Skeletal Radiology (2022) 51:345–353
https://doi.org/10.1007/s00256-021-03739-2

SCIENTIFIC ARTICLE

Deep learning detection of subtle fractures using staged algorithms to mimic radiologist search pattern

Mark Ren[1] · Paul H. Yi[1,2,3]

Received: 18 November 2020 / Revised: 25 January 2021 / Accepted: 7 February 2021 / Published online: 12 February 2021

Abstract

Objective To develop and evaluate a two-stage deep convolutional neural network system that mimics a radiologist's search pattern for detecting two small fractures: triquetral avulsion fractures and Segond fractures.

Materials and methods We obtained 231 lateral wrist radiographs and 173 anteroposterior knee radiographs from the Stanford MURA and LERA datasets and the public domain to train and validate a two-stage deep convolutional neural network system: (1) object detectors that crop the dorsal triquetrum or lateral tibial condyle, trained on control images, followed by (2) classifiers for triquetral and Segond fractures, trained on a 1:1 case:control split. A second set of classifiers was trained on uncropped images for comparison. External test sets of 50 lateral wrist radiographs and 24 anteroposterior knee radiographs were used to evaluate generalizability. Gradient-class activation mapping was used to inspect image regions of greater importance in deciding the final classification.

Results The object detectors accurately cropped the regions of interest in all validation and test images. The two-stage system achieved cross-validated area under the receiver operating characteristic curve values of 0.959 and 0.989 on triquetral and Segond fractures, compared with 0.860 ($p = 0.0086$) and 0.909 ($p = 0.0074$), respectively, for a one-stage classifier. Two-stage cross-validation accuracies were 90.8% and 92.5% for triquetral and Segond fractures, respectively.

Conclusion A two-stage pipeline increases accuracy in the detection of subtle fractures on radiographs compared with a one-stage classifier and generalized well to external test data. Focusing attention on specific image regions appears to improve detection of subtle findings that may otherwise be missed.

Keywords Artificial intelligence · Machine learning · Neural network · Convolutional neural network · Deep convolutional neural network · Triquetral fracture · Segond fracture · Fracture detection · Fracture

Introduction

Over the last decade, deep convolutional neural networks (DCNNs) have significantly advanced computer vision. DCNNs can be trained on a large set of pre-labeled images, learn discerning features by which to discriminate between image classes, and classify new data much faster than human counterparts [1]. This technique has been applied to various medical imaging tasks, including the automated detection of radiological findings in cancer and tuberculosis [2–4], as well as musculoskeletal imaging tasks, such as automated fracture detection, with studies demonstrating performance and reliability comparable to radiologists and orthopedic surgeons [5–8].

Although prior studies have shown promise for automated fracture detection, previously described DCNNs have generally been trained to detect and classify relatively large and/or obvious abnormalities, such as distal radius fractures [9, 10]. Identifying more subtle imaging findings, such as small avulsion fractures, is a more difficult task. To maximize sensitivity for subtle radiological findings, radiologists learn to adopt specific search patterns for a given type of image and clinical scenario to better identify these findings, which are often small relative to the rest of the image [11–13]. For example, in the setting of acute trauma to an extremity, radiologists will often

✉ Paul H. Yi
pyi10@jhmi.edu

[1] The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, MD Baltimore, USA

[2] University of Maryland Intelligent Imaging Center, Department of Radiology, University of Maryland School of Medicine, MD Baltimore, USA

[3] Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA

zoom into specific regions of an image to assess for abnormalities that may not be readily visible in the full-size image, such as a triquetral avulsion fracture. Focusing on a specific part of a radiograph increases the relative size of the abnormality (in relation to the image encompassed in the observer's visual field) and allows the observer to better evaluate for specific findings by reducing surrounding distractors or "noise."

A similar approach to increase the "signal" present within a given image from a subtle abnormality by zooming in on the area of interest would logically help a DCNN learn to identify these subtle abnormalities. When evaluating images cropped to the humeral head, DCNNs have demonstrated fracture detection performance comparable with or better than physicians [14]. However, in this study, cropping was done manually, which precludes the primary DCNN benefits of speed and efficiency in detecting abnormalities. We sought to mimic the radiologist search pattern by developing a two-stage deep convolutional neural network system that "zooms in" to parts of an image and evaluates those specific regions for two small fractures: triquetral avulsion fractures and Segond fractures. We hypothesize that, by increasing the signal-to-noise ratio of the abnormality, this method can improve DCNN performance in detecting subtle radiological findings.

Methods

All images used to train our DCNNs and to validate their performance during training were either obtained from the MURA [5] and LERA [15] datasets or are in the public domain and were found using the Google internet search engine (http://www.google.com) and Radiopaedia (http://radiopaedia.org). Images were saved as Portable Network Graphics files in their original available resolution. External testing data was obtained from trauma radiographs obtained at our institution. All images used in this study were deidentified and compliant with the Health Insurance Portability and Accountability Act (HIPAA). In accordance with 45 CFR 46.102(f), our institutional review board classified this study as non-human subject research, because all images were anonymized before the time of the study.

Datasets

Object detection

To train our DCNN to localize the region containing the dorsal triquetrum, we used 200 lateral wrist radiographs obtained from the MURA dataset and 82 images obtained elsewhere from the public domain, for a total of 282 lateral wrist radiographs which did not have a triquetral fracture. To train a separate DCNN to localize the lateral tibial condyle, we used 71 images from the LERA dataset and 49 images obtained elsewhere from the public domain, for a total of 120 AP knee radiographs which did not have a Segond fracture. Raw images were cropped to exclude extraneous information (e.g., borders) or to separate bilateral radiographs into two unilateral images where appropriate. Each image was manually annotated by a member of the research team and verified by a radiologist with over 6 years of experience in reviewing musculoskeletal radiographs.

Classification

To train our DCNNs to detect triquetral fractures, we used 106 lateral wrist images (53 triquetral fracture, 53 control). To train our DCNNs to detect Segond fractures, we used 102 AP knee images (51 Segond fracture, 51 control). Fracture images were selected from the public domain, and control images were selected from the public domain and the MURA and LERA datasets. Each of the four datasets was normalized to have the same mean and standard deviation pixel values on each RGB channel for each image within the dataset.

To externally test the results of our two-stage system and compare them with a one-stage classifier, we also curated test sets from our tertiary care center comprised of 50 radiographs (25 triquetral fracture, 25 no fracture) for triquetral fractures and 24 radiographs (12 Segond fracture, 12 no fracture) for Segond fractures. These images were used solely for testing and were not presented to the DCNNs during the training or validation phases.

Deep learning system development

Two-stage algorithm approach

We designed a two-stage algorithm approach to identifying small avulsion fractures that mimics a radiologist search pattern of first zooming in on the area of interest followed by identification of the presence or absence of a fracture in that area through use of sequential object detector DCNN and image classifier DCNNs (Fig. 1).

We trained our object detector and classifier DCNNs using a transfer learning approach on DCNNs previously trained on large non-radiographic image datasets. The structure and function of DCNNs have been extensively described; in brief, by being passed iterations of a large, pre-annotated dataset, DCNNs are designed to recognize patterns and features that enable them to classify or localize features in novel data inputs [16]. Transfer learning is an approach in which a DCNN trained for many iterations on a large dataset can be adapted for a new task by fine-tuning on a smaller dataset. This technique has successfully been applied to commonly used image detection models for use on medical imaging [9].

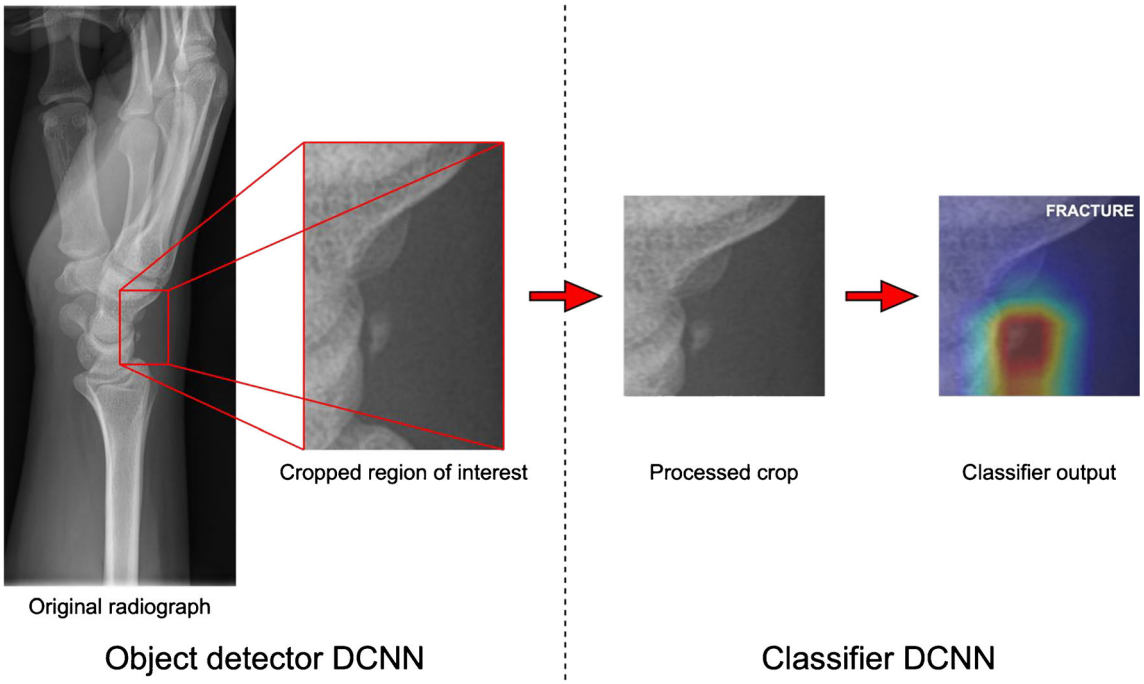

Fig. 1 Schematic of our two-stage system for avulsion fracture detection/classification. A full-size radiograph is inputted into an object detector DCNN, which identifies the area of interest and automatically crops it. This crop is then inputted into the classifier DCNN, which processes the image and determines if it contains the specified abnormality (a triquetral avulsion fracture in this schematic)

Localization (1st stage of the system)

Lateral wrist and AP knee datasets were randomly divided into training and validation sets, with an 85%:15% training:validation split. No images overlapped between the training and validation sets. Both training sets were augmented using standard techniques applied randomly, including random rotation, random flipping, and affine transformations, for a total of 10× and 16× augmentation for lateral wrist and AP knee x-rays, respectively. No additional pre-processing was performed prior to feeding images to the DCNNs, which automatically resize inputs to have lengths and widths between 800 and 1333 pixels. Transfer learning was performed using the RetinaNet object detection DCNN based on a ResNet-50 backbone and pretrained on the MS COCO dataset [17]. The parameters used for our DCNN training were 10 epochs and Adam with a learning rate of 1×10^{-5} and gradient norm clipping of 0.001. Training loss and validation mAP were monitored for hyperparameter optimization. To evaluate "correctness" of localization on test images, an observer with over 6 years of experience evaluating musculoskeletal radiographs determined if the bounding box included the relevant anatomic area of interest.

Classification (2nd stage of the system)

Lateral wrist and AP knee datasets were inputted into their respective object detector DCNNs. Copies of each image were cropped to the predicted location of the triquetrum and lateral tibial condyle, respectively, and they were used to train classifier DCNNs. Each classifier was trained using 5-fold cross-validation, resulting in each dataset (uncropped wrist, cropped wrist, uncropped knee, cropped knee) divided into training and validation sets, with an 80%/20% training/validation split and no overlap between splits. Each training set was augmented as described for the localization datasets. No additional pre-processing was performed prior to feeding images to the DCNNs, which automatically resize inputs to 224×224 pixels. Transfer learning was performed using the ResNet-50 classification DCNN, which was pretrained on the ImageNet dataset [18]. The parameters used for our DCNN training were 16 epochs and stochastic gradient descent with an initial learning rate of 0.001, momentum of 0.9, and learning rate decay of 0.1 with a step size of 6. After each training epoch, each DCNN was evaluated against its validation set in order to select the highest-performing model (as determined by accuracy) and discourage overfitting. Training and validation loss were monitored for hyperparameter optimization. The DCNNs performing highest on 5-fold cross-validation accuracy were also tested on the external test set.

Comparison of the staged system for fracture detection to non-staged system

To compare the performance of our two-stage system with a one-stage (standard) method, we trained two additional

classifier DCNNs, using the uncropped lateral wrist and AP knee images for classification of triquetral and Segond fractures, respectively. The one-stage classifiers were trained and evaluated using the same methods as the classifier of the two-stage system. We then tested these DCNNs on the same external test set as the staged DCNN system.

DCNN decision-making visualization

In order to visualize and compare the parts of each image used by the DCNNs to classify images as fracture or no fracture, we used a technique called gradient-weighted class activation mapping (Grad-CAM) to create heatmaps showing the regions of each image important for DCNN decision-making. Grad-CAM utilizes the values flowing into the last layers of a DCNN to compute each region of an image's relative importance in making the final classification [19]. The relative importances can then converted into a colorized heatmap; in our Grad-CAM color scheme, red indicates the greatest effect on the classification decision and blue indicates the least. For each test image, an observer with over 6 years of experience evaluating musculoskeletal radiographs assessed the Grad-CAM outputs to determine if the DCNNs learned to discriminate fractures as opposed to confounding features in the image.

All image processing and DCNN development was performed via remote connection on servers using NVIDIA K80 GPUs. Image augmentation was performed with the *albumentations* package (version 0.4.6), DCNN development for object detection/localization was done with the *keras-retinanet* package (version 0.4.1), and Grad-CAM heatmaps were generated with the *pytorch-gradcam* package (https://github.com; version 0.2.1). Development for classification/fracture detection was done using the PyTorch deep learning framework (version 1.6.0, https://pytorch.org).

Statistical analysis

All statistical analyses were performed using R version 3.4 with the pROC package (http://cran.r-project.org). Object detector performance was evaluated using mean average precision (mAP) and the number of images whose region of interest was correctly localized. The performance of each classification model on validation and test datasets was described using receiver operating characteristic (ROC) curve, and summarized using the area under the ROC curve (AUC) or cross-validated area under the ROC curve (cvAUC), calculated as the mean of each fold AUC. Optimal threshold points were determined using Youden's J-statistic to calculate sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy, which are derived from means of each fold value for cross-validation results. The DeLong non-parametric method was used to compare DCNN performance.

Results

Localization results (1st stage of the system)

The localizer/object detection DCNNs achieved a validation mAP of 0.970 for the dorsal triquetrum and 0.934 for the lateral proximal tibia. Each localizer DCNN correctly found and cropped the region of interest for 100% of test images for both triquetral and Segond fractures (Fig. 2).

Classification results (2nd stage of the system)

On fivefold cross-validation, the classifier DCNN for triquetral fractures in our two-stage system achieved a cvAUC of 0.954 (standard deviation across folds [SD_{AUC}]

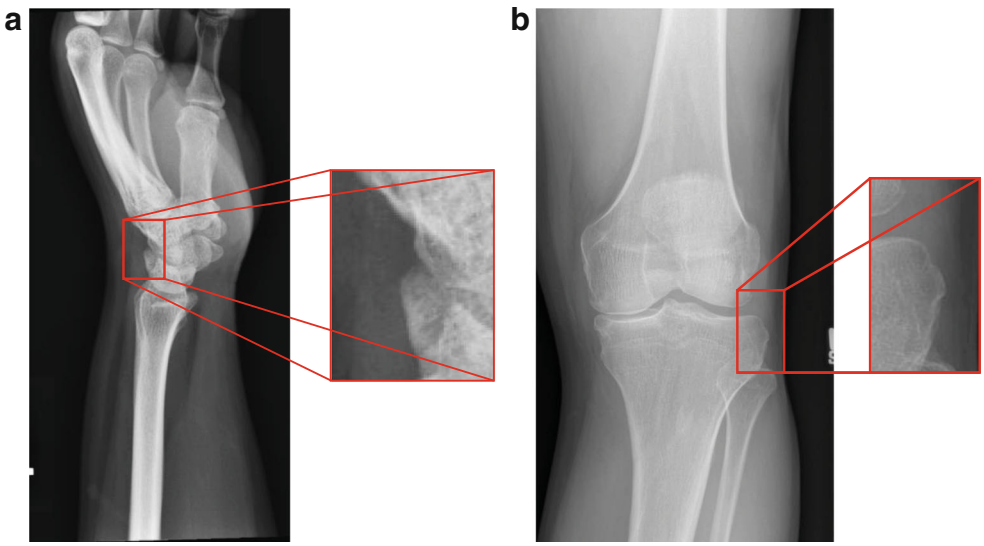

Fig. 2 Representative output of object detector DCNN on **a** lateral wrist and **b** AP knee radiographs

Table 1 Performance measures of two-stage system and single-stage classifier for triquetral and Segond fractures on cross-validation

	cvAUC	SD_{AUC}	Sensitivity	Specificity	PPV	NPV	Accuracy
Triquetral fracture (n=106)							
2-stage	0.954	0.038	91.2%	90.4%	90.5%	91.1%	90.8%
1-stage	0.870	0.067	76.0%	79.2%	78.5%	76.7%	77.6%
p value	0.0086*						
Segond fracture (n=102)							
2-stage	0.986	0.014	91.7%	93.3%	93.2%	91.8%	92.5%
1-stage	0.909	0.067	61.7%	70.0%	67.3%	64.6%	65.8%
p value	0.0074*						

Threshold values were chosen using Youden's J-statistic

SD_{AUC} is the standard deviation of AUC values across cross-validation folds

= 0.038), with sensitivity and specificity of 0.912 and 0.904, respectively. The overall accuracy in triquetral fracture detection was 90.8%. The classifier DCNN for Segond fractures achieved a cvAUC of 0.986 (SD_{AUC} = 0.067) with optimal sensitivity of 0.917, specificity of 0.933, and accuracy of 92.5% (Table 1).

Two-stage system evaluation and comparison to one-stage classifier

On the external test set, our two-stage system pipeline achieved an AUC of 0.952 for triquetral fracture detection and an AUC of 0.965 for Segond fracture detection (Table 2).

The two-stage system demonstrated statistically significantly greater performance on cross-validation for triquetral fracture detection (cvAUC = 0.954, 95% confidence interval [95% CI]: 0.923–0.995) compared with a one-stage classifier trained using the same methods (cvAUC = 0.870, 95% CI: 0.791–0.930, p = 0.0086). The two-stage system also achieved higher performance in detecting Segond fractures (cvAUC = 0.986, 95% CI: 0.977–1) than a one-stage classifier (cvAUC = 0.909, 95% CI: 0.853–0.965, p = 0.0074) (Table 1).

Table 2 Performance measures of two-stage system and single-stage classifier for triquetral and Segond fractures on external test set

	AUC	Sensitivity	Specificity	PPV	NPV	Accuracy
Triquetral fracture (n=50)						
2-stage	0.952	96.0%	88.0%	88.9%	95.7%	92.0%
1-stage	0.899	96.0%	64.0%	72.7%	94.1%	80.0%
p value	0.32					
Segond fracture (n=24)						
2-stage	0.965	91.7%	91.7%	91.7%	91.7%	91.7%
1-stage	0.660	58.3%	83.3%	77.8%	66.7%	70.8%
p value	0.0028*					

Threshold values were chosen using Youden's J-statistic

On the external test sets, the two-stage AUC (0.952) was greater than the one-stage AUC (0.899) for triquetral fracture detection; however, this difference was not statistically significant (p = 0.32). The two-stage AUC (0.965) was significantly greater than the one-stage AUC (0.660) for Segond fracture detection on the external test set (p = 0.0028) (Tables 2 and 3).

Grad-CAM heatmaps of the cropped triquetral fracture classifier showed that the DCNNs identified the dorsal part of the triquetrum from which bone fragments are typically avulsed in fractures, as well as the fragments themselves. For the uncropped, one-stage classifier, heatmaps indicated that the DCNN emphasized the region of the wrist in general,

Table 3 Confusion matrices of two-stage system and single-stage classifier for triquetral and Segond fractures on external test set

	Actual		
Triquetrum two-stage	Fracture	No fracture	Total
Fracture	24	3	27
No fracture	1	22	23
Total	25	25	50
	Actual		
Triquetrum one-stage	Fracture	No fracture	Total
Fracture	24	9	33
No fracture	1	16	17
Total	25	25	50
	Actual		
Segond two-stage	Fracture	No fracture	Total
Fracture	11	1	12
No fracture	1	11	12
Total	12	12	24
	Actual		
Segond one-stage	Fracture	No fracture	Total
Fracture	7	2	9
No Fracture	5	10	15
Total	12	12	24

without specifically considering the region containing the triquetrum (Fig. 3). Heatmaps of the cropped Segond classifier also showed that the DCNNs focused specifically on fracture fragments and the proximal lateral region of the tibia from which fragments are avulsed, while those of the uncropped Segond classifier demonstrated general activation at the tibiofemoral joint, often not including lateral proximal region of the tibia where Segond fractures occur (Fig. 3).

Discussion

Subtle avulsion fractures can be difficult to identify, even for trained eyes, due to the small relative "signal" present in the midst of a relatively large radiograph. Accordingly, radiologists frequently zoom in on anatomic areas of interest to search for these fractures, thereby increasing their "signal:noise" ratio. Based on this observation, we developed a two-stage system utilizing a localizer/object detection DCNN, which crops radiographs to the anatomic region of interest, followed by a classifier DCNN to identify subtle avulsion fractures. Using a small training dataset, our two-stage system outperformed a single-stage fracture detection DCNN in detecting both triquetral fractures and Segond fractures, achieving higher diagnostic performance during five-fold cross-validation. Using Grad-CAM heatmaps, we found that the cropped classifier DCNN was able to emphasize and make fracture detection decisions based on the specific regions and radiological findings characterizing the fractures of interest, whereas one-stage classifiers did not focus precisely on the fracture site, highlighting the importance of not only evaluating DCNNs based on diagnostic performance measures but also on the specific areas of an image that weigh into DCNN decision-making.

In clinical practice, radiologists have the advantage of integrating clinical information to guide their differential diagnosis and search patterns, allowing them to focus on looking for small findings such as triquetral and Segond fractures. While machine learning algorithms cannot understand a patient's overall clinical picture like a human physician, systematically replicating a focused search pattern may improve sensitivity for small radiologic findings, with the added advantages of speed, consistency, and lack of fatigability of automated algorithms. Therefore, the first task that we sought to perform using DCNNs was identifying focused anatomic regions of interest. After being trained on a small set of images, our object detection DCNNs demonstrated perfect accuracy in identifying and cropping their respective regions of interest, which mimics the radiologist's search pattern of zooming in on relevant anatomic areas of interest to identify subtle findings like avulsion fractures. Our DCNNs achieving high accuracy at finding specific regions are consistent with prior

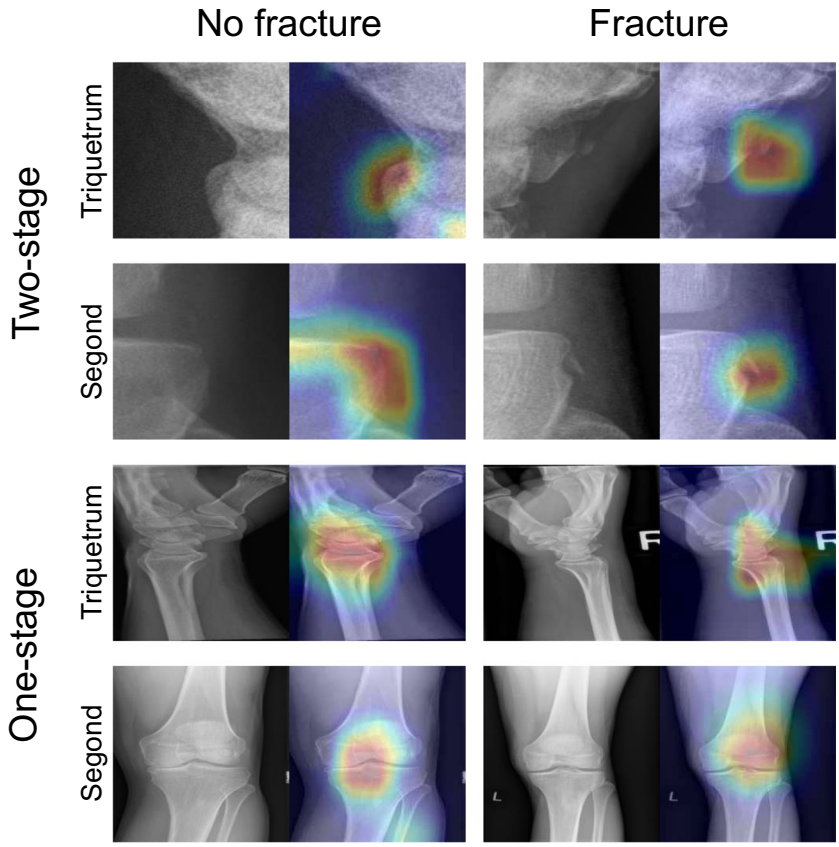

Fig. 3 Representative processed radiograph and superimposed Grad-CAM heatmap from the two-stage and one-stage DCNN classifiers for triquetral and Segond fractures. Images shown from the two-stage system were automatically cropped by the respective object detector DCNN from the first stage. Red color indicates greater weight in assigning the final classification

studies that have shown DCNNs to have excellent results in anatomic region detection and segmentation tasks on radiographs [20, 2122]. Importantly, because accurate identification of avulsion fractures is predicated on focusing in on the correct anatomic area, having an accurate anatomic region detector DCNN is critical to developing a staged algorithm pipeline like the one developed in this study.

The second stage of our pipeline (and the more difficult task) was determining whether a subtle fracture is present or not within the localized anatomic region of interest. Previous studies using DCNNs have focused on detecting the presence of large abnormalities, such as displaced long bone fractures and orthopedic implants [10, 1423]. While these deep learning models are promising, their ability to discriminate small fractures, such as triquetral and Segond fractures, has not been specifically evaluated. Such fractures may be easily missed due to the subtlety of their radiologic findings. By training DCNNs on automatically cropped images focused on the specific anatomic areas of interest for the two avulsion fractures, we demonstrated high diagnostic performance for both Segond and triquetral avulsion fractures on both internal cross-validation and external testing. Furthermore, the two-stage DCNN system had improved performance over a one-stage system that was trained on full (non-cropped radiographs), which is consistent with the idea of increasing the signal:noise ratio of the avulsion fractures by cropping the image to the anatomic region of interest.

By using Grad-CAMs, we are able to visualize the features of an image which DCNNs prioritize in deciding the classification. This technique provides evidence that algorithms

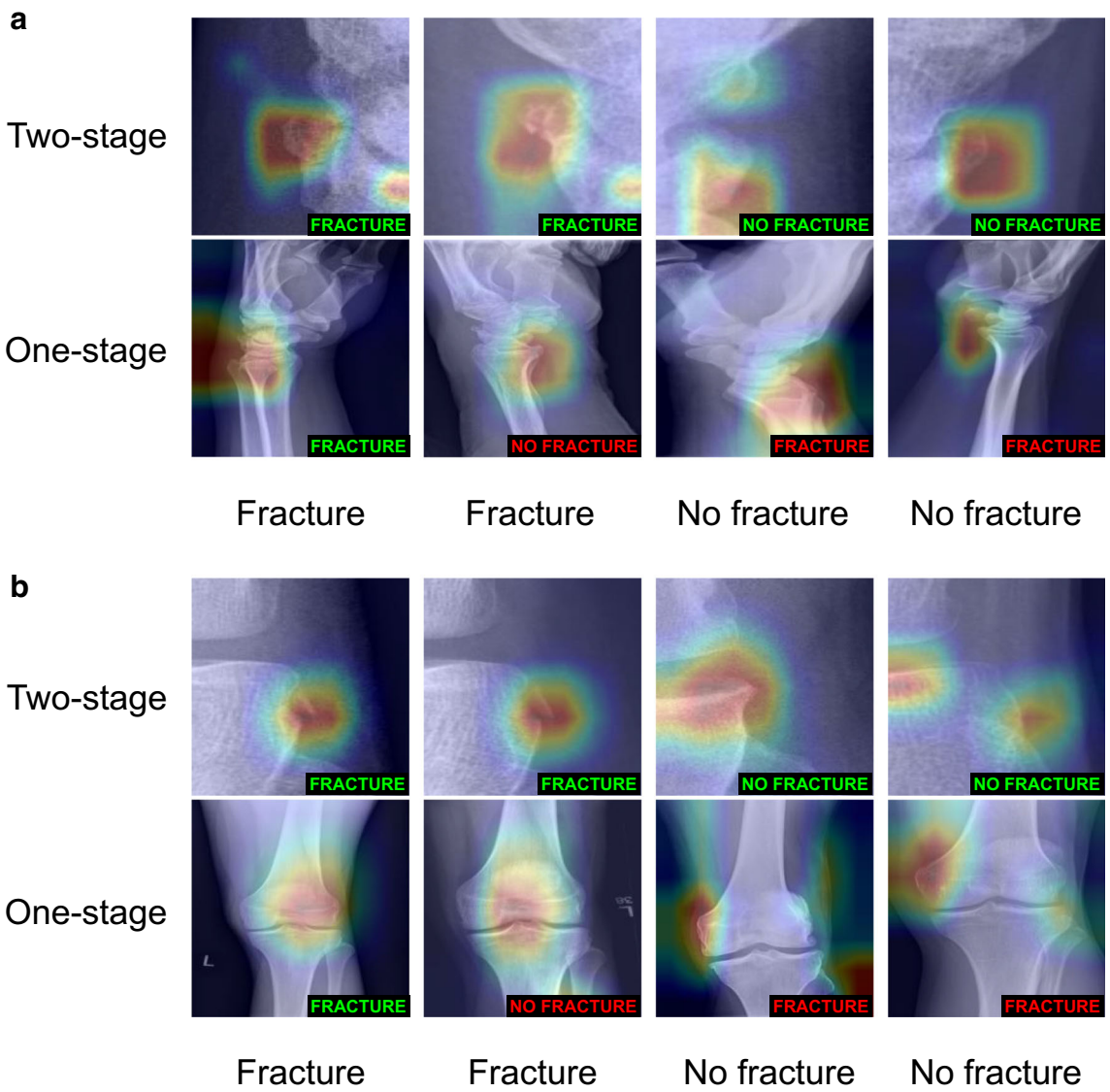

Fig. 4 Side-by-side examples of classifications made by the two-stage system and one-stage classifier for **a** triquetral avulsion fractures and **b** Segond fractures. True classifications are noted below each pair of images, and predicted classifications are noted on each image. Red color indicates greater weight in assigning the final classification

correctly classifying images are using features consistent with the given pathology. Importantly, the Grad-CAM heatmaps demonstrated that the two-stage classifiers precisely localized the fracture, while the one-stage classifiers had less precise localizations, and sometimes focused on regions far from the actual fracture, which suggests some degree of overfitting in the latter and the leveraging of confounding factors in the images to make the right diagnosis. For the single-stage triquetral fracture classifier, Grad-CAMs showed variable areas of strong activation, which included the wrist for most images but did not consistently focus on the dorsal part of the wrist, where the fracture would be expected to be found. However, all Grad-CAMs for the cropped classifier showed strong activation over the fracture fragment, if present, and/or the dorsal part of the triquetrum, if absent. Likewise, the one-stage DCNN for Segond fracture detection showed variable regions of activation, mostly at the center of the tibiofemoral joint, while all Grad-CAMs for the two-stage system showed strong activation just lateral to the lateral tibial plateau where the fracture fragment is or would be expected to be avulsed from, or at the fragment itself (Fig. 4).

Our study has several limitations. First, our training sample size was relatively small. However, we used standard data augmentation techniques to dramatically increase our training set volume and diversity, which resulted in strong results for our two-stage DCNN pipeline, especially when compared with a one-stage algorithm, which demonstrated that our two-stage system is accurate, generalizable to external data, and provides improvements over a standard one-stage DCNN classifier. Second, we used publicly available radiographs for the training and validation of our DCNNs, rather than a dataset made with known, controlled, and consistent selection and processing. In addition, neither the publicly available data from MURA, LERA, and the public domain nor the external test data from our institution had available clinical parameters for further description. However, the strong performance of our two-stage system across both datasets and improvement over a one-stage classifier indicate that our results may be generalizable, which may reflect the heterogeneity of data sources (which may have helped our DCNNs be resilient to differences in image acquisition protocols between sites), as well as the relatively constant appearance of avulsion fractures between different populations. Third, DCNNs do not explicitly indicate the precise features that they use to determine their decisions. By using Grad-CAMs, we are able to better understand if DCNN models are highlighting expected radiologic findings, such as fracture fragments, or confounding features, such as osseous anatomy distant from the fracture, although this is still a limited technique to fully "explain" the algorithms' decisions. Last, we only trained and compared DCNN models on two specific types of fracture, which limits generalizability of these findings to other types of fractures. Nevertheless, our results demonstrate a proof-of-concept of using staged algorithms to mimic a radiologist's search pattern, demonstrating that DCNN image classification accuracy can be improved by using such a strategy. We propose that a similar pipeline can be used for other subtle avulsion fractures, such as those in the foot and ankle.

Conclusion

Drawing inspiration from the common radiologist practice of zooming in on areas of interest to identify subtle avulsion fractures, we developed and evaluated a two-stage deep learning pipeline for the identification of Segond and triquetral avulsion fractures. We found that a two-stage pipeline increases accuracy in the detection of subtle fractures on radiographs compared with a one-stage DCNN classifier and generalized well to external test data. This staged pipeline could be applied to other subtle findings, such as other avulsion fractures, as well as non-traumatic findings, such as accessory bones or unfused ossification centers, which are similarly subtle osseous findings. By focusing attention on specific image regions in a manner mimicking a radiologist search pattern, deep learning algorithms appear to improve detection of subtle findings that may otherwise be missed.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skelet Radiol. 2020;49(2):183–97. https://doi.org/10.1007/s00256-019-03284-z.
2. Gao F, Wu T, Li J, et al. SD-CNN: a shallow-deep CNN for improved breast cancer diagnosis. Comput Med Imaging Graph. 2018;70:53–62. https://doi.org/10.1016/j.compmedimag.2018.09.004.
3. Kim TK, Yi PH, Hager GD, Lin CT. Refining dataset curation methods for deep learning-based automated tuberculosis screening. J Thorac Dis. 2019;3(2). https://doi.org/10.21037/jtd.2019.08.34.
4. Tsehay YK, Lay NS, Roth HR, et al. Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images. Med Imaging 2017 Comput Diagnosis. 2017;10134(March 2017):1013405. https://doi.org/10.1117/12.2254423.
5. Rajpurkar P, Irvin J, Bagul A, et al. MURA: large dataset for abnormality detection in musculoskeletal radiographs. 2017;(Midl 2018):1–10. http://arxiv.org/abs/1712.06957.
6. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skelet Radiol. 2019;48(2):239–44. https://doi.org/10.1007/s00256-018-3016-3.
7. Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures:

deep learning system versus radiologists. Eur J Radiol. 2020;126(February):108925. https://doi.org/10.1016/j.ejrad.2020.108925.
8. 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(45):11591–6. https://doi.org/10.1073/pnas.1806905115.
9. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018;73(5):439–45. https://doi.org/10.1016/j.crad.2017.11.015.
10. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell. 2019;1(1):e180001. https://doi.org/10.1148/ryai.2019180001.
11. Leong JJH, Nicolaou M, Emery RJ, Darzi AW, Yang GZ. Visual search behaviour in skeletal radiographs: a cross-speciality study. Clin Radiol. 2007;62(11):1069–77. https://doi.org/10.1016/j.crad.2007.05.008.
12. Manning D, Ethell S, Donovan T, Crawford T. How do radiologists do it? The influence of experience and training on searching for chest nodules. Radiography. 2006;12(2):134–42. https://doi.org/10.1016/j.radi.2005.02.003.
13. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. Radiographics. 2018;38(1):236–47. https://doi.org/10.1148/rg.2018170107.
14. 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):468–73. https://doi.org/10.1080/17453674.2018.1453714.
15. Varma M, Lu M, Gardner R, et al. Automated abnormality detection in lower extremity radiographs using deep learning. Nat Mach Intell. 2019;1(12):578–83. https://doi.org/10.1038/s42256-019-0126-0.
16. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology. 2019;290(3):590–606. https://doi.org/10.1148/radiol.2018180547.
17. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318–27. https://doi.org/10.1109/TPAMI.2018.2858826.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016-Decem:770–8. https://doi.org/10.1109/CVPR.2016.90.
19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Vol 17. IEEE; 2017:618–626. https://doi.org/10.1109/ICCV.2017.74.
20. Arbabshirani MR, Dallal AH, Agarwal C, Patel A, Moore G. Accurate segmentation of lung fields on chest radiographs using deep convolutional networks. Med Imaging 2017 Image Process. 2017;10133(February 2017):1013305. https://doi.org/10.1117/12.2254526.
21. Tuzoff DV, Tuzova LN, Bornstein MM, et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. Dentomaxillofacial Radiol. 2019;48(4):1–10. https://doi.org/10.1259/dmfr.20180051.
22. Yi PH, Kim TK, Wei J, et al. Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning. Pediatr Radiol. 2019;49(8):1066–70. https://doi.org/10.1007/s00247-019-04408-2.
23. Yi PH, Wei J, Kim TK, et al. Automated detection & classification of knee arthroplasty using deep learning. Knee. 2020;27(2):535–42. https://doi.org/10.1016/j.knee.2019.11.020.