**Original Article** | Thoracic Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2019.0651
Korean J Radiol 2020;21(7):869-879

Korean Journal of Radiology
KJR

# Automatic Detection and Classification of Rib Fractures on Thoracic CT Using Convolutional Neural Network: Accuracy and Feasibility

Qing-Qing Zhou, MS[1], Jiashuo Wang, MS[2], Wen Tang, MS[3], Zhang-Chun Hu, MS[1], Zi-Yi Xia, BS[1], Xue-Song Li, MS[1], Rongguo Zhang, PhD[3], Xindao Yin, MD, PhD[4], Bing Zhang, MD, PhD[5], Hong Zhang, BS[1]

[1]Department of Radiology, The Affiliated Jiangning Hospital of Nanjing Medical University, Nanjing, China; [2]Research Center of Biostatistics and Computational Pharmacy, China Pharmaceutical University, Nanjing, China; [3]FL 8, Ocean International Center E, Beijing, China; [4]Department of Radiology, Nanjing First Hospital, Nanjing Medical University, Nanjing, China; [5]Department of Radiology, The Affiliated Nanjing Drum Tower Hospital of Nanjing University Medical School, Nanjing, China

**Objective:** To evaluate the performance of a convolutional neural network (CNN) model that can automatically detect and classify rib fractures, and output structured reports from computed tomography (CT) images.

**Materials and Methods:** This study included 1079 patients (median age, 55 years; men, 718) from three hospitals, between January 2011 and January 2019, who were divided into a monocentric training set (n = 876; median age, 55 years; men, 582), five multicenter/multiparameter validation sets (n = 173; median age, 59 years; men, 118) with different slice thicknesses and image pixels, and a normal control set (n = 30; median age, 53 years; men, 18). Three classifications (fresh, healing, and old fracture) combined with fracture location (corresponding CT layers) were detected automatically and delivered in a structured report. Precision, recall, and F1-score were selected as metrics to measure the optimum CNN model. Detection/diagnosis time, precision, and sensitivity were employed to compare the diagnostic efficiency of the structured report and that of experienced radiologists.

**Results:** A total of 25054 annotations (fresh fracture, 10089; healing fracture, 10922; old fracture, 4043) were labelled for training (18584) and validation (6470). The detection efficiency was higher for fresh fractures and healing fractures than for old fractures (F1-scores, 0.849, 0.856, 0.770, respectively, $p$ = 0.023 for each), and the robustness of the model was good in the five multicenter/multiparameter validation sets (all mean F1-scores > 0.8 except validation set 5 [512 x 512 pixels; F1-score = 0.757]). The precision of the five radiologists improved from 80.3% to 91.1%, and the sensitivity increased from 62.4% to 86.3% with artificial intelligence-assisted diagnosis. On average, the diagnosis time of the radiologists was reduced by 73.9 seconds.

**Conclusion:** Our CNN model for automatic rib fracture detection could assist radiologists in improving diagnostic efficiency, reducing diagnosis time and radiologists' workload.

**Keywords:** *Rib fractures; Convolutional neural networks; Deep learning; Artificial intelligence; Multidetector computed tomography; Structured report*

## INTRODUCTION

Rib fractures are the most frequently observed injury following thoracic blunt trauma, occurring in approximately 40–80% of cases (1, 2) and representing an important indicator of trauma severity (3). In one series, 81%

**Received:** August 31, 2019 **Revised:** January 10, 2020 **Accepted:** January 21, 2020
This study has received funding by Nanjing Health Commission (YKK17226).
**Corresponding author:** Hong Zhang, BS, Department of Radiology, The Affiliated Jiangning Hospital of Nanjing Medical University, No.168, Gushan Road, Nanjing 211100, China.
• Tel: (8625) 52281848-80080 • Fax: (8625) 52281256 • E-mail: jnyyfsk@126.com

of patients with two or more rib fractures had either pneumothorax or hemothorax (4). With the increased use of chest multislice CT scans, the detection rate of rib fractures has improved remarkably (5, 6). However, it is time-consuming and labor-intensive to detect rib fractures in the "24 ribs" on hundreds of thin-slice CT images and missed rib fractures are not uncommon (1, 7). Cho et al. (1) reported that the rate of missed diagnosis of rib fracture on initial CT images reached 20.7%, significantly higher than those of the thoracic vertebrae or sternum (nearly 100% sensitivity) (8, 9), which could lead to poor patient prognosis or adverse medicolegal disputes (10). Therefore, it is necessary to improve the accuracy of clinical diagnosis and reduce the rate of missed diagnoses.

Currently, the convolutional neural network (CNN), a deep learning technique, is widely used in the medical field due to its aid in reaching an accurate diagnosis, reducing medical errors, and improving productivity (11-13). Furthermore, CNNs have also been successfully employed in thoracic CT, for example, in the automated classification of pulmonary nodules (14, 15), carcinoma, or tuberculosis (16). However, to our best knowledge, automatic classification and localization of rib fractures and the output of structured reports of thoracic CT images using CNN methods have not been reported.

In our study, a CNN model was developed and three goals were pursued: 1) to verify the robustness of the optimal model with multicenter/multiparameter validation sets; 2) to merge the multilayer results to one fracture and output a structured report generated from the CNN model; and 3) to compare diagnostic efficiency among the structured report and experienced radiologists diagnosing assisted and unassisted by the CNN model. It was expected that our CNN model would improve diagnostic accuracy, reduce diagnosis time, and reduce the required manpower.

## MATERIALS AND METHODS

### Dataset and Classification Criteria

The local Institutional Review Board approved this multicenter retrospective study and waived the requirement for informed consent. A total of 1079 patients (25054 annotations) from three different hospitals (A, B, C) were included in this study. The monocentric data of 1004 patients (974 patients with rib fracture and 30 healthy controls) were collected using keyword searches in the picture archiving and communication system (PACS) from hospital A between January 2011 to January 2019 (Tables 1, 2). Moreover, data from 75 patients with rib fracture were collected from hospital B and C in January 2019 using the same method as the multicenter data. Among the 974 rib fracture patients (median age, 55 years; male, 643), 90% were treated as training (n = 876; median age, 55 years; male, 582) and others as validation (n = 98; median age, 58 years; male, 61) (Table 1). In addition, 5 independent multicenter/multiparameter validation sets (n = 173; median age, 59 years; male, 118) and a control set (n = 30; median age, 53 years; male, 18) were added to test the performance of the CNN model (Table 2). Figure 1 shows the flow chart of the study process.

In this study, rib fractures were classified into one of three main types: fresh fracture, healing fracture, and old fracture. Fresh fracture was defined based on its sharp margin, lack of periosteal reaction or callus formation, and imaged within approximately 3 weeks of trauma (17, 18). The healing fracture, intermediate between the fresh and old fracture, was imaged in the period with blurring of the fracture margins to callus formation after the trauma (19).

**Table 1. Clinical and Radiologic Information of Rib Fracture Patients from Monocentric Data**

| Variables | Total | Training Set | Validation Set | *P* |
|---|---|---|---|---|
| No. of patients | 974 | 876 (90) | 98 (10) | - |
| No. of thick slices (5 mm) | 679 | 614 (70.1) | 65 (66.3) | 0.442 |
| No. of thin slices (1 mm) | 295 | 262 (29.9) | 33 (33.7) | 0.442 |
| Median age (range) | 55 (20–97) | 55 (20–97) | 58 (22–89) | 0.190 |
| Sex (male:female) | 643:331 | 582:294 | 61:37 | 0.472 |
| No. of annotations | 20064 | 18584 | 1480 | - |
| Fresh fracture | 8179 | 7699 (41.4) | 480 (32.5) | - |
| Healing fracture | 8723 | 8112 (43.7) | 611 (40.7) | - |
| Old fracture | 3162 | 2773 (14.9) | 389 (26.8) | - |
| Pixels | 1024 x 1024 | 1024 x 1024 | 1024 x 1024 | - |

Numerical data were reported as median (range). Percentages were shown inside parentheses.


https://doi.org/10.3348/kjr.2019.0651

Korean Journal of Radiology
KJR

**Table 2. Clinical and Radiologic Information of Five Multicenter/Multiparameter Validation Sets**

| Variables | Validation Sets | | | | | Control Set | *P* |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| No. of patients | Hospital A (n = 33) | Hospital A (n = 65) | Hospital B (n = 25) | Hospital C (n = 25) | Hospital B/C (n = 25) | Hospital A (n = 30) | - |
| Median age (range) | 62 (26–80) | 59 (24–89) | 59 (24–87) | 53 (28–73) | 61 (29–73) | 53 (32–71) | 0.625 |
| Sex (male:female) | 19:14 | 42:23 | 19:6 | 18:7 | 20:5 | 18:12 | 0.384 |
| Slice thickness (mm) | 1 | 5 | 1 | 2 | 1/2 | 1 | - |
| Pixels | 1024 x 1024 | 1024 x 1024 | 1024 x 1024 | 1024 x 1024 | 512 x 512 | 1024 x 1024 | - |
| No. of CT images* | 809 | 491 | 1468 | 1006 | 1667 | 9917 | - |
| Annotations | 881 | 599 | 1708 | 1150 | 2132 | - | - |
| Fresh fracture | 214 (24.3) | 266 (44.4) | 567 (33.2) | 270 (23.5) | 1073 (50.3) | - | - |
| Healing fracture | 418 (47.4) | 193 (32.2) | 1001 (58.6) | 388 (33.7) | 810 (38.0) | - | - |
| Old fracture | 249 (28.3) | 140 (23.4) | 140 (8.2) | 492 (42.8) | 249 (11.7) | - | - |

Numerical data were reported as median (range). Percentages were shown inside parentheses. *Validation set 1–5 contained only images with annotations, and control set contained patients' all CT images.

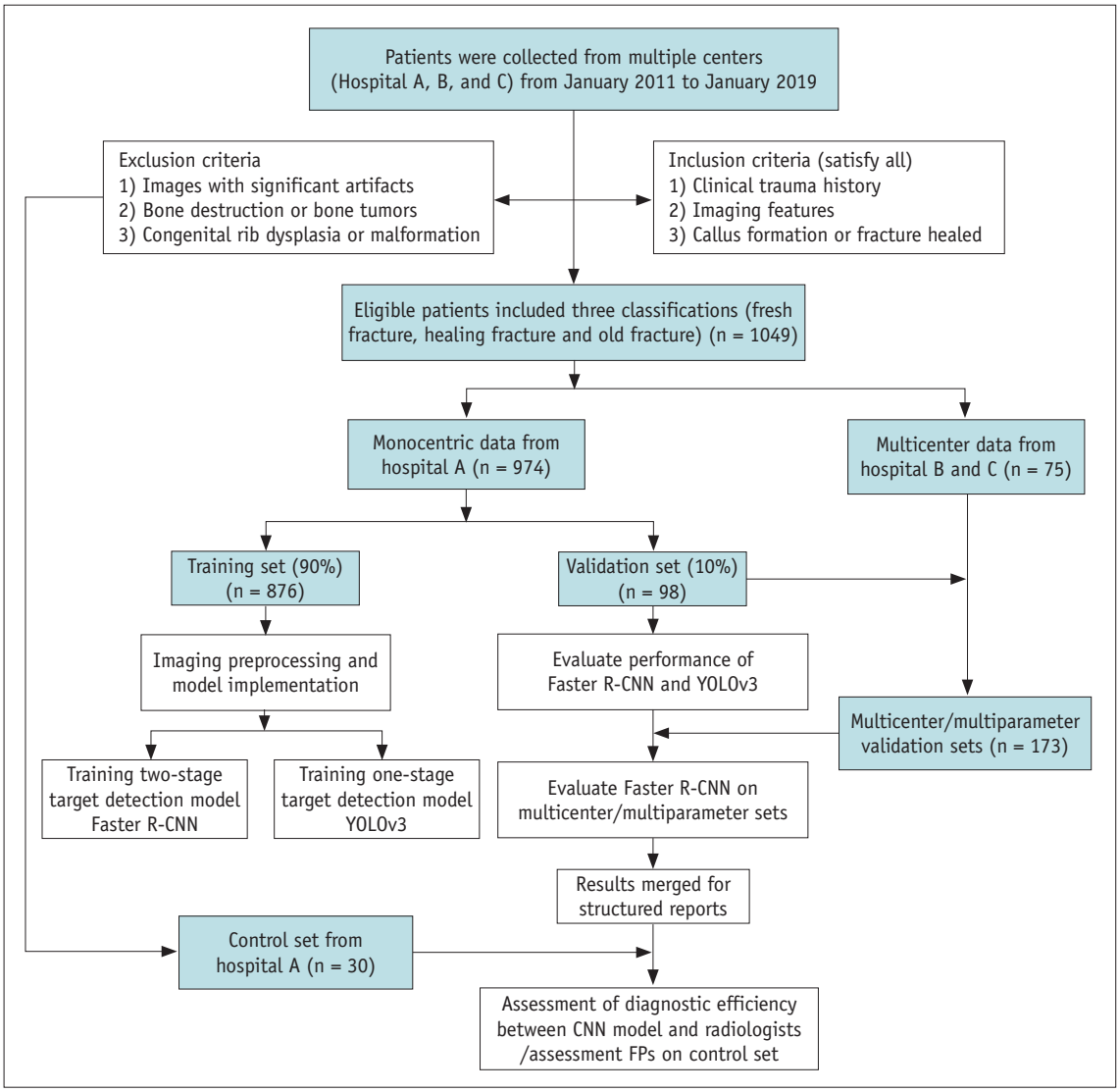


**Fig. 1. Flow chart showing overall study process.** CNN = convolutional neural network, Faster R-CNN = faster region-based convolutional neural network, FP = false positive, YOLOv3 = you only look once v3


https://doi.org/10.3348/kjr.2019.0651

Bone healing in rib fractures typically takes approximately 12 weeks (20); therefore, the fracture was defined as old if it was imaged approximately 3 months after trauma and mature callus, bony remodeling, non-visualization of the fracture line (19), and no change in follow-up scans were seen.

Data were only collected from patients with rib fracture who met all three inclusion criteria: 1) a medical history of trauma (obtained from electronic medical records); 2) imaging features of rib fracture; and 3) callus formation or healed fracture on follow-up CT scans. As for the healthy controls, the inclusion criteria were: 1) CT examination including 24 ribs without a history of trauma and 2) no rib fractures found by two senior radiologists. The exclusion criteria applicable to all participants were as follows: 1) images with significant radial or motion artefacts affecting the diagnosis of rib fracture; 2) bone destruction or bone tumor; and 3) congenital rib dysplasia or malformation.

### Imaging Annotation and Preprocessing

The CT images, including monocentric (training and validation dataset) and multicenter validation datasets, were annotated by two experienced musculoskeletal radiologists (8 and 9 years of experience in CT diagnosis) and checked by two senior radiologists (20 and 14 years of experience in CT diagnosis, respectively). If the conclusion was inconsistent, one thoracic surgeon was invited to participate in the discussion, and the final discussion result was established as the gold standard (ground truth [GT]) for diagnosis and classification. A rectangular bounding box, approximately 0.7–1.5 cm in size, was drawn on every CT slice of rib fractures using labelImg (version: 1.8.1, available at https://github.com/tzutalin/labelImg). CT scanners, scanning parameters, and details of CT image preprocessing are presented in Supplementary Material 1.

### Model Architecture and Implementation

The faster region-based convolutional neural network (Faster R-CNN) (21) (Fig. 2) and you only look once v3 (YOLOv3) model (22) were used in this study. The interpretation of those two models, and architecture and implementation of the CNN models are described in Supplementary Material 2.

### Model Comparison and Validation

Patients with rib fractures in hospital A were split into a monocentric training set (90%, n = 876) and validation set (10%, n = 98) using the random function of Python (version 2.7.15, available at http://www.python.org). The training set was used to fine tune the pre-trained model to fit a specific fracture image. The validation set was employed to evaluate the performance of the model, which included the accuracy of the classification and location. In this study, GT included the category of rib fracture and true bounding box. Supplementary Material 3 details the judgment criteria of the CNN models.

To objectively evaluate the performance between the Faster R-CNN and YOLOv3, three evaluation indicators, precision, recall, and F1-score, were calculated to select

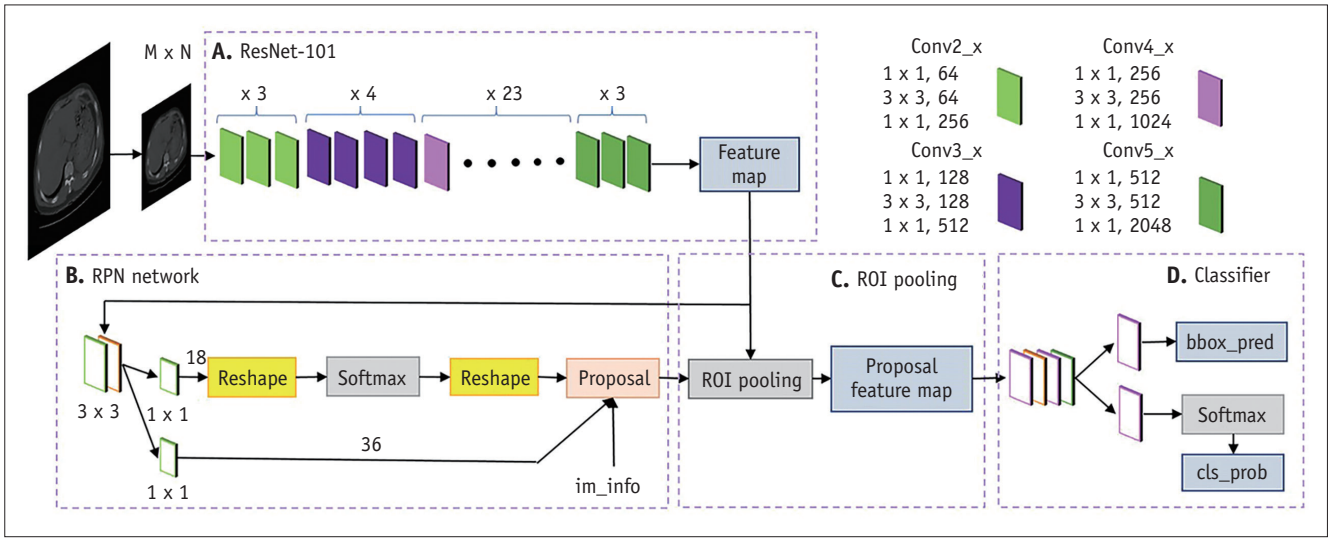


**Fig. 2. Schematic illustration of Faster R-CNN architecture.**
**A.** ResNet-101. **B.** RPN network was mainly used to generate regional proposals. **C.** ROI pooling. **D.** Classifier. ROI = region of interest, RPN = region proposal network


https://doi.org/10.3348/kjr.2019.0651

Korean Journal of Radiology
KJR

the optimal model. Supplementary Material 4 provides an explanation of these indicators.

In order to evaluate the generalizability, stability, and robustness of the CNN model, five different validation sets spanning different hospitals and including various parameters (slice thickness and pixels) were added to the model. Validation set 1 (n = 33, 1-mm slice thickness) and validation set 2 (n = 65, 5-mm slice thickness) were from the monocentric validation set. Validation set 3 (n = 25, 1-mm slice thickness) was from hospital B and validation set 4 (n = 25, 2-mm slice thickness) was from hospital C. The CT images in validation sets 1–4 were 1024 x 1024 pixels, and those in validation set 5 (n = 25, 1-mm or 2-mm slice thickness) from hospital B/C were 515 x 512 pixels (Table 2). All training and validation sets only contained images with annotations; images without fractures were not included.

### Merging Results for Structured Report

A program, which was able to combine several adjacent annotations of thin CT images (1-mm or 2-mm slice thickness) into one result, was designed to output structured reports. The reports contained the location and classification of the rib fractures. Localization included marking several rectangular boxes at the fracture site and outputting the numbers of the corresponding CT layers. We employed the Dice value to judge whether detection results in different layers or different parts of one image belonged to the same fracture. The detailed explanation is presented in Supplementary Material 5 and the structured report from CNN model is shown in Figure 3.

### Comparison of Diagnostic Efficiency between the CNN Model and Radiologists

To compare the efficiency of the CNN model and that of experienced radiologists in diagnosing and classifying rib fractures, three settings were included in our study: structured report, conventional diagnosis by experienced radiologists, and comprehensive results based on artificial intelligence (AI)-assisted diagnosis. The dataset for the test encompassed all CT images (data from validation set 1, 1-mm slice thickness, 1024 x 1024 pixels) from 33 patients. Five attending radiologists (no overlap with the radiologists who labelled and checked the annotations), each with 6–8 years of experience in CT diagnosis, participated in the study. They were informed of the gold standard criteria for rib fracture classification. They were required to read and record the localization (corresponding CT layers of the rib fracture) and category of rib fracture using the bone window with thoracic CT images. The second test was conducted

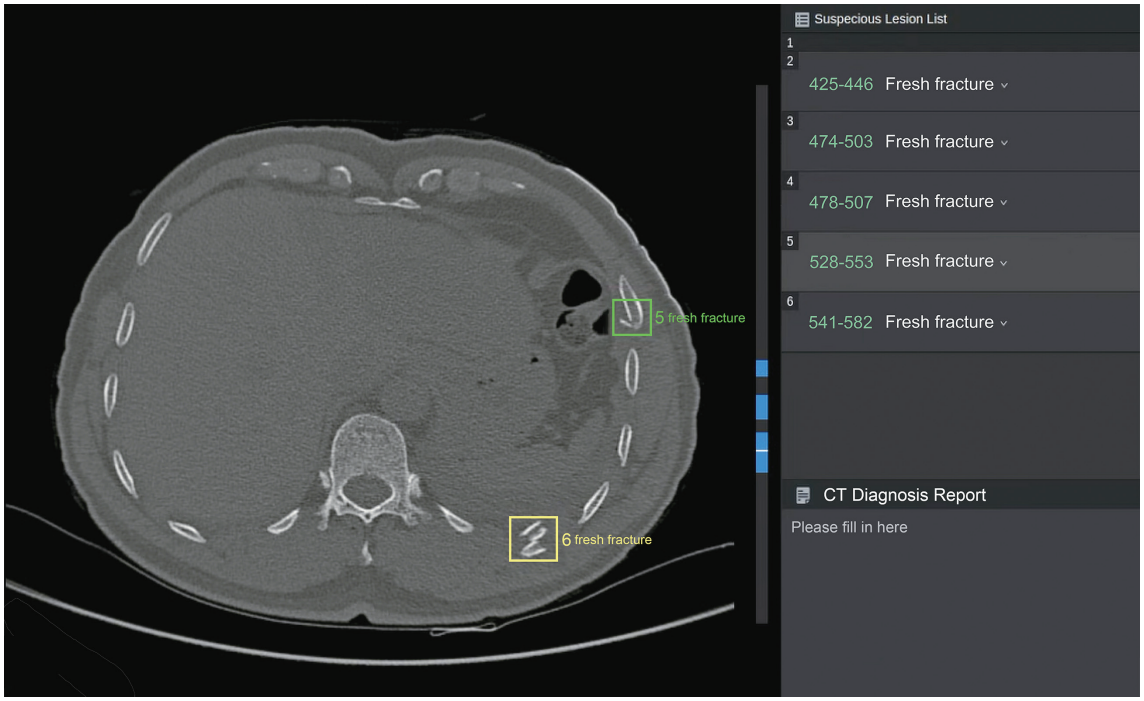


**Fig. 3. CT image with rectangular boxes and corresponding CT image reports.** All detected fractures were listed in sequence with numbers of corresponding CT layers (green numbers) (left). Preceding small white numbers correspond to fractures labelled in CT image (right).


https://doi.org/10.3348/kjr.2019.0651

two months later and diagnosed based on the structured report. The diagnosis time was recorded using a stopwatch and the detection time of the CNN model was obtained from the computer terminal.

A control set contained 30 subjects without rib fractures was added to the study in order to validate the performance of the CNN model in images without fractures. The false positives (FPs) and their frequency in patient and lesion level were calculated.

The number of true negatives for the ribs was too large to be calculated, as one rib may have myriad normal parts; therefore, the receiver operating characteristic (ROC) curve and specificity were not applicable. We used a free-response ROC (fROC) curve to evaluate the comprehensive performance of the model, as described by Setio et al. (23), because fROC allows for many lesions and normal appearance on one image. Detection/diagnosis time, precision, and sensitivity were selected as evaluation indices.

### Statistical Analysis

Precision, recall, and F1-score were selected as performance metrics of the CNN models, and the corresponding 95% confidence intervals were estimated using bootstrapping with 1000 bootstraps. The sensitivity and average number of FPs per patient in the CNN model of whole CT images from 33 patients without merging results were analyzed using fROC curves and the 11 points

**Table 3. Performance Metrics for Multicenter and Multiparameter Validation**

| Indicators | Validation Set | | | | | Mean |
|---|---|---|---|---|---|---|
| | 1 (n = 33) | 2 (n = 65) | 3 (n = 25) | 4 (n = 25) | 5 (n = 25) | |
| Precision | | | | | | |
| Fresh fractures | 203/259 = 0.784 (0.722–0.832) | 231/250 = 0.924 (0.895–0.955) | 487/600 = 0.812 (0.746–0.843) | 225/282 = 0.798 (0.726–0.852) | 853/961 = 0.888 (0.857–0.912) | 0.841 |
| Healing fractures | 365/388 = 0.941 (0.908–0.961) | 171/206 = 0.830 (0.803–0.868) | 840/913 = 0.920 (0.896–0.943) | 343/461 = 0.744 (0.696–0.787) | 660/778 = 0.848 (0.812–0.876) | 0.857 |
| Old fractures | 200/237 = 0.844 (0.791–0.901) | 116/144 = 0.806 (0.773–0.859) | 129/182 = 0.709 (0.620–0.782) | 367/415 = 0.884 (0.832–0.922) | 144/230 = 0.626 (0.537–0.696) | 0.774 |
| Mean | 0.856 | 0.853 | 0.814 | 0.809 | 0.787 | 0.824 |
| Recall | | | | | | |
| Fresh fractures | 203/214 = 0.949 (0.897–0.978) | 231/266 = 0.868 (0.841–0.898) | 487/567 = 0.859 (0.846–0.871) | 225/270 = 0.833 (0.811–0.866) | 853/1073 = 0.795 (0.780–0.808) | 0.861 |
| Healing fractures | 365/418 = 0.873 (0.860–0.884) | 171/193 = 0.886 (0.857–0.926) | 840/1001 = 0.839 (0.830–0.852) | 343/388 = 0.884 (0.869–0.899) | 660/810 = 0.815 (0.804–0.831) | 0.859 |
| Old fractures | 200/249 = 0.803 (0.774–0.830) | 116/140 = 0.829 (0.795–0.884) | 129/140 = 0.921 (0.901–0.939) | 367/492 = 0.746 (0.727–0.774) | 144/249 = 0.578 (0.521–0.631) | 0.775 |
| Mean | 0.875 | 0.861 | 0.873 | 0.821 | 0.729 | 0.832 |
| F1-score | | | | | | |
| Fresh fractures | 1.488/1.733 = 0.859 (0.824–0.886) | 1.604/1.792 = 0.895 (0.868–0.914) | 1.395/1.671 = 0.835 (0.798–0.863) | 1.329/1.631 = 0.815 (0.785–0.844) | 1.412/1.683 = 0.839 (0.823–0.847) | 0.849 |
| Healing fractures | 1.643/1.814 = 0.906 (0.890–0.917) | 1.471/1.716 = 0.857 (0.827–0.885) | 1.544/1.759 = 0.878 (0.869–0.888) | 1.315/1.628 = 0.808 (0.783–0.834) | 1.382/1.663 = 0.831 (0.817–0.848) | 0.856 |
| Old fractures | 1.355/1.647 = 0.823 (0.796–0.852) | 1.336/1.635 = 0.817 (0.792–0.848) | 1.306/1.630 = 0.801 (0.757–0.845) | 1.319/1.630 = 0.809 (0.792–0.830) | 0.724/1.204 = 0.601 (0.552–0.643) | 0.770 |
| Mean | 0.863 | 0.856 | 0.840 | 0.811 | 0.757 | 0.825 |
| Total FPs* | 116 | 82 | 239 | 223 | 312 | 194 |
| Fresh fractures | 56 (41–78) | 19 (11–27) | 113 (91–166) | 57 (39–85) | 108 (82–142) | 71 |
| Healing fractures | 23 (15–37) | 35 (26–49) | 73 (51–97) | 118 (93–150) | 118 (93–153) | 73 |
| Old fractures | 37 (22–53) | 28 (19–34) | 53 (36–79) | 48 (31–74) | 86 (63–124) | 50 |
| Mean | 39 | 27 | 80 | 74 | 104 | 65 |

Corresponding 95% confidence intervals, shown inside parentheses, were estimated by using bootstrapping with 1000 bootstraps and randomly sampled at annotations level. *Number of FPs was total number of FPs annotations. Validation set 1 and 2 were from monocentric validation set, and validation set 3–5 were from multicenter data. FPs = False positives


https://doi.org/10.3348/kjr.2019.0651

Korean Journal of Radiology
KJR

of the structured report and diagnoses of radiologists with and without AI assistance were overlaid on the curves. Supplementary Material 6 provides the detailed statistical methods.

## RESULTS

### Patient Characteristics

There was no significant difference in age, sex, or slice thickness between the monocentric training and validation set, and no difference of age or sex was observed between the validation sets and control set (all $p > 0.05$) (Table 2).

### Comparison of the Classification Models

The mean precision (0.862 > 0.670), mean recall (0.867 > 0.554), and mean F1-score (0.864 > 0.603) were significantly better for the Faster R-CNN ($p = 0.037$, 0.0002, and 0.002, respectively). The results are tabulated in Supplementary Material 7. Therefore, Faster R-CNN was chosen as our study model and proceeded to the next step of validation.

### Multicenter and Multiparameter Validation of the CNN Model

As shown in Table 3, except validation set 5 (1-mm slice thickness; 512 x 512 pixels) and old fractures, the mean precision, recall, and F1-score for the five different validation sets and three different classifications were all > 0.8. Based on the Kruskal-Wallis H test and least significant difference post hoc test (rank conversion), there were no significant differences in the performance regarding fresh and healing fractures (mean F1-score: 0.849 and 0.856, respectively; $p = 0.999$), and both were identified better than old fractures (mean F1-score: 0.849, 0.856, 0.770, respectively; all $p = 0.023$).

### Comparison of Diagnostic Efficacy of CNN Model and Radiologists

The 11 points of the structured report and radiologists with and without AI assistance were all above the fROC curve, and the five points representing diagnoses after AI assistance had the highest value (Fig. 4). The precision of the structured report and five radiologists without AI assistance were comparable (mean precision: 0.642 < 0.870, 0.803 < 0.848 and 0.826 > 0.692 for fresh, healing, and old fractures, respectively; $p = 0.001$, 0.578, and 0.117, respectively) and the sensitivity was higher for the three types of structured report (mean recall: 0.956 > 0.725, 0.875 > 0.614, and 0.704 > 0.533 for fresh, healing, and old fractures; all $p < 0.05$) (Table 4).

The mean sensitivity of the five radiologists' diagnoses increased from 0.624 to 0.863 (increased by 23.9%) after AI assistance ($p = 0.008$), and the mean precision of diagnosis improved from 0.803 to 0.911 ($p = 0.008$) (Table 4). Moreover, the detection/diagnosis time of the structured report, and radiologists with and without AI assistance were

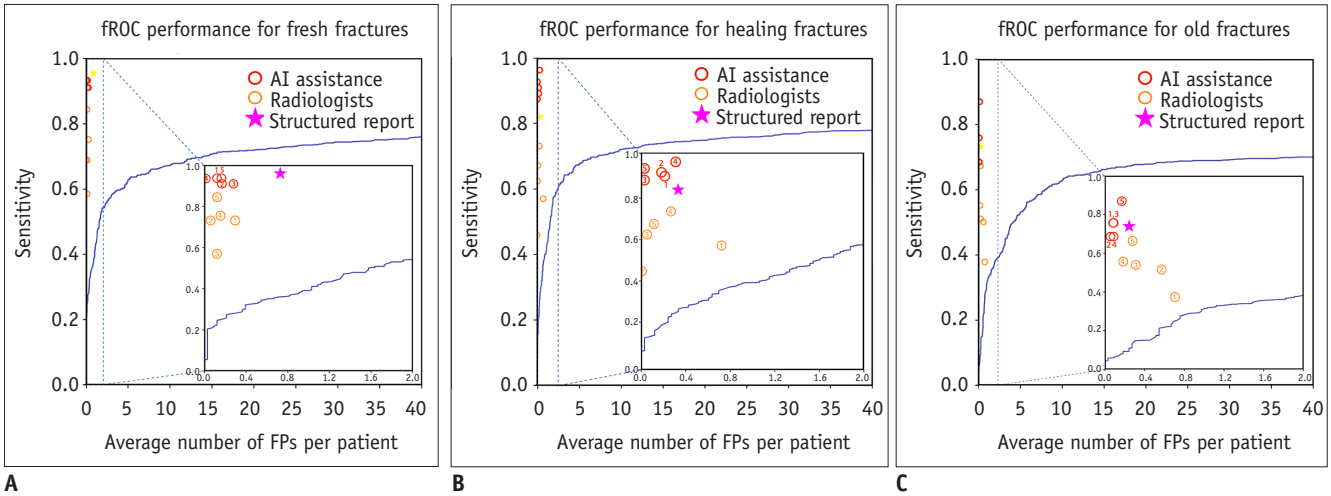


**Fig. 4. Comparison of diagnostic efficiency for different fractures in different situations on fROC curves.**
True positive rate and average number of FP per scan of fresh fractures **(A)**, and healing fractures **(B)**, old fractures **(C)** on whole CT images from 33 patients without merging results are shown by fROC curves. From enlarged inset, 11 points of structured report (yellow star) and radiologists with and without AI assistance (red and orange circles, respectively) were all above curve; among them, five points representing AI-assisted diagnosis (red circles) were greatest (all located in upper left corner). AI = artificial intelligence, fROC = free-response receiver operating characteristic


https://doi.org/10.3348/kjr.2019.0651

**Table 4. Comparison of Precision and Sensitivity of Different Fractures in Different Situations**

| Indicators | Fresh Fractures | Healing Fractures | Old Fractures | Mean |
|---|---|---|---|---|
| Mean precision | | | | |
| Five validation sets with fracture images | 4.206/5 = 0.841 | 4.283/5 = 0.857 | 3.869/5 = 0.774 | 0.824 |
| Full images without results merged | 165/283 = 0.583 | 345/483 = 0.714 | 168/296 = 0.568 | 0.622 |
| Structured report | 43/67 = 0.642 | 49/61 = 0.803 | 38/46 = 0.826 | 0.757 |
| Radiologists without AI assistance* | 4.351/5 = 0.870 | 4.243/5 = 0.848 | 3.459/5 = 0.692 | 0.803 |
| Radiologists with AI assistance* | 4.457/5 = 0.891 | 4.577/5 = 0.915 | 4.642/5 = 0.928 | 0.911 |
| Mean sensitivity | | | | |
| Five validation sets with fracture images | 4.304/5 = 0.861 | 4.297/5 = 0.859 | 3.877/5 = 0.775 | 0.832 |
| Full images without results merged | 165/214 = 0.771 | 345/418 = 0.825 | 168/249 = 0.675 | 0.757 |
| Structured report | 43/45 = 0.956 | 49/56 = 0.875 | 38/54 = 0.704 | 0.845 |
| Radiologists without AI assistance† | 3.623/5 = 0.725 | 3.071/5 = 0.614 | 2.667/5 = 0.533 | 0.624 |
| Radiologists with AI assistance† | 4.621/5 = 0.924 | 4.570/5 = 0.914 | 3.758/5 = 0.752 | 0.863 |

*Precision of radiologists' diagnoses increased 10.8% after AI assistance ($p$ = 0.008), †Sensitivity of diagnosis increased 23.9% after AI assistance ($p$ = 0.008). AI = artificial intelligence

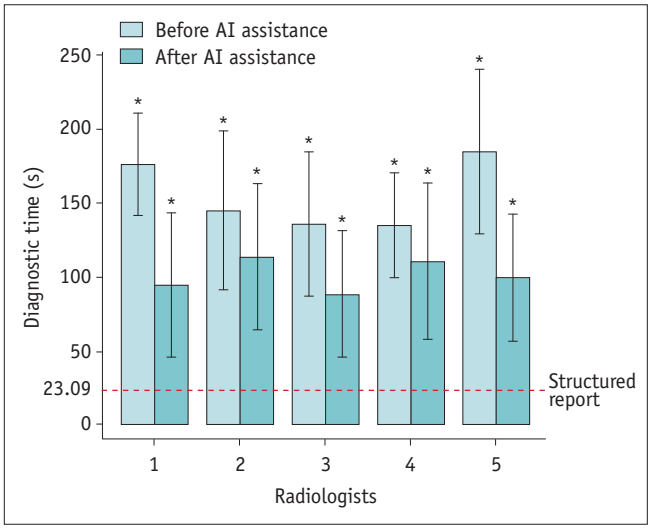


**Fig. 5. Bar graph of time to diagnosis.** Time to diagnosis of five different radiologists decreased when AI assistance was used (all $p$ < 0.01) and average time decrease was 73.9 seconds.

significantly different (23.08 ± 8.15 seconds, 101.25 ± 47.75 seconds, and 155.15 ± 50.34 seconds, respectively; $p$ < 0.01) (Fig. 5). Diagnosis time was reduced by an average of 73.9 seconds after AI assistance. Supplementary Material 8 provides the detailed performances of individual radiologists in the reading test. In the control set, FPs were inevitable. The FP frequency (FPs per patient) for fresh, healing, and old fractures was 0.200, 0.167, and 0.033, respectively, at a per-patient level and 0.333, 0.200, and 0.067, respectively, at a per-lesion level (Table 5). Figure 6 displays the CT images detected/diagnosed correctly by CNN model and radiologists (Fig. 6A-C), misdiagnosed by radiologists (Fig. 6D, E) and with FPs in control set by CNN model (Fig. 6F-I).

## DISCUSSION

We presented a method for the fully automatic detection and classification of rib fractures based on Faster R-CNN and assessed the performance of the algorithm twice, first using the raw output from the CNN and second using the merged structured report. The results demonstrate that our model has good performance in classifying rib fractures into three different categories, as verified by multicenter/multiparameter validation sets. In addition, after AI assistance was implemented, both the precision and sensitivity improved remarkably and the diagnosis time was reduced observably.

In the current study, the Faster R-CNN showed better performance in fracture detection and classification than YOLOv3. The Faster R-CNN is a two-stage algorithm with real-time performance and superior detection accuracy (24), whereas YOLOv3 is the most advanced one-stage algorithm and is focused on detection speed. In terms of detection efficiency of the different categories, detection of healing fractures and fresh fractures was better than that of old fractures. This was probably because the proportion of healing fractures and fresh fractures was higher than that of old fractures. Patients with fresh or healing fractures usually went to the hospital for CT examination and were followed up many times. However, old fractures were seldom re-examined. In addition, old fractures were similar to the surrounding healthy ribs, with mature calluses, no visible fracture lines (19), so it was difficult to distinguish


https://doi.org/10.3348/kjr.2019.0651

Korean Journal of Radiology
KJR

**Table 5. FPs and Frequency of Rib Fracture of Structured Report in Control Set (n = 30)**

| Category | Per-Patient Level | | Per-Lesion Level | |
|---|---|---|---|---|
| | FPs | Frequency (FPs/Patient) | FPs | Frequency (FPs/Patient) |
| Fresh fractures | 6 (2–9) | 6/30 = 0.200 (0.067–0.300) | 10 (3–16) | 10/30 = 0.333 (0.100–0.533) |
| Healing fractures | 5 (1–8) | 5/30 = 0.167 (0.033–0.267) | 6 (1–10) | 6/30 = 0.200 (0.033–0.333) |
| Old fractures | 1 (0–2) | 1/30 = 0.033 (0.000–0.067) | 2 (0–4) | 2/30 = 0.067 (0.000–0.133) |
| Total | 10 (3–19) | 10/30 = 0.333 (0.100–0.633) | 18 (4–30) | 18/30 = 0.600 (0.133–1.000) |

Correspondence 95% confidence intervals were shown inside parentheses. In patient level, there were 2 patients who have PFs of both fresh and healing fractures.

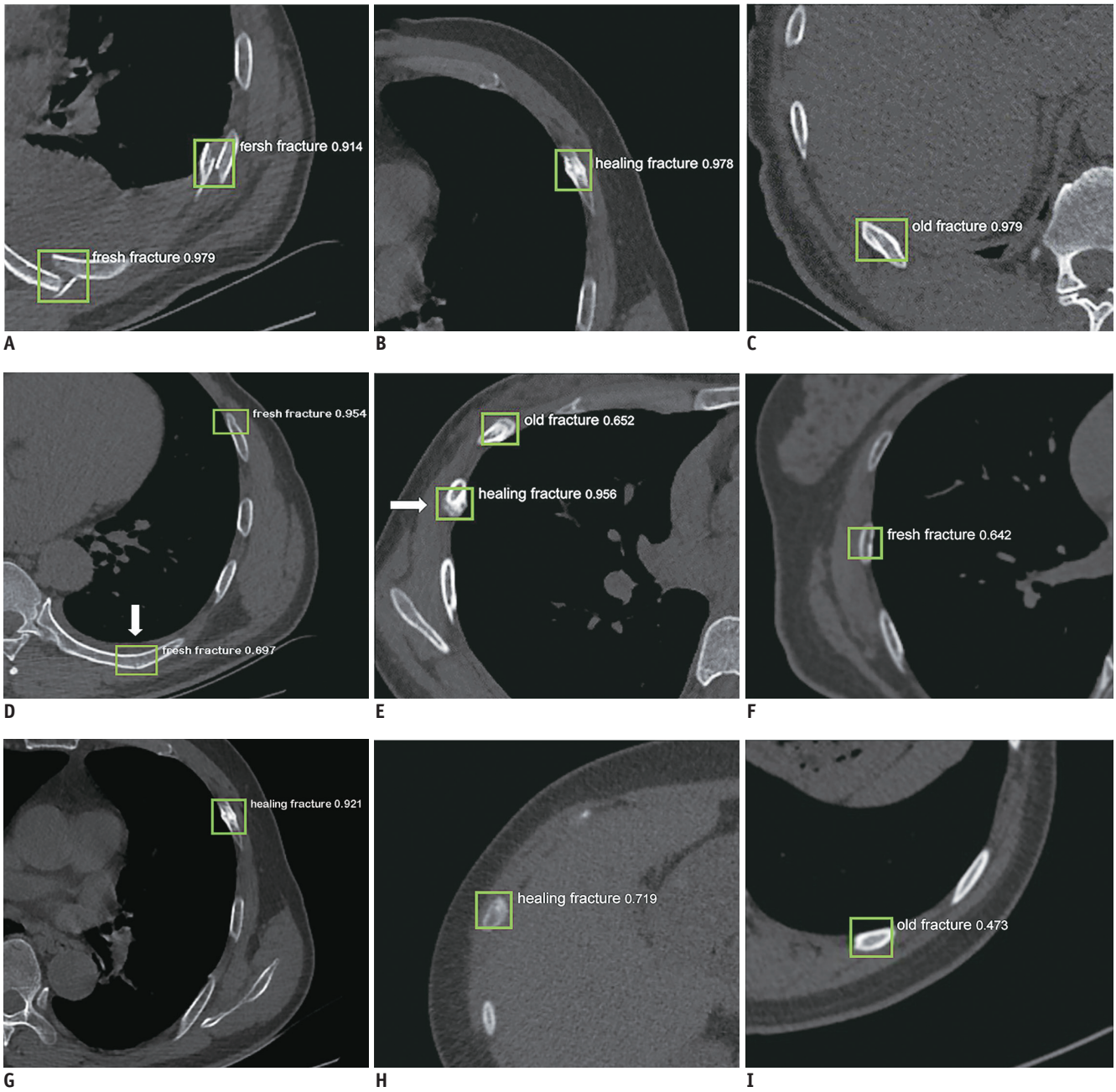


**Fig. 6. Detection/diagnosis results of different fractures shown on CT images.**

**A-C.** Rib fractures were detected/diagnosed by CNN model and radiologists correctly. **D.** Two fresh fractures were diagnosed by CNN model, while subtle fresh fracture in posterior rib was missed by some radiologists (arrow). **E.** These two healing fractures were misdiagnosed as old fractures by radiologists, and rear one was detected correctly by CNN model (arrow). **F-I.** FPs were detected on healthy ribs by CNN model. CNN = convolutional neural network


https://doi.org/10.3348/kjr.2019.0651

them from healthy ribs. White callus formation in healing fractures or fracture lines in fresh fractures were evident on CT images.

We used five different multicenter/multiparameter validation sets including diverse CT scanners, three reconstructed slice thicknesses (5, 1, and 2 mm) and two image resolutions (1024 x 1024 and 512 x 512 pixels). The results revealed that the robustness of the Faster R-CNN in these validation sets was good (mean F1-score > 0.8) except for validation set 5 (512 x 512 pixels). It was speculated that image resolution might affect image recognition. In a pilot experiment, we found that a 1024 x 1024-pixel resolution led to a better detection performance than a 512 x 512-pixel resolution as the 512 x 512-pixel images did not reflect the features of rib fractures well. In brief, this CNN model could be applied in different hospitals using various scanners, different reconstructed slice thicknesses, and high-resolution CT images in PACS.

In order to avoid incorrectly identifying a fracture in one location as many separate fractures, a program to merge the results and output a structured report was developed. It can reduce false negatives and FPs because the CNN model could output correct results identified correctly over a few layers and ignore predicted boxes only on one or two layers. Surprisingly, when we compared the diagnostic efficiency to that of experienced radiologists, the structured report performed comparably to the radiologists, and the diagnoses after AI assistance had the highest diagnostic efficiency. An analysis of the radiologists' diagnoses showed that they often missed diagnoses with multiple fractured ribs or subtle fractures. Some relatively confounding fractures, for instance, fresh fractures vs. healing fractures or healing fractures vs. old fractures, were also misdiagnosed. However, the Faster R-CNN extracted the feature map for each input image by means of a region proposal network and sliding-window M x N feature map (25). This led to the accurate detection of rib fractures in accordance with the diagnosis of radiologists, and the model could detect many subtle fractures that the radiologists missed. Moreover, the diagnosis time was obviously shortened after AI assistance. In the control set, FPs were inevitable, and the causes of misdiagnosis included the identification of uneven bone density/local defects as fresh fractures, identification of bone island/costal cartilage calcification as healing fractures, and local bone enlargement as old fractures.

Our preliminary study had several limitations. First, although our model could mark the fractures on CT images and output a structured report, the current model cannot show the anatomical location of the rib fractures (right or light, number of ribs, anatomical name of fractured rib). In future, a three-dimensional deep learning and tracking method may be used to identify the anatomical location. In addition, the precision and recall of this model were not particularly high, especially for old fractures, and some FPs existed in the ribs without fractures. We will increase the number of different shapes of fractures and introduce some common FP annotations into the model. Finally, the size of the validation set was relatively small, and additional data and prospective studies should be pursued to verify the CNN model.

In conclusion, our CNN model achieved fully automatic detection and classification of rib fractures and output structured reports. AI-assisted diagnosis attained a precision of 91.1% (increased by 10.8%) and sensitivity of 86.3% (increased by 23.9%), measurably surpassing the unaided work of experienced radiologists and requiring significantly less time. Furthermore, our method has a certain degree of generalizability, stability, and robustness based on the multicenter/multiparameter validation. In summary, our model suggest the feasibility of AI-assisted diagnosis of rib fractures, which could improve diagnostic efficiency, and reduce diagnosis time and radiologists' workload.

## Supplementary Materials

The Data Supplement is available with this article at https://doi.org/10.3348/kjr.2019.0651.

### Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

### Acknowledgments

The authors would like to especially acknowledge Qian Gao, MS and Yucai Li, MS (Ocean International Center E, Chaoyang Rd Side Rd, ShiLiPu, Chaoyang Qu, Beijing Shi) for their support in training and testing the convolutional neural network model.

### ORCID iDs

Hong Zhang
https://orcid.org/0000-0002-1770-6582
Qing-Qing Zhou
https://orcid.org/0000-0002-6699-9617

Korean Journal of Radiology
KJR

Rongguo Zhang
https://orcid.org/0000-0001-6566-8843
Xindao Yin
https://orcid.org/0000-0001-6958-5201
Bing Zhang
https://orcid.org/0000-0002-3953-0290

## REFERENCES

1. Cho SH, Sung YM, Kim MS. Missed rib fractures on evaluation of initial chest CT for trauma patients: pattern analysis and diagnostic value of coronal multiplanar reconstruction images with multidetector row CT. *Br J Radiol* 2012;85:e845-e850
2. Lin FC, Li RY, Tung YW, Jeng KC, Tsai SC. Morbidity, mortality, associated injuries, and management of traumatic rib fractures. *J Chin Med Assoc* 2016;79:329-334
3. Talbot BS, Gange CP Jr, Chaturvedi A, Klionsky N, Hobbs SK, Chaturvedi A. Traumatic rib injury: patterns, imaging pitfalls, complications, and treatment. *Radiographics* 2017;37:628-651
4. Liman ST, Kuzucu A, Tastepe AI, Ulasan GN, Topcu S. Chest injury due to blunt trauma. *Eur J Cardiothorac Surg* 2003;23:374-378
5. Murphy CE 4th, Raja AS, Baumann BM, Medak AJ, Langdorf MI, Nishijima DK, et al. Rib fracture diagnosis in the panscan era. *Ann Emerg Med* 2017;70:904-909
6. Langdorf MI, Medak AJ, Hendey GW, Nishijima DK, Mower WR, Raja AS, et al. Prevalence and clinical import of thoracic injury identified by chest computed tomography but not chest radiography in blunt trauma: multicenter prospective cohort study. *Ann Emerg Med* 2015;66:589-600
7. Kim J, Kim S, Kim YJ, Kim KG, Park J. Quantitative measurement method for possible rib fractures in chest radiographs. *Healthc Inform Res* 2013;19:196-204
8. Kim EY, Yang HJ, Sung YM, Hwang KH, Kim JH, Kim HS. Sternal fracture in the emergency department: diagnostic value of multidetector CT with sagittal and coronal reconstruction images. *Eur J Radiol* 2012;81:e708-e711
9. Sollmann N, Mei K, Hedderich DM, Maegerlein C, Kopp FK, Löffler MT, et al. Multi-detector CT imaging: impact of virtual tube current reduction and sparse sampling on detection of vertebral fractures. *Eur Radiol* 2019;29:3606-3616
10. van Laarhoven JJEM, Hietbrink F, Ferree S, Gunning AC, Houwert RM, Verleisdonk EMM, et al. Associated thoracic injury in patients with a clavicle fracture: a retrospective analysis of 1461 polytrauma patients. *Eur J Trauma Emerg Surg* 2019;45:59-63
11. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402-2410
12. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30-36
13. Choi JS, Han BK, Ko ES, Bae JM, Ko EY, Song SH, et al. Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography. *Korean J Radiol* 2019;20:749-758
14. Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM Trans Comput Biol Bioinform* 2018;15:1968-1978
15. Onishi Y, Teramoto A, Tsujimoto M, Tsukamoto T, Saito K, Toyama H, et al. Automated pulmonary nodule classification in computed tomography images using a deep convolutional neural network tained by generative adversarial networks. *Biomed Res Int* 2019;2019:6051939
16. Gao XW, Qian Y. Prediction of multidrug-resistant TB from CT pulmonary images based on deep learning techniques. *Mol Pharm* 2018;15:4326-4335
17. Zura R, Xu ZJ, Della Rocca GJ, Mehta S, Steen RG. When is a fracture not "fresh"? Aligning reimbursement with patient outcome after treatment with low-intensity pulsed ultrasound. *J Orthop Trauma* 2017;31:248-251
18. Wu X, Jiang Y. [Old fracture]. *Zhonghua Wai Ke Za Zhi* 2015;53:460-463
19. Wootton-Gorges SL, Stein-Wexler R, Walton JW, Rosas AJ, Coulter KP, Rogers KK. Comparison of computed tomography and chest radiography in the detection of rib fractures in abused infants. *Child Abuse Negl* 2008;32:659-663
20. Arredondo-Gómez E. [Treatment of traumatic clavicular pseudoarthrosis with the Hunec Colchero nail]. *Acta Ortop Mex* 2007;21:63-68
21. Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, et al. Speed/accuracy trade-offs for modern convolutional object detectors [updated April 2017]. Available at: https://arxiv.org/abs/1611.10012. Accessed June 21, 2019
22. Redmon J, Farhadi A. YOLOv3: an incremental improvement. Cornell University, 2018. Available at: https://arxiv.org/abs/1804.02767. Accessed June 22, 2019
23. Setio AAA, Traverso A, de Bel T, Berens MSN, Bogaard CVD, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Med Image Anal* 2017;42:1-13
24. Wang Y, Yan F, Lu X, Zheng G, Zhang X, Wang C, et al. IILS: intelligent imaging layout system for automatic imaging report standardization and intra-interdisciplinary clinical workflow optimization. *EBioMedicine* 2019;44:162-181
25. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks [updated January 2016]. Cornell University, 2015. Available at: https://arxiv.org/abs/1506.01497. Accessed June 22, 2019

 https://doi.org/10.3348/kjr.2019.0651 