Original Article

Application of convolutional neural networks for distal radio-ulnar fracture detection on plain radiographs in the emergency room

Min Woong Kim[1], Jaewon Jung[2], Se Jin Park[2], Young Sun Park[1], Jeong Hyeon Yi[1], Won Seok Yang[1], Jin Hyuck Kim[3], Bum-Joo Cho[2,4], Sang Ook Ha[1]

[1]Department of Emergency Medicine, Hallym University Sacred Heart Hospital, Hallym University Medical Center, Anyang, Korea
[2]Medical Artificial Intelligence Center, Hallym University Sacred Heart Hospital, Hallym University Medical Center, Anyang, Korea
[3]Department of Neurology, Hallym University Sacred Heart Hospital, Hallym University Medical Center, Anyang, Korea
[4]Department of Ophthalmology, Hallym University Sacred Heart Hospital, Hallym University Medical Center, Anyang, Korea

Objective Recent studies have suggested that deep-learning models can satisfactorily assist in fracture diagnosis. We aimed to evaluate the performance of two of such models in wrist fracture detection.

Methods We collected image data of patients who visited with wrist trauma at the emergency department. A dataset extracted from January 2018 to May 2020 was split into training (90%) and test (10%) datasets, and two types of convolutional neural networks (i.e., DenseNet-161 and ResNet-152) were trained to detect wrist fractures. Gradient-weighted class activation mapping was used to highlight the regions of radiograph scans that contributed to the decision of the model. Performance of the convolutional neural network models was evaluated using the area under the receiver operating characteristic curve.

Results For model training, we used 4,551 radiographs from 798 patients and 4,443 radiographs from 1,481 patients with and without fractures, respectively. The remaining 10% (300 radiographs from 100 patients with fractures and 690 radiographs from 230 patients without fractures) was used as a test dataset. The sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of DenseNet-161 and ResNet-152 in the test dataset were 90.3%, 90.3%, 80.3%, 95.6%, and 90.3% and 88.6%, 88.4%, 76.9%, 94.7%, and 88.5%, respectively. The area under the receiver operating characteristic curves of DenseNet-161 and ResNet-152 for wrist fracture detection were 0.962 and 0.947, respectively.

Conclusion We demonstrated that DenseNet-161 and ResNet-152 models could help detect wrist fractures in the emergency room with satisfactory performance.

Keywords Wrist; Fractures, bone; Deep learning; Neural networks, computer

Capsule Summary

What is already known

Recently, studies using various deep learning models as an assistive method for more accurate and efficient diagnosis of fractures have been actively conducted, and the models have shown satisfactory performance. Prior studies collected and analyzed heterogeneous image data. However, there are limitations in the possibility of their clinical use in the emergency room.

What is new in the current study

The study developed a model that is representative of emergency room settings with satisfactory performance to recognize fractures of the wrist.

INTRODUCTION

Wrist fractures are commonly diagnosed using simple radiographic images,[1] and corresponding treatment depends on the shape and stability of the fracture. Computed tomography can be used to provide a more accurate assessment of the presence and type of fractures and a better joint evaluation[2]; nevertheless, radiographs remain a rapid and low-cost primary method for early wrist trauma evaluation.[3] However, up to 30% of wrist fractures are misdiagnosed using radiographs.[4] This can result in mistreatment.

Recently, studies have used various deep-learning models as assistive methods for more accurate and efficient fracture diagnoses; several models (e.g., VGGNet, Inception-ResNet, Faster RCNN, and ViDi) have shown satisfactory performance.[5-10] However, because previous studies collected and analyzed heterogeneous image data, there were limitations to the clinical use of these models, especially in emergency room (ER) scenarios.

In the present study, we collected image data of emergency room patients with wrist trauma. Two types of convolutional neural networks (CNN) (i.e., DenseNet-161 and ResNet-152) were applied. The purpose of this study was, therefore, to evaluate the performance of the fracture detection model by analyzing the accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC) of each CNN.

METHODS

Study participants

We included data of ER patients with wrist trauma who underwent plain radiography between January 2018 and May 2020. Basic metabolic panel images of $1,776 \times 2,132$ pixels were retrieved from the hospital's picture archiving and communication system (PACS) using INFINITT PACS M6 software (INFINITT Healthcare, Seoul, Korea). Poor quality images and those lacking radiolo-

gist classifications were excluded. Annotations and personal information were omitted. Three image views (i.e., anteroposterior [AP] and bilateral oblique) were included for each patient. Images were classified into non-fracture and fracture groups based on dual radiological reporting. The fracture group images included those of radial, ulnar, and radio-ulnar fractures. The participation flowchart is presented in Fig. 1. Approval from the institutional review board of Hallym University Sacred Heart Hospital was obtained (2020-07-030), and participant informed consent was waived because of the retrospective nature of the study; moreover, this study adhered to the principles of the Declaration of Helsinki.

Dataset construction

We split the dataset into two subsets: training and testing. Radiographs taken between January 2018 and December 2019 and between January 2020 and May 2020 were included in the training and testing datasets, respectively. Furthermore, we allocated 10% of the training dataset, using patient identification numbers,
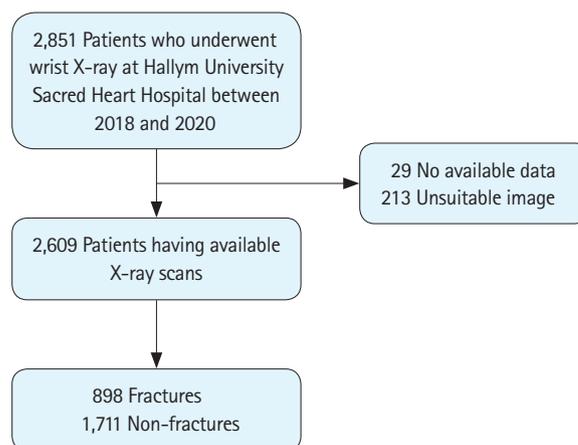

Fig. 1. Flow diagram for participant enrollment.

to a tuning dataset for hyperparameter tuning. The three datasets were separated from each other. The number of radiographs included in the fracture group was approximately half that of the non-fracture group. This fact presented a limitation, which was mitigated by oversampling; the number of radiographs in the fracture group in the training dataset was doubled by zooming-in 10%. Otherwise, no image flipping or rotation was applied.

Data preprocessing

Images were preprocessed using a contrast-limited adaptive histogram equalization (CLAHE) algorithm to enhance local contrast.[11] The CLAHE algorithm resolves noise amplification problems in small image regions. These problems often occur when adaptive histogram equalization via contrast limitation is applied. Before calculating a cumulative distribution function, histogram values were clipped to a predefined value to limit noise amplification. Consequently, the processed images more clearly revealed fractures. The CLAHE algorithm was implemented using OpenCV ver. 4.1.2.30 in Python. All images were then reduced to 550×660 pixels in consideration of memory capacity, batch sizes, training times, and model performance.

CNN

In this study, we adopted two CNN architectures: DenseNet-161 and ResNet-152. DenseNet-161 consisted of dense blocks. For this test, all output feature maps were propagated to all deeper layers as input to the blocks. The architecture utilized all previous feature maps to classify target objects without adding layers.[12] ResNet-152 was designed with a residual block.[13] Thus, a skip connection, in which an input feature map is added to the output feature map of a deeper layer, enabled the CNN model to learn residual features and have deeper layers.[13] This type of CNN model won the ImageNet Large Scale Visual Recognition Challenge in 2015 in the field of image classification, detection, and localization.[14] The two CNN models used in this study were pretrained with the ImageNet dataset and were fine-tuned during training.

The Adam optimizer was used to train the two CNN models, with a beta-1 of 0.9 and a beta-2 of 0.999, using a binary cross-entropy loss function. The initial learning rate was 1e-4, and a learning rate decay policy was used. Every 10 epochs, the learning rate was decreased by 90% until it reached 1e-7. The batch size was 4. The weight decay was 1e-4, and early stopping was used, with a starting point of 30 and a patience of 20, to avoid overfitting. Dropout was not applied to either CNN, and both models were implemented on the Pytorch deep-learning framework and trained on the NVIDIA GeForce Titan RTX graphics-processing unit (NVIDIA, Santa Clara, CA, USA).

Gradient-weighted class activation mapping (Grad-CAM) was used to present the region of the radiograph scan that contributed to the classification decision of the artificial intelligence (AI) model. Grad-CAM is a generalized version of class activation mapping.[15] Grad-CAM results were obtained using feature maps of the last layer in a CNN model generated from an input image and its gradient.[16] The gradient of feature maps was averaged using global average pooling, and the feature maps were multiplied using the averaged gradient along the channel side.[16] The final color map was achieved using an element-wise summation of the feature maps and a clipping of negative values to zero.[16]

Statistical analysis

The normality of data distributions was evaluated using the Kolmogorov-Smirnov test to select the appropriate parametric and non-parametric statistical methods. Categorical variables were analyzed using the chi-square test. Continuous variables were expressed as mean ± standard deviation and analyzed using the Student's t-test. Performance of the CNN models was evaluated using the separate test dataset. AUROCs were then computed. The accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were calculated, in terms of Youden index, using receiver operating characteristic (ROC) curves at the maximum point. Youden index, J, is calculated using the formula J = sensitivity + specificity–1; it is a criterion for finding the optimal threshold of ROC curves, regardless of prevalence.[6,17] We used DeLong's test to compare the performances of the two models. Two-tailed tests were used for all comparisons, and group differences with P < 0.05 were considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics ver. 21.0 (IBM Corp., Armonk, NY, USA).

RESULTS

Dataset details

Patients' demographic characteristics are summarized in Table 1. We included and analyzed the radiographs of 2,609 patients with wrist trauma admitted to the ER from January 2018 to May 2020.

Table 1. Demographic characteristics of enrolled patients

Factor	Total (n = 2,609, 100%)	Non-fracture (n = 1,711, 65.6%)	Fracture (n = 898, 34.4%)	P-value
Age (yr)	42.1 ± 22.9	41.5 ± 22.8	44.1 ± 23.2	0.004
≤ 15	396 (15.2)	259 (15.1)	137 (15.3)	0.936
> 16	2,213 (84.8)	1,452 (84.9)	761 (84.7)	
Sex, male	1,332 (51.1)	920 (53.8)	412 (45.9)	< 0.01

Values are presented as mean ± standard deviation or number (%).

Among 898 patients with fractures, 22 (2.4%), 482 (53.7%), and 394 (43.9%) had ulnar fractures alone, radial fractures alone, and radio-ulnar fractures, respectively. The overall, non-fracture group, and fracture group mean ages were 42.1, 41.5, and 44.2 years old (P = 0.004), respectively. There was no statistically significant difference in age groups between patients with and without fractures (P = 0.936). However, there was a significant difference in sex ratio between patients in the non-fracture and fracture groups (53.8% vs. 45.9% men in the non-fracture vs. fracture groups, respectively; P < 0.010). For model training, we used 4,551 radiographs from 798 patients with fractures and 4,443 radiographs from 1,481 patients without fractures. The remaining 10% of the whole dataset (300 radiographs from 100 patients with fractures

and 690 radiographs from 230 patients without fractures) was used as a test dataset (Table 2).

Performance of DenseNet-161 and ResNet-152 in wrist fracture detection

The sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of DenseNet-161 and ResNet-152 with the test dataset are shown in Table 3 (90.3%, 90.3%, 80.3%, 95.6%, and 90.3%, respectively vs. 88.6%, 88.4%, 76.9%, 94.7%, and 88.5%, respectively). The confusion matrix and ROC curve for the test dataset are shown in Figs. 2 and 3. The AUROCs of DenseNet-161 and ResNet-152 for wrist fracture detection were 0.962 and 0.947, respectively. DeLong's test demonstrated that DenseNet-161 had

Table 2. Details of training and test datasets

|	Whole dataset (100%)		Training dataset (90%)		Test dataset (10%)	
	No. of radiographs	No. of patients	No. of radiographs	No. of patients	No. of radiographs	No. of patients
Overall	9,984	2,609	8,994	2,279	990	330
Non-fracture	5,133	1,711	4,443	1,481	690	230
Fracture	4,851	898	4,551	798	300	100

Table 3. Performance of fracture detection by DenseNet-161 and ResNet-152 in test dataset

Model	Diagnostic performance (%)				
	Sensitivity	Specificity	PPV	NPV	Accuracy
DenseNet-161	90.3 ± 1.4	90.3 ± 1.3	80.3 ± 2.4	95.6 ± 0.7	90.3 ± 1.3
ResNet-152	88.6 ± 1.0	88.4 ± 1.0	76.9 ± 1.8	94.7 ± 0.5	88.5 ± 1.0

Values are presented as mean ± standard deviation.
PPV, positive predictive value; NPV, negative predictive value.
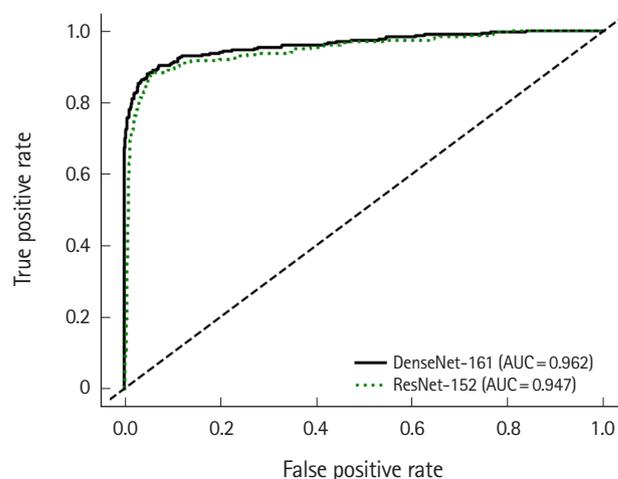
Fig. 3. Receiver operating characteristic curve of DenseNet-161 and ResNet-152 models for the detection of wrist fracture. The area under the receiver operating characteristic curves of DenseNet-161 and ResNet-152 for wrist fracture detection were 0.962 and 0.947, respectively. AUC, area under the curve.
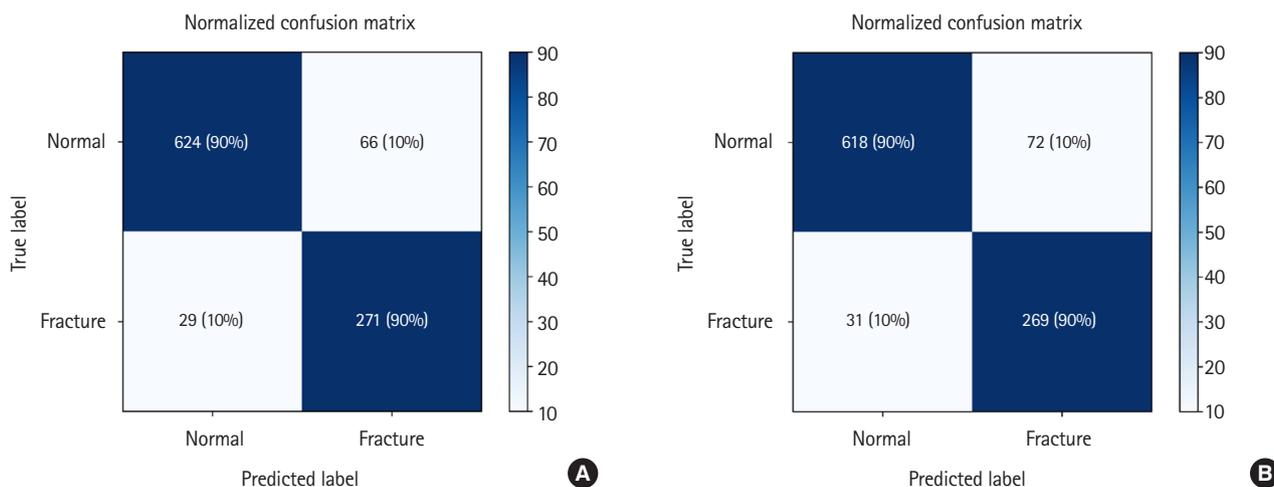
Fig. 2. Heatmap of confusion matrix for fracture detection by the best performing (A) DenseNet-161 and (B) ResNet-152.

Fig. 4. Gradient-weighted class activation mapping (Grad-CAM) for the detection of wrist fractures by DenseNet-161 and ResNet-152. The use of Grad-CAM demonstrates the importance of certain areas for accurate image classification. The percentage of Grad-CAM indicates the probability of wrist fracture occurrence. (A) DenseNet-161, fracture: 100%; (B) ResNet-152, fracture: 100%; (C) DenseNet-161, fracture: 100%; and (D) ResNet-152, fracture: 99.4%.

Fig. 5. (A) False-negative and (B) false-positive fracture detection examples. The model shows lower sensitivity for undisplaced or minimally-displaced fractures and ulnar styloid process fractures (A). Old fractures and artifacts on images were common causes of false-positive detections (B).

a significantly different AUROC from ResNet-152 (P < 0.050).

Fracture localization

The Grad-CAM algorithm of DenseNet-161 and ResNet-152 emphasized the most important area for image detection and classification (Fig. 4). The percentages indicated the probability of wrist fracture occurrence. The probabilities in Fig. 4A and 4B images of a pediatric patient were both 100%. The probabilities in Fig. 4C and 4D images of an adult patient were 100% and 99.4%, respectively.

Missed diagnosis

Fig. 5 presents false-negative and false-positive detection of wrist fractures. Findings determined to be false negatives were mainly undisplaced, minimally displaced, or ulnar styloid process fractures (Table 4). However, the deep learning models mainly classi-

Table 4. Misdiagnosis of fractures by DenseNet-161 and ResNet-152 in the displacement and non-displacement groups

Model	Displacement (n = 231)	Non-displacement (n = 69)
DenseNet-161	13 (5.6)	19 (27.5)
ResNet-152	18 (7.8)	22 (31.9)

Values are presented as number (%).

fied old fractures or artifacts as new fractures, which led to false-positive results.

DISCUSSION

The present study demonstrated that DenseNet-161 and ResNet-152 models could be trained to satisfactorily detect wrist fractures in the ER. Fractures are frequent issues in medical litigation, and misdiagnoses or delays can result in prolonged pain and long-term

Table 5. Previous studies using deep-learning models for the detection of wrist fractures

Study	No. of images	Model	Sensitivity	Specificity	mAP (m)/accuracy (A) (%)	AUROC
Olczak et al.[9]	25,658	VGG 16	-	-	(A) 0.83	-
Kim et el.[7]	1,389	Inception v3	0.90	0.88	-	0.95
Lindsey et al.[8]	34,990	U-Net	0.94	0.95	-	0.97/0.98
Yahalomi et al.[19]	4,476	VGG 16	-	-	(m) 0.87/(A) 96.0	-
Gan et al.[6]	2,340	Inception v4	0.90	0.96	-	0.96
Thian et al.[18]	14,614	Inception-ResNet	0.96	0.83	-	0.90
Bluthgen et al.[10]	524	ViDi	0.64-0.92	0.60-0.90	-	0.80-0.96
Present study	9,984	DenseNet-161	0.90	0.90	(A) 96.6	0.96
		ResNet-152	0.87	0.89	(A) 88.5	0.95

mAP, mean average precision; AUROC, areas under the receiver-operating characteristic curve.

complications. Thus, the presence of a fracture should be judged quickly and carefully. Several studies have been conducted using deep learning in real clinical cases beyond obtaining simple data results. In Table 5, we summarize such study results for the detection of wrist fractures.[6-10,18,19] Olczak et al.[9] supported the use of AI to identify orthopedic radiographs, providing human-level accuracies of 83% and 84% for AI compared to 82% for clinicians. In clinical practice, during overnight shifts in small ERs when radiologists and orthopedic surgeons are absent, AI tools can be used for triage. Lindsey et al.[8] evaluated the utility of a trained model by measuring its effect on the diagnostic accuracy of a group of emergency medicine clinicians. They reported that the ability of clinicians to diagnose wrist fractures could be improved with the aid of a deep-learning model. The average clinician sensitivity and specificity were improved from 80.8% to 91.5% and from 87.5% to 93.9%, respectively. These findings suggest that AI can be used efficiently in clinical practice to aid in fracture diagnosis.

In the present study, we reported DenseNet-161 and ResNet-152 AUROCs of 0.962 and 0.947, respectively, which were similar or somewhat lower than those reported in previous studies (with AUROCs of 0.80–0.98).[6-8,10] We suggest the following reasons for these results. First, results can be interpreted using only one dataset obtained from the confined ER environment. However, the prevalence of specific environment data use affects the actual learning performance, and therefore the process may be more suitable for use in real settings. Thian et al.[18] collected data from an ER and found an AUROC of 0.90, which was lower than those found in other studies. Second, we used a relatively small dataset, which had a greater impact compared to those of other studies. Lindsey et al.[8] used 31,490 wrist radiographs during model training. Similarly, Olczak et al.[9] obtained 65,264 wrist images. Third, fracture interpretation is typically radiologist-dependent, which can explain the differences in model performance across studies. Fourth, differences in learning methods could have affected the

results. The DenseNet-161 and ResNet-152 used in the present study were trained only with images with and without fractures for classification. However, in recurrent CNN models,[6] the fracture areas were indicated, and the models were trained directly for localization. Fifth, our data included radiographs of children. Different classifications are sometimes required for children because the degree of growth plate adhesion can vary, depending on bone age. The distal radial and ulnar epiphyses appear at ages of 1 and 5 years, respectively, and all tend to close between the ages of 16 and 18 years. Furthermore, fracture types vary in children. Sixth, we included radiographs with splints, which were easily interpretable, to increase their resemblance to real clinical data. During the initial learning phase, the model recognized fractures by first recognizing the splints. This method could produce false-positive results. Additionally, the splint acts as an artifact that can affect image interpretation.

Our study had the following strengths. First, we developed a model that is representative of the real-world clinical environment; it included all wrist radiographs from ER patients over a certain period, unlike previous studies,[7,9] wherein AI was trained using heterogeneous datasets. Second, DenseNet comprised dense blocks with densely connected layers.[12] DenseNet showed an improvement in accuracy, without performance degradation or overfitting, as parameters increased. This model encourages feature reuse, and substantially reduces the number of parameters and the amount of computation required to achieve state-of-the-art performance. In previous studies, DenseNet was applied in the diagnosis of ankle and hip fractures.[20,21] This study was the first to apply the technique in wrist fracture diagnosis. Third, previous studies[6,18] were conducted using radiographs comprising AP and lateral views. In the current study, the radiograph data consisted of AP and bilateral oblique views. Thus, accuracy was improved over the assessment of two views.

This study had some limitations. First, the dataset was small because the study was conducted over a short period in a single

research institute. To mitigate this limitation, we used data augmentation methods to increase the sample size. Subsequent studies having large sample datasets are required for external validation and more accurate model development. The second limitation resulted from DenseNet-161 and ResNet-152 characteristics. These models allowed algorithms to create good visualizations of fractures; moreover, they identified features previously missed by humans because they learned the most predictable features. In particular, it is possible that wrist angle or splint application at the time of radiography played a significant role in the judgment. Third, this study did not compare the ability of DenseNet-161 and ResNet-152 model with the ability of a clinician to diagnose wrist fractures.

In summary, this study demonstrated that DenseNet-161 and ResNet-152 models could be trained to detect wrist fractures in the ER with satisfactory performance.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

REFERENCES

1. Mauffrey C, Stacey S, York PJ, Ziran BH, Archdeacon MT. Radiographic evaluation of acetabular fractures: review and update on methodology. J Am Acad Orthop Surg 2018;26:83-93.

2. Fotiadou A, Patel A, Morgan T, Karantanas AH. Wrist injuries in young adults: the diagnostic impact of CT and MRI. Eur J Radiol 2011;77:235-9.

3. Newberg A, Dalinka MK, Alazraki N, et al. Acute hand and wrist trauma. American College of Radiology. ACR Appropriateness Criteria. Radiology 2000;215 Suppl:375-8.

4. Ootes D, Lambers KT, Ring DC. The epidemiology of upper extremity injuries presenting to the emergency department in the United States. Hand (N Y) 2012;7:18-22.

5. Adams M, Chen W, Holcdorf D, McCusker MW, Howe PD, Gaillard F. Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. J Med Imaging Radiat Oncol 2019;63:27-32.

6. Gan K, Xu D, Lin Y, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthop 2019;90:394-400.

7. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73:439-45.

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;115:11591-6.

9. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017; 88:581-6.

10. Bluthgen 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;126:108925.

11. Pizer SM, Johnston RE, Eriksen JP, Yankaskas BC, Muller KE. Contrast-limited adaptive histogram equalization: speed and effectiveness. Proceedings of the First Conference on Visualization in Biomedical Computing; 1990; Atlanta, GA, USA. New York, NY: IEEE; 1990. p.337-45.

12. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; Honolulu, HI, USA. New York, NY: IEEE; 2017. p.4700-8.

13. He K, Zhang X, Ren S, Sun J. Paper presented at: Identity mappings in deep residual networks. European Conference on Computer Vision; 2016; Amsterdam, Netherlands. Berlin: Springer; 2016. p.630-45.

14. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-52.

15. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, NV, USA. New York, NY: IEEE; 2016. p. 2921-9.

16. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:336-59.

17. 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;89:468-73.

18. Thian YL, Li Y, Jagmohan P, Sia D, Chan VE, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 2019;1: e180001.

19. Yahalomi E, Chernofsky M, Werman M. Detection of distal radius fractures trained by a small set of X-ray images and Faster R-CNN. Intelligent Computing-Proceedings of the Computing Conference; 2019; London, United Kingdom. Berlin: Springer; 2019. p.971-81.

20. Kitamura G, Chung CY, Moore BE 2nd. Ankle Fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. J Digit Imaging 2019;32:672-7.

21. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med 2019;2:31.