided and 94.1% (95% CI, 92.8–95.2%) aided. The average emergency medicine PA's sensitivities were 78.0% (95% CI, 71.5–83.7%) unaided and 89.9% (95% CI, 86.5–92.5%) aided, and specificities were 87.5% (95% CI, 84.4–90.3%) unaided and 93.6% (95% CI, 91.5–95.0%) aided. The average clinician experienced a relative reduction in misinterpretation rate of 47.0% (95% CI, 37.4–53.9%). Almost every clinician showed an improvement in both sensitivity and specificity. Additionally, the unaided accuracies observed are consistent with the limited literature for clinicians in controlled studies of radiographic wrist fracture detection (21) and with retrospective studies of fracture detection (2, 4, 22, 23). For comparison, on the same images, the model operated at 93.9% sensitivity (95% CI, 83.2–98.0%) and 94.5% specificity (95% CI, 90.6–97.2%) under its predetermined decision threshold and at 0.990 AUC (95% CI, 0.971–0.997) (Fig. 4).

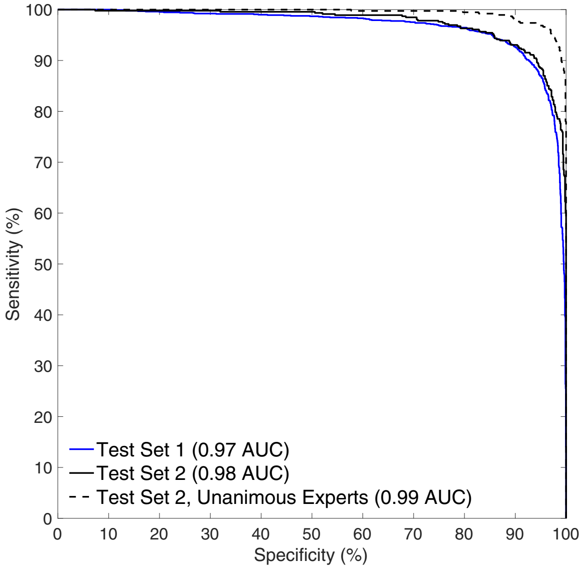

Fig. 3. The model accurately detects the presence of visible fractures in wrist radiographs on two separate test datasets. When given a radiograph, one of the model's outputs is a probability that the patient has a fracture visible in the radiograph. A decision threshold t has to be chosen such that, for any probability value greater than the threshold, the CAD system alerts the clinician. The above curves show, for all possible values of $t \in [0, 1]$, what the corresponding sensitivity (true positive rate) and specificity (true negative rate) of the system would be on that test dataset. The dashed black line restricts the analysis to the subset of Test Set 2, on which there was no interexpert disagreement about the presence or absence of a visible fracture (1,243 of 1,400 radiographs).

Fig. 5 shows the relationship between reading time and diagnostic accuracy in the aided and unaided conditions. Radiographs that took little time on average to read unaided were generally read accurately. The longer that it took to read a radiograph, generally the worse the diagnostic accuracy became for both conditions. However, the difference in accuracy between the aided and unaided reading conditions increased with the unaided reading time. This suggests that emergency medicine workers with relatively hard, time-consuming caseloads would benefit more from the CAD software.

Discussion

This study showed that a deep learning model can be trained to detect wrist fractures in radiographs with diagnostic accuracy similar to that of senior subspecialized orthopedic surgeons. Additionally, this study showed that, when emergency medicine clinicians are provided with the assistance of the trained model, their ability to detect wrist fractures can be significantly improved, thus diminishing diagnostic errors and also improving the clinicians' efficiency.

Misinterpretation of radiographs may have grave consequences, resulting in complications including malunion with restricted range of motion, posttraumatic osteoarthritis, and joint collapse, the latter of which may require joint replacement. Misdiagnoses are also the primary cause of malpractice claims or litigation (1, 3–6). There are multiple factors that can contribute to radiographic misinterpretations of fractures by clinicians, including physician fatigue, lack of subspecialized expertise, and inconsistency among reading physicians (2, 4, 5, 24). The approach of this investigation is to apply machine learning algorithms trained by experts in the field to less experienced clinicians (who are at particular risk for diagnostic errors yet responsible for primary patient care and triage) to improve both their performance and efficiency. The learning model presented in this study mitigates these factors. It does not become fatigued, it always provides a consistent read, and it gains subspecialized expertise by being provided with labeled radiographs from human experts. The experiments described in this paper showed that the proposed model can be used to assist practicing clinicians and help improve their performance in identifying fractures in radiographs. Every measure that we used to characterize clinician performance showed a statistically significant improvement in clinician performance with a large effect size. Notably, the misinterpretation rate of the practicing emergency medicine clinicians was reduced by approximately one-half through the assistance of the model.

A common criticism of CAD systems in oncology is that they increase sensitivity at the expense of lowering specificity, which results in unnecessary procedures and increased costs. Importantly, the increased sensitivity observed in this study did not come at the expense of a lower specificity. This is likely

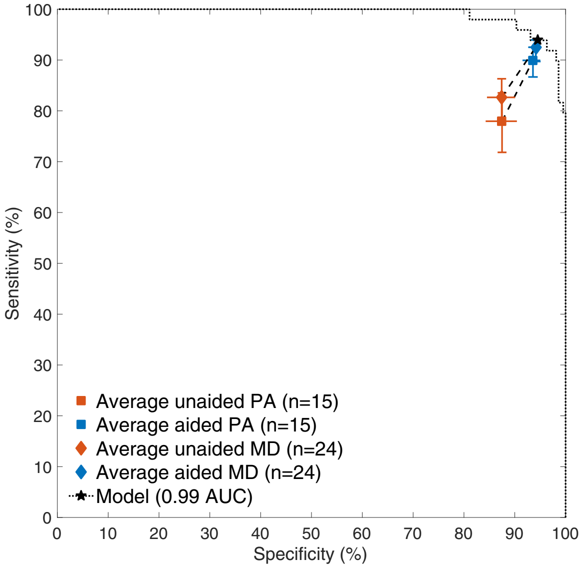

Fig. 4. Performance of the emergency medicine clinicians in the experiment. Each clinician read each radiograph first unaided (without the assistance of the model) and then aided (with the assistance of the model). The average clinician's sensitivities were 80.8% (95% CI, 76.7–84.1%) unaided and 91.5% (95% CI, 89.3–92.9%) aided, and specificities were 87.5% (95% CI, 85.3–89.5%) unaided and 93.9% (95% CI, 92.9–94.9%) aided. The model operated at 93.9% sensitivity and 94.5% specificity (shown as the star) using a decision threshold set on the model development dataset.

attributable to the high standalone diagnostic accuracy of the model. To further mitigate the specificity–sensitivity tradeoff, a soft highlighting visualization method has recently been proposed (25). This method modulates the heat map according to the model's probability that a fracture is present somewhere in the image instead of making a hard present/absent decision. This method could easily be incorporated into our paradigm.

A differentiator of the modeling and experimental approach presented in this paper is the reliance on senior subspecialized experts to provide the ground truth labels. The purpose of building CAD software is to use it to improve the diagnostic accuracy of practicing clinicians, not simply to report the highest AUC possible in the CAD system's underlying model. A model trained on labels provided by less experienced clinicians might be able to achieve a very high AUC when evaluated on those clinicians' labels (i.e., it might accurately predict what clinical interpretations the less experienced clinicians would provide), but its clinical expertise and relevance would be limited based on the expertise of the labeling clinicians. Whatever subtle findings that the inexperienced clinicians would have systematically missed, the model would have learned to miss too, and yet, these misses would not be reported as errors. Hence, we argue that it is important to train and evaluate deep learning models on datasets that have ground truth labels worth emulating, ones in which the hardest cases have not been systematically mislabeled.

One of the keys to the success of our proposed approach is the establishment of the rigorous "ground truth" for the presence and location of a fracture. We used the expertise of multiple orthopedic hand surgeons with many years of experience to establish the ground truth with which the model was trained. An alternative approach might be to take large numbers of radiographic reports from busy trauma centers and extract the ground truth from those reports. Within such reports, however, there exists a distribution of accuracy and expertise. On the lower end of the curve are the junior radiology trainees interpreting emergency department radiographs under the tutelage of attending radiologists who are not specialized in orthopedics but rather, contribute to the interpretation of emergency radiographs based on a rotating assignment. On the upper end of the performance curve are the fellowship-trained musculoskeletal radiologists who are experts in the area. One could argue then that the "noise" of a less accurate standard is provided to such algorithms, which are based on simply the input of radiographic reports. Based on our results, we speculate that our model would have lower noise based on a more rigorously established ground truth.

There are several limitations of this study. First, the experiment was a retrospective evaluation conducted through a web interface resembling a Picture Archiving and Communication System (PACS) used by clinicians for medical imaging. A prospective study in a real-world clinical environment with a real PACS system would need to be conducted to know the exact unaided and aided accuracies of practicing emergency medicine clinicians. Second, the diagnostic accuracy of clinicians and the model in this study is limited to determining what is visible within a radiograph. A more clinically valid study end point would instead be based on the clinician's overall assessment of the patient, taking into account information available outside the one radiograph (e.g., other radiographs, physical examination). Third, performance in the aided condition was driven not only by the diagnostic accuracy of the deep learning model but also by the way in which the model's output was displayed (i.e., as a green heat map and a text recommendation). The experiment's design does not allow for these factors to be separated during analysis; however, the fact that the average aided clinician's performance is slightly worse than the model's standalone performance suggests that the way that the model's output was presented was suboptimal.

As a proof of concept, we focused our evaluations on wrist fractures, but the models are not limited to learning from wrist radiographs. Given enough training data and a suitably designed model, they can in principle be taught to detect any condition on radiographs that a human clinician could identify. This study shows that deep learning models offer potential for subspecialized clinicians (without machine learning experience) to teach computers how to emulate their diagnostic expertise and thereby help patients on a global scale. Although teaching the model is a laborious process requiring collecting thousands of radiographs and carefully labeling them, making a prediction using the trained model takes less than a second on a modern computer.

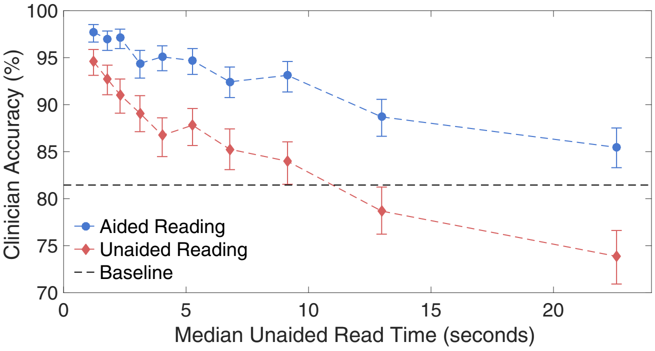

Fig. 5. Each point represents a bin containing one-10th of the radiographs used in the experiment. The horizontal location of a point indicates the median unaided response time in seconds for the radiographs within the bin. The vertical location of a point indicates the across-clinician average diagnostic accuracy on the radiographs within the bin. The difference in accuracy between the aided and unaided reading conditions increases with unaided reading time, which is a proxy for the radiograph's difficulty. The dashed horizontal black line indicates the accuracy that a clinician would have achieved had he or she reported "no fracture" on every radiograph. The aided reading condition never has an average accuracy worse than baseline guessing.

Thus, we speculate that, someday, technology may permit any patient whose clinician has computer access to receive the same high-quality radiographic interpretations as those received by the patients of senior subspecialized experts.

ACKNOWLEDGMENTS. We thank Andrew Sama for helping to run the study, Thomas Kho for programming the software for the human experiment, and Sophia Paul for helping with reader recruitment. The project is funded by Imagen Technologies. The work presented in this manuscript is for research purposes only and is not for sale within the United States.

1. Berlin L (2001) Defending the "missed" radiographic diagnosis. *Am J Roentgenol* 176:317–322.
2. Hallas P, Ellingsen T (2006) Errors in fracture diagnoses in the emergency department: Characteristics of patients and diurnal variation. *BMC Emerg Med* 6:4.
3. Kachalia A, et al. (2007) Missed and delayed diagnoses in the emergency department: A study of closed malpractice claims from 4 liability insurers. *Ann Emerg Med* 49:196–205.
4. Wei CJ, et al. (2006) Systematic analysis of missed extremity fractures in emergency radiology. *Acta Radiologica* 47:710–717.
5. Guly HR (2001) Diagnostic errors in an accident and emergency department. *Emerg Med J* 18:263–269.
6. Whang JS, Baker SR, Patel R, Luk L, Castro A (2013) The causes of medical malpractice suits against radiologists in the United States. *Radiology* 266:548–554.
7. Williams SM, Connelly DJ, Wadsworth S, Wilson DJ (2000) Radiological review of accident and emergency radiographs: A 1-year audit. *Clin Radiol* 55:861–865.
8. Leeper WR, et al. (2013) The role of trauma team leaders in missed injuries: Does specialty matter? *J Trauma Acute Care Surg* 75:387–390.
9. Lehman C, et al. (2015) Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 175:1828–1837.
10. Taylor P, Potts HW (2008) Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. *Eur J Cancer* 44:798–807.
11. Khoo LA, Taylor P, Given-Wilson RM (2005) Computer-aided detection in the United Kingdom national breast screening programme: Prospective study. *Radiology* 237:444–449.
12. Azavedo E, Zackrisson S, Mejàre I, Heibert Arnlind M (2012) Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review. *BMC Med Imaging* 12:22.
13. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
14. Gulshan V, et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J Am Med Assoc* 304:649–656.
15. Esteva A, et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:151–118.
16. Sirinukunwattana K, et al. (2016) Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans Med Imaging* 35:1196–1206.
17. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J (2013) Mitosis detection in breast cancer histology images with deep neural networks. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Lecture Notes in Computer Science, eds Mori K, Sakuma I, Sato Y, Barillot C, Navab N (Springer, Berlin), vol 8150, pp 411–418.
18. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, eds Navab N, Hornegger J, Wells W, Frangi A (Springer, Cham, Germany), vol 9351, pp 234–241.
19. Kingma DP, Ba JL (2015) Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations 2015*. arXiv:1412.6980v9. Preprint, posted December 22, 2014.
20. DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. *Stat Sci* 11:189–228.
21. Doyle AJ, Le Fevre J, Anderson GD (2005) Personal computer versus workstation display: Observer performance in detection of wrist fractures on digital radiographs. *Radiology* 237:872–877.
22. Espinosa JA, Nolan TW (2000) Reducing errors made by emergency physicians in interpreting radiographs: Longitudinal study. *Br Med J* 320:737–740.
23. Lufkin KC, Smith SW, Matticks CA, Brunette DD (1998) Radiologists' review of radiographs interpreted confidently by emergency physicians infrequently leads to changes in patient management. *Ann Emerg Med* 31:202–207.
24. Juhl M, Møller-Madsen B, Jensen J (1990) Missed injuries in an orthopaedic department. *Injury* 21:110–112.
25. Kneusel RT, Mozer MC (2017) Improving human-machine cooperative visual search with soft highlighting. *ACM Trans Appl Percept* 15:1–21.