IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 15, 2021, accepted May 15, 2021, date of publication May 24, 2021, date of current version June 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3082952

Fine-Tuned Deep Convolutional Networks for the Detection of Femoral Neck Fractures on Pelvic Radiographs: A Multicenter Dataset Validation

LIN MU[1], TAIPING QU[2], DONG DONG[1], XIULI LI[2], YUN PEI[3], YUCHONG WANG[4], GUANGYAO SHI[5], YONGRUI LI[1], FUJIN HE[2], AND HUIMAO ZHANG[1]

[1]Department of Radiology, The First Hospital of Jilin University, Changchun 130021, China
[2]Deepwise AI Lab, Deepwise Inc., Beijing 100080, China
[3]College of Electronic Science and Engineering, Jilin University, Changchun 130012, China
[4]Department of Radiology, Jilin Province FAW General Hospital (The Fourth Hospital of Jilin University), Changchun 130011, China
[5]Department of Radiology, The Jilin People's Hospital, Jilin 132012, China

Corresponding author: Huimao Zhang (huimaozhanglinda@163.com)

This work was supported in part by the Foundation of Jilin Provincial Department of Finance under Grant 2018SCZWSZX-02, in part by the Foundation of Health Commission of Jilin Province under Grant 2017J073, and in part by the Foundation of Scientific and Technological Developing Scheme of Jilin Province under Grant 2017C020.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee of the First Hospital of Jilin University under Approval No. 2020-362, and performed in line with the Declaration of Helsinki.

ABSTRACT In this study, we aim to provide a deep convolutional network based femoral neck fracture detection system on radiographs for emergency patients. We retrospectively collected 1,491 frontal pelvic radiographs from three institutions and assigned them to the following data sets: primary dataset (710 radiographs, to fine-tune and validate the initial model called the Digital Radiography Fracture Detection System [DR-FDS]), internal test set 1 (189 radiographs) and 2 (235 radiographs), and external test set 1 (189 radiographs) and 2 (168 radiographs). Per-bounding box recall and precision and per-image sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC) were computed. We randomly extracted 300 radiographs from the above test sets and compared their effect on the diagnostic accuracy and efficiency of fine-tuned model-assisted and unassisted clinicians. The fine-tuned DR-FDS showed a better overall performance in detecting femoral neck fractures than did the initial DR-FDS. The fine-tuned DR-FDS achieved AUC values of 0.9526 (95%CI, 0.9048–0.9767) and 0.9633(95%CI, 0.9346-0.9797) in internal test sets 1 and 2. In external test sets 1 and 2, this model also achieved promising results with AUC values of 0.9231 (95%CI, 0.8779–0.9520), and 0.9937 (95%CI 0.9739–0.9985), respectively. The clinicians showed a statistically significant increase in specificity, sensitivity, and accuracy for the identification of minimal/undisplaced fracture and a decrease in the average reading time. The object detection model that is fine-tuned has high sensitivity and specificity and the universal ability to detect and locate femoral neck fractures on pelvic radiographs.

INDEX TERMS Femoral neck fractures, convolutional neural network, radiographs, small sample, fine-tuning.

I. INTRODUCTION

The femur is the longest and strongest bone in the body. The femoral neck, the upper part of the femur, often suffers severe fractures [1]. Femoral neck fractures are generally considered as one of the most serious osteoporotic fractures, which are considered as the main cause of disability in elderly individuals worldwide [2], [3]. Fractures of the neck of the femur can seriously affect the quality of life of patients. Survivors may require considerable social and nursing care, which increases social and economic burdens [4]. Especially in emergency scenarios, timely and accurate diagnosis and

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Wang.

evaluation of femoral neck fractures are critical, and delayed surgical repair may lead to increased morbidity and mortality [5] Conventionally, frontal pelvic radiograph (PXR) is an economical and widely used tool to evaluate the location and types of the fractures. While displaced fractures diagnosis is a relatively plain task, minimal or undisplaced fractures are challenging for less experienced clinicians or radiologists with few musculoskeletal imaging experiences. Some previous studies showed that the initial misdiagnosis rate was as high as 7–14% [6], [7], and delayed diagnosis and treatment exacerbated the prognosis [8]. Therefore, development of an efficient femoral neck fracture detection system is of great significance for emergency radiology work.

Integrating artificial intelligence (AI) into the existing medical workflow is a very promising trend [9]. Deep learning (DL), which is an even more specialized subfield of AI, has shown great potential in medical imaging because of its high performance in diagnosis, classification, and prediction [10]–[12]. Several studies have demonstrated the application of DL for fracture detection. In a study by Olczak *et al.* [13], the deep convolutional neural network (DCNN) performed as well as or better than orthopedic surgeons in the detection of proximal humerus, hand, wrist, and ankle fractures on radiographs. A study by Adams *et al.* [14] suggests that as impressive as recognizing fractures is for a DCNN, similar learning can be achieved by top-performing medically-naive humans with less than 1 hour of perceptual training. Three studies revealed that DCNN not only detected fractures on radiographs but also used hot maps to localize fracture lesions and visualize the results [11], [15], [16]. Two studies developed excellent DL networks for detecting the fracture position; they use a bounding box to show the results instead of hot maps, which indicates a remarkable advancement in the field of fracture detection using AI [17], [18].

The purpose of the study is to detect and locate femoral neck fractures on radiographs by using a DL algorithm. Simulate clinical situation to explore the feasibility of applying a fine-tuned DCNN, and compare the accuracy and average radiograph reading time of doctors with two experience levels with or without the assistance of the model to further verify the clinical feasibility.

II. MATERIALS AND METHODS

A. DATASET

Three local ethics committees approved this study and waived the requirement for informed consent owing to the retrospective nature of this study. We retrospectively collected 1,491 patients (with only one PXR per patient) in the form of Digital Imaging and Communications in Medicine (DICOM) from three institutions' Picture Archiving and Communication Systems identified through the Radiology Information System. Among them, 710 PXRs (568 normal and 142 femoral neck fractures) were consecutively acquired from outpatients (OPs) in Radiology departments of First Hospital of Jilin University (JLU-1) between January 2016 and April 2019 and constituted the primary dataset. We randomly split the primary dataset into the training set (n = 610) which was used to fine-tune the DR-FDS, and tuning set (n = 100) which was used to select the final model. We also collected 189 consecutive PXRs (151 normal and 38 femoral neck fractures) from emergency patients (EPs) in Emergency Radiology department of JLU-1 between October 2018 and April 2019, which constituted internal test set 1. Additionally, in order to validate our model performance in real-world clinical environment, we split 253 consecutive PXRs from EPs in Emergency Radiology department of JLU-1 and considered them as internal test set 2, which were consisted of all kinds of pelvic fractures over a three-month period in 2020 (July to September), this data set was included to investigate the feasibility of extending this method to other types of proximal hip fractures. We acquired 189 (propensity matching normal, 115; femoral neck fractures, 74) and 168 (propensity matching normal, 75; femoral neck fractures, 93) PXRs from OPs in Radiology departments of Forth Hospital of Jilin University (JLU-4) and the Jilin People's Hospital (JLP) in 12 months period in 2019, respectively, and geographically split the two datasets as external test sets 1 and 2. We used all four datasets to test the model. We checked normal PXRs to confirm the absence of positive findings and extracted the demographic data, report data, and images of patients with hip trauma who underwent PXRs on the date of injury; we excluded postoperative images with internal fixation and arthroplasty. We used only one image/person to decrease the overperformance of model in each of training and test sets. A cross-sectional analysis to assess the functional outcome of femoral neck fractures in JLU-1. Using case-control analysis of femoral neck fracture of patients in JLU-4 and JLP Propensity matching for femoral neck fractures was performed by selecting non-fracture cases with a similar distribution of patients in the same hospital (controls). The cohort characteristics amongst the radiographs of all datasets are shown in Table 1.

B. OUTCOME LABLES

All these radiographs were assessed by two radiologists (Lin Mu and Dong Dong: 13 years and 15 years of experience, respectively) and were separated into the normal and fracture groups. All fracture locations were manually annotated with bounding boxes by the two radiologists in consensus, and these constituted the reference standards. Each image in the fracture group was assigned 2 binary diagnostic labels for the presence or absence of (1) displaced fracture and (2) minimal/undisplaced fracture (defined: angulation $< 15°$, tilt $< 20°$, axial neck of femur shortening or the distance of the displaced bone chip < 5 mm).

C. IMAGE PREPROCESSING

The DL networks directly accepted each image without cropping, which helped retain maximum information. During the model training process, all input data were augmented to

TABLE 1. Cohort characteristics of different dataset.

	JLU-1			JLU-4	JLP	JLU-1+JLU4+JLP
	Primary dataset (OPs)	Internal test set 1 (EPs)	Internal test set 2 (EPs)	External test set 1 (OPs)	External test set 2 (OPs)	Clinical test dataset (OPs+EPs)
No. of radiographs	710	189	235	189	168	300
Sampling	Cross-sectional	Cross-sectional	Cross-sectional	Case-control	Case-control	Case-control
Propensity matching	NA	NA	NA	Femoral neck fractures	Femoral neck fractures	Femoral neck fractures
Partition	Train	Test	Test	Test	Test	Test
No. of scanners	2	2	2	4	2	7
No. of scanner manufacturers	2	2	2	3	1	4
Age, mean (SD), years	54.4 (21.5)	55.7 (21.8)	51.9 (23.0)	53.5 (19.9)	58.1 (20.3)	56.59 (20.6)
Female frequency, no. (%)	266 (37.5)	65 (34.4)	135 (57.4)	103 (54.5)	104 (61.9)	164 (54.7)
Normal, no. (%)	568 (80.0)	151 (79.9)	158 (67.2)	74 (39.2)	93 (55.4)	145 (48.3)
Hip fracture, no. (%)	NA	NA	77 (32.8)	NA	NA	NA
Femoral neck fracture, no. (%)	142 (20.0)	38 (20.1)	NA	115 (60.8)	75 (44.6)	155 (51.7)

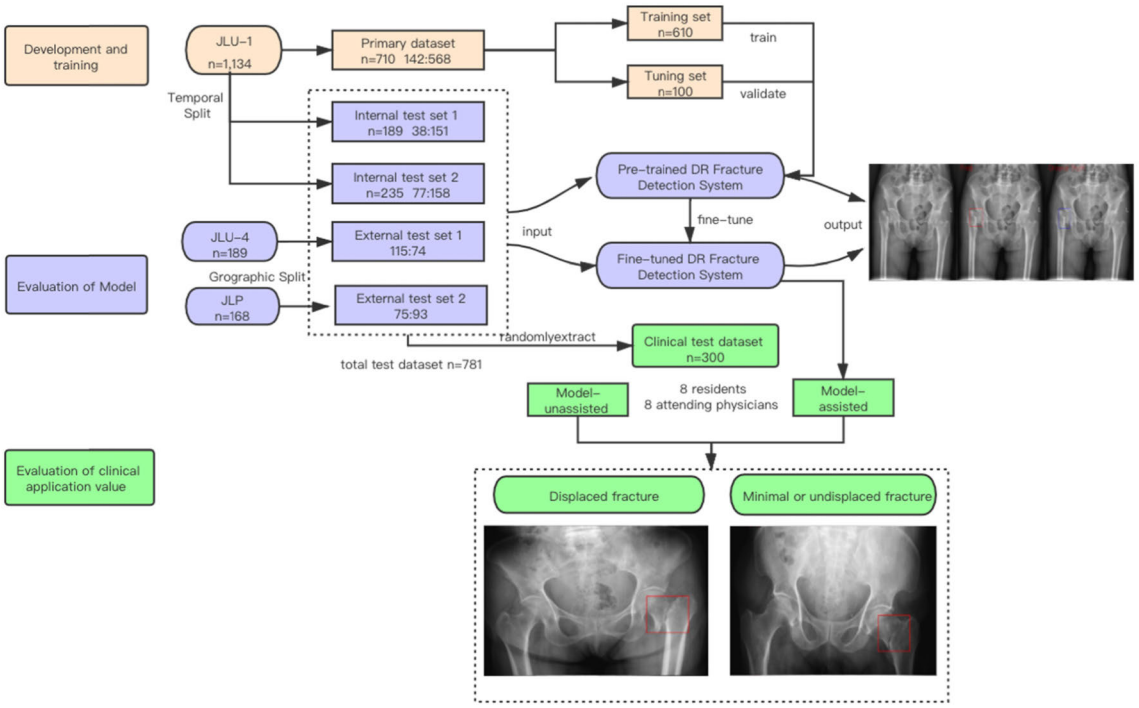

FIGURE 1. Flow chart of the whole research design. The development and evaluation process of DR-FDS was shown in orange and purple shade. The evaluation process of clinical application value was shown in green shade.

improve the accuracy of the proposed method for fracture detection. In this study, we employed the two augmentation methods: (i) flipping: flips input radiographs horizontally with a probability of 0.5; (ii) rotating: rotates input radiographs by $-180°$ to $180°$.

D. FRACTURE DETECTION MODEL TRAINING DETAILS

Figure 1 shows the flow chart of the whole experiment. The DR-FDS was inspired by Multi-domain Fracture Detection Network (MFDN) [19], which was improved on the basis of Faster R-CNN [20]. The DR-FDS was pre-trained with 7338 X-ray images of the wrist, feet, hand, ankle, elbow, shoulder, hip, and knee. Because the current task involves only the detection of hip fracture, we preserved only the fracture detection network of MFDN and removed the domain classification network of MFDN. Therefore, as shown in Figure 2 (A), this model was mainly composed of three structures: backbone, feature enhancement modules (FEMs), region proposal network (RPN), and R-CNN. The function of the backbone network was to extract advanced features from input radiographs. In our model, a pre-trained Faster R-CNN with ResNet-50 [21] was introduced as the backbone network. The input of the backbone was a series of pre-processed radiographs. Figure 2 (B) shows the architecture of FEMs, which contained two consecutive sub-modules. Firstly, FEM-1 was added to the output of multi-scale layers (A2, A3, A4, A5) in the backbone to form new multi-scale layers (M1, M2, M3, M4). FEM-1 was used to implement multi-scale predictions so that the RPN could be used at different scales (B1, B2, B3, B4, B5) to generate region proposals. Then, followed FEM-1, FEM-2 formed new multi-scale layers (N1, N2, N3, N4, N5) and further extracted more

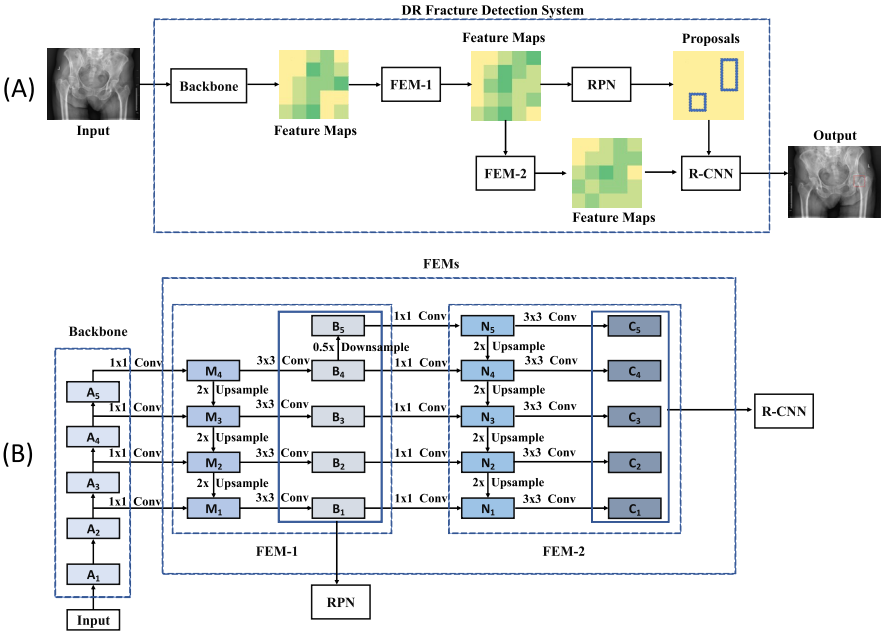

FIGURE 2. (A) is the overview of DR fracture detection system, and (B) is the structure of the feature enhancement modules.

valuable multi-scale information (C1, C2, C3, C4, C5) to enhance feature expression. Finally, the enhanced multi-scale features were input to R-CNN to obtain the final femoral neck fracture detection results. The loss function consists of the RPN loss and the R-CNN loss. It is defined as follows:

$$Loss = \zeta \cdot Loss_{RPN} + \eta \cdot Loss_{R-CNN} \tag{1}$$

where ζ and η are the weights. ζ is set to 1 and η is set to 0.1. Both losses contain regression loss and classification loss. The regression loss is designed as a smoothed-L1 loss to evaluate the accuracy of fracture area detection, and the classification loss is a cross-entropy function to assess the accuracy of the fracture area classification. Before fine-tuning the model, we first test the training data with the original DR-FDS, adding the positive samples with detection errors to the training set as difficult samples and input them into the original DR-FDS for fine-tuning. Initially, we also add the negative samples with detection errors to the training set as difficult samples. However, the detection results of the model were more biased to miss detection, thereby ensuring a lower probability of false detection. For fracture detection tasks, a lower missed rate is more clinically significant. Therefore, we eliminated these difficult samples generated by negative samples during the training process. The fine-tuning process of our model can be divided into two steps. First, we fixed the parameters of all layers; subsequently, were reduced the learning rate by 10 times to train for 50 epochs and selected the optimal model parameters according to the results of the validation set.

E. EVALUATION OF THE FRACTURE DETECTION MODEL

Before and after fine-tuning, the DR-FDS was tested using the previously unseen four test sets. We set two levels (per-image level and per-bounding box level) to evaluate the performance of CNN. In the per-image level, the DL model will output the result of whether or not there is a fracture in the image regardless of the specific fracture location. The true-positive determination required at least one true-positive fracture mark on the image. In the per-bounding box level, the Intersection over Union (IoU) from the bounding box predicted by the DL model and the reference standard bounding box was calculated to determine whether there is a possibility of a fracture. For example, when one region of interest localized by the CNN as the fracture with more than 50% probability overlapped with a reference box, the network would output the true-positive mark; the other annotations by the CNN were considered false-positive.

F. OBSERVER EVALUATION AND COMPARISON WITH FRACTURE DETECTION MODEL

We evaluate the utility of our fine-tuned model, we measured its effect on the diagnostic accuracy and efficiency of clinicians. We recruited 16 doctors from the departments of radiology, osteoarthrosis surgery, and emergency, including 8 residents (experience: 1–3 years) and 8 chief physicians (experience: 5–10 years). This study was performed to ensure a multicenter comparison and also to save the time cost of 16 doctors in outlining the fracture areas. Therefore, we randomly selected 300 x-ray sequences from all test sets except the internal test set 2 to explore the value of the practical clinical application. All the clinicians evaluated the clinical test dataset, unaided by the model, and checked the annotation with bounding boxes on the fracture areas. After a 30-day washout period, they reviewed the same test dataset with the assistance of the fine-tuned DR-FDS and added or removed bounding boxes on the images. Then,

TABLE 2. Performance of DR-FDS model and fine-tuned DR-FDS. (CI:95% confidence interval; M: Median).

	Model	Sensitivity (%)	Specificity（%）	AUC	Recall (%)	Precision (%)
			per-image level		per-bounding box	
Internal test set 1 （n=189）	DR-FDS	96.59 (90.36-99.29)	78.22 (68.90-85.82)	0.9646 (0.9294-0.9824)	95.56 (89.01-98.78)	73.50 (64.55-81.23)
	Fine-tuned DR-FDS	97.73 (92.03-99.72)	81.19 (72.19-88.28)	0.9526 (0.9048-0.9767)	95.56 (89.01-98.78)	81.13 (72.38-88.08)
	P value	0.500	0.600	0.279	0.064	0.176
Internal test set 2 （n=235）	DR-FDS	98.72 （96.17-100.00）	65.23 (57.82-72.65)	0.9651 (0.9350-0.9824)	91.68 (86.08-97.29)	65.29 (57.83-72.68)
	Fine-tuned DR-FDS	98.72 （96.17-100.00）	73.40 (66.55-80.36)	0.9633 (0.9346-0.9797)	87.48 (80.90-94.08)	73.45 (66.50-80.39)
	P value	1.000	<0.001	0.998	<0.001	<0.001
External test set 1 (n=189)	DR-FDS	71.62 (59.95-81.50)	54.78 (45.23-64.08)	0.6996 (0.6107-0.7712)	62.16 (50.13-73.19)	60.53 (48.65-71.56)
	Fine-tuned DR-FDS	91.89 (83.18-96.97)	74.78 (65.83-82.42)	0.9231 (0.8779-0.9520)	90.54 (81.48-96.11)	81.71 (71.63-89.38)
	P value	0.001	0.001	<0.001	<0.001	0.003
External test set 2 （n=168）	DR-FDS	98.92 (94.15-99.97)	60.00 (48.04-71.15)	0.9732 (0.9470-0.9865)	97.85 (92.45-99.74)	77.12 (68.48-84.35)
	Fine-tuned DR-FDS	98.92 (94.15-99.97)	80.00 (69.17-88.35)	0.9937 (0.9739-0.9985)	98.92 (94.15-99.97)	82.88 (74.57-89.37)
	P value	0.751	0.008	0.021	0.500	0.277

we divided these images into two subgroups according to the diagnostic labels: displaced fracture subgroup (normal, 81; displaced fractures, 79) and minimal/undisplaced fracture subgroup (normal, 74; minimal/undisplaced fractures, 66); we performed a statistical analysis of the subgroup results.

G. STATISTICAL ANALYSIS

Statistical analysis was conducted with SPSS 25.0. (SPSS Inc, Chicago, IL, USA). We determined per-image sensitivity, specificity, accuracy, and AUC analysis, and determined per-bounding box recall and precision. We added the average reading time of clinicians. Comparison of AUCs of the model before and after fine-tuning in multicentric validation set were performed using DeLong's Test. Statistically significant differences in sensitivity, specificity, accuracy, recall, precision, and average reading time were evaluated using $\chi 2$ analysis and Fisher's exact probability. A p value of < 0.05 was considered to indicate a statistically significant difference.

III. RESULTS

A. FRACTURE DETECTION MODEL PERFORMANCE

The sensitivity, specificity, and AUC at the per-image level and the recall and precision at the per-bounding box level in the model before and after fine-tuning are shown in Table 2, for the four test sets. The best compromise between sensitivity and specificity was observed at a cut-off threshold of 0.4. The performance of DR-FDS before and after fine-tuning showed no statistically significant difference in internal test set 1 in terms of all metrics. In internal test set 2, the fine-tuned DR-FDS showed a good performance in terms of specificity (0.7340; 95%CI, 0.6655–0.8036), AUC (0.9633; 95%CI, 0.9346–0.9797). The fine-tuned model showed statistically significant improvement in the detection of fractures

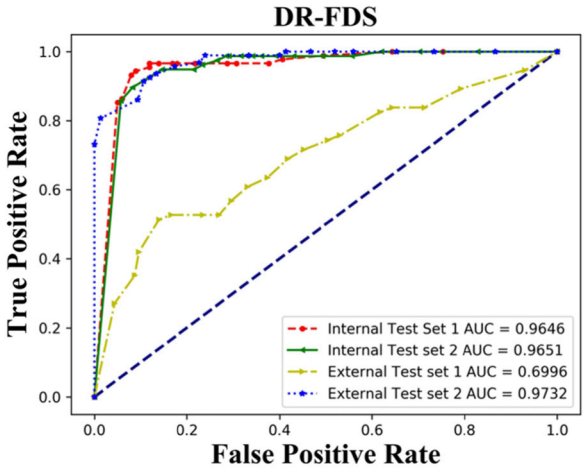

FIGURE 3. The area under the receiver operating characteristic (ROC) curves for detection of fracture using DR-FDS at the per-image level.

in external test set 1, on all metrics. With regard to the performance of the fine-tuned model in external test set 2, there were statistically significant increases in the specificity and AUC at the per-image level. The specificity increased from 0.6000 (95%CI, 0.4804–0.7115) to 0.8000 (95%CI, 0.6917–0.8835), and AUC increased from 0.9732 (95%CI, 0.9470–0.9865) to 0.9937 (95%CI, 0.9739–0.9985). The receiver operating characteristic (ROC) curves for the detection of fracture at per-image level are shown in Figure 3 (DR-FDS) and Figure 4 (fine-tuned DR-FDS). Examples of performance are shown in Figure 5 and Figure 6.

B. OBSERVER PERFORMANCE AND COMPARISON WITH FRACTURE DETECTION MODEL

A comparison of unassisted and model-assisted performance metrics of clinicians on two types of fractures is shown

TABLE 3. Comparison of unassisted and model-assisted performance metrics of clinicians on two types of fracture (CI:95% confidence interval; M: Median).

Prediction /Subgroup	*Displaced fracture*			*Minimal or undisplaced fracture*		
micro-average	Unassisted	Model-assisted	*p* value	Unassisted	Model-assisted	*p* value
Specificity (95%CI)	0.9219(0.9124-0.9314)	0.9584(0.9513-0.9655)	0.383	0.8909(0.8798-0.9020)	0.9239(0.9145-0.9333)	0.014
Sensitivity(95%CI)	0.93.92(0.9307-0.9477)	0.9734(0.9677-0.9791)	0.313	0.8237(0.8102-0.8372)	0.8854(0.8741-0.8967)	<0.001
Accuracy(95%CI)	0.9.06(0.9240-0.9372)	0.9659(0.9613-0.9705)	0.187	0.8571(0.8476-0.8666)	0.9046(0.8968-0.9124)	<0.001
Reading Time(second) M(P_{25}-P_{75})	14(9-26)	12(8-18)	<0.001	15(10-26)	10(7-16)	<0.001

TABLE 4. Comparison of unassisted and model-assisted performance metrics of residents, chief physicians and DR-FDS on clinical test set. (CI:95% confidence interval; M: Median).

Metric Total clinical test set (n=300 PXRs)	Residents(n=8)			Chief physicians(n=8)			DR-FDS
	Unassisted	Model-assisted	p value	Unassisted	Model-assisted	p value	
Specificity [95%CI]	0.8906 (0.8815-0.8997)	0.9108 (0.9025-0.9191)	0.964	0.9655 (0.9602-0.9708)	0.9714 (0.9666-0.9762)	<0.001	0.8311 (0.7707-0.8914)
Sensitivity [95%CI]	0.8258 (0.8148-0.8368)	0.9276 (0.9201-0.9351)	<0.001	0.9386 (0.9316-0.9456)	0.9478 (0.9413-0.9543)	<0.001	0.9660 (0.9367-0.9953)
Accuracy [95%CI]	0.8582 (0.8505-0.8659)	0.9192 (0.9134-0.9250)	<0.001	0.9520 (0.9475-0.9565)	0.9596 (0.9555-0.9637)	<0.001	0.8983 (0.8638-0.9328)
Reading Time(second) M(P_{25}-P_{75})	20(14-32)	13(8-20)	<0.001	10(8-14)	8(6-12)	<0.001	——

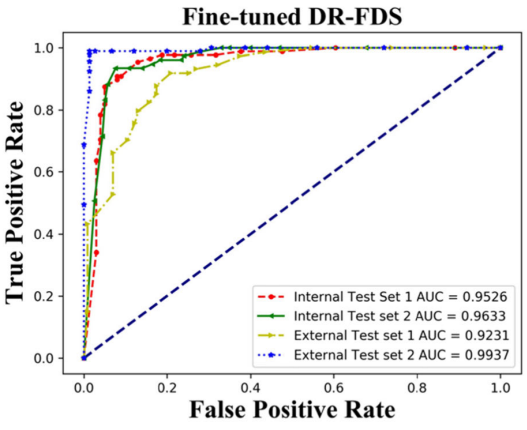

FIGURE 4. The area under the receiver operating characteristic (ROC) curves for detection of fracture using fine-tuned DR-FDS at the per-image level.

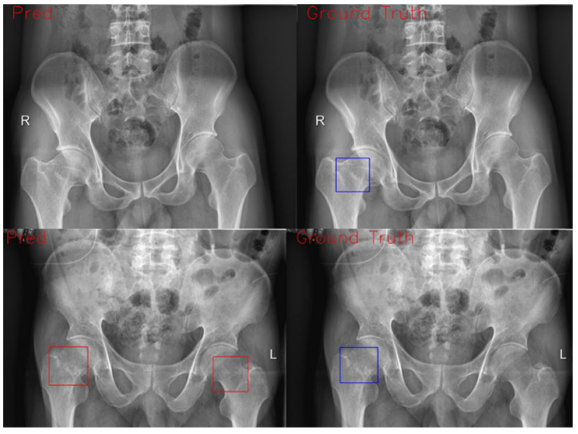

FIGURE 5. Radiographs show selected true-positive examples of femoral neck fracture. Red boxes are made by the fine-tuned DR-FDS to detect and localize fractures. IOU means Intersection of the Union.

in Table 3. All the fine-tuned model-assisted clinicians showed statistically significant increases fracture identification indices, namely specificity, sensitivity, and accuracy, in the minimal/undisplaced fracture subgroup; their average reading time decreased from 15 s to 10 s. The fine-tuned model-assisted clinicians achieved significantly higher accuracy (0.9046; 95%CI, 0.8968-0.9124) and spent lesser time on reading each radiograph than did the unassisted clinicians in the displaced fracture subgroup. The numerical values of the clinical utility of the model to the two different kinds of clinicians are provided in Table 4. The fine-tuned model-assisted residents showed a significant increase in the diagnostic sensitivity (0.9276; 95%CI, 0.9201–0.9351), and accuracy (0.9192; 95%CI, 0.9134–0.9250); there was no statistically significant improvement in specificity (0.9108; 95%CI, 0.9025–0.9191). The average reading time of the residents decreased from 20 s to 13 s. The fine-tuned model-assisted chief physicians also achieve statistically significant improvement in all the performance metrics.

IV. DISCUSSION

This study involved a preliminary examination of the diagnostic value of DL in the detection of the neck of femur fractures. Accurate and efficient detection of femoral neck fractures is essential for clinical diagnosis. Since the limitations associated with the human eye's observation power, rapid identification of uneven or nondisplaced femoral neck fractures is challenging. The experimental results of this study suggest that DCNN can accurately detect displaced

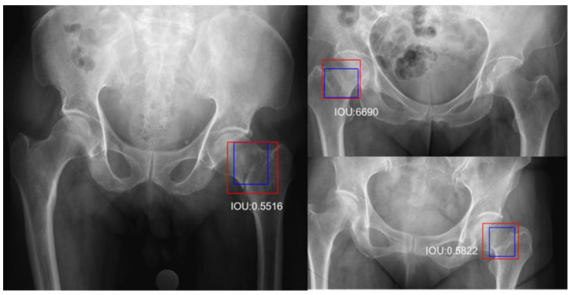

FIGURE 6. Radiographs show selected false-negative example (above two radiographs. The model had lower sensitivity for minimal or undisplaced fracture), false-positive example (below two radiographs, the model had lower specificity for poor exposure of femoral neck). Red boxes are made by the fine-tuned DR-FDS, and blue boxes are made by radiologists as the ground truth.

fractures and has good performance for minimal or non-displaced fractures as well. Besides, DCNN will help clinicians lower the misdiagnosis rate and prevent subsequent misdiagnosis events. We also show that fine-tuned the generic fracture detection system using a small number of location-specific fracture images can achieve more accurate detection performance for location-specific fractures.

The current techniques for object detection are mainly divided into one-stage and two-stage object detection networks. The framework used in this study is based on a two-stage object detection network Faster R-CNN instead of a one-stage object detection network, such as YOLO. The reason is twofold, firstly, medical images are extremely imbalanced between positive and negative samples, most of which are negative. The two-stage object detection network can alleviate this sample imbalance to some extent. Secondly, the two-stage object detection network can also be viewed as a cascade process, which uses RPN to remove a large amount of background information so that the network can focus on differentiating lesion information.

In previous studies, single detection tasks as an image classification problem in radiographs were commonly used in DCNN research [13], [22], the features extracted by the classified model pay considerable attention to high-level semantics. However, we used a DL object detection network for image analysis. The task of object detection involves two fundamental questions, semantic information and the location information of the target. Cheng *et al.* [11] used DCNN for the detection of hip fractures and with a low false-negative rate, which is noninferior to the performance of the experts. Mutasa *et al.* [23] explored two data augmentation approaches to improve the detection of femoral neck fractures and focused on Garden fracture classification. Krogue *et al.* [24] investigated the use of DL for automatic identification and classification of hip fractures and found that DL improved outcomes by reducing diagnostic errors. In our study, we used whole PXRs for training and testing, and the model output the PXRs which have one or more boxes to visualize the regions as the fracture sites as a detection result. Visualization of the DCNN may convince doctors to accept the results and make the results become explicable. In addition, another advantage of DL object detection is the ability to review the false-positive and false-negative cases, which helps refine the training of the network for the detection of image features. However, the diagnostic ability of DL cannot be overstated, and there is great scope for improvement in the field of auxiliary physician diagnosis.

We trained a fine-tuned model on a relatively small dataset (710 PXRs) and found that it improved the ability to detect fractures in specific locations. Because the fine-tuned model was specifically optimized for the hip joint, fine-tuning strategies can effectively extract task-specific characteristics from a small amount of training data. At the same time, we collected the continuous data that simulates real clinical scenarios. The fine-tuned DR-FDS improves the detection performance of femoral neck fractures. Particularly, the AUC value increased significantly from 0.6971 to 0.9216 in external test 1 and from 0.9732 to 0.9937 in external test 2; the highest AUC value was found in external test 2. Our study demonstrates that a limited amount of data and specific and homogeneous datasets can help the DL networks achieve high-level automated detection performance, and this finding indicates that our DR-FDS can be used in clinical practice, simply after fine-tuning.

Verifying how well the model performs in multicentric validation datasets is a major challenge. The result in internal test set 2 showed that the fine-tuned DR-FDS is able to identify and classify any kind of fracture in the pelvis, which includes femoral neck fracture, intertrochanteric fracture, and fracture of the ilium, ischium, pubis, and acetabulum. However, there is no better than the original model in identifying multitype of fractures (such as internal test set 2). Considering that the fine-tuned DR-FDS was optimized for femoral neck fractures, the original model can be trained in the same way for any kind of fracture. In the future, we can develop different optimization strategies to adjust and enhance the fracture detection ability of the model according to different location of the fracture. In the external test sets, the fine-tuned DR-FDS presented a statistically significant improvement of all metrics in external test set 1, and the specificity and AUC of the fine-tuned model in external test set 2 demonstrated a significantly superior capacity in detecting fractures. On the one hand, the reliability of external verification shows that the generalization ability of the model is improved through limited sample fine-tuning. In particular, the results of the external test set 2 are comparable to those of the internal test set 1, which also indicates the better robustness of the model; On the other hand, this result revealed a good prospect for the generalized application of our model in clinical work.

The purpose of the fracture detection system is to improve the diagnostic accuracy of practicing clinicians, but not simply to achieve the highest AUC possible. The reason behind designing the two types of fracture subgroups was to observe how the fine-tuned DR-FDS performed at different levels of diagnostic difficulty. We considered that minimal/undisplaced fractures were easy to misdiagnosis and not easily detected in daily work. Fine-tuned model-assisted

clinicians were significantly more sensitive and specific in detecting minimal/undisplaced fractures than were unassisted clinicians. This result revealed the clinicians' diagnostic ability improved further in difficult cases, with the assistance of the computer. Fracture detection models can be of great benefit to clinicians, and our results showed that there was statistically significant improvement in the performance metrics. It suggests that not only experienced doctors (such as chief physicians), but also inexperienced doctors (such as residents), the both will benefit from the fine-tuned DR-FDS. The added value of this model to assist primary doctors can increase confidence and reduce the occurrence of missed diagnoses. Collectively, the fine-tuned DR-FDS can assist doctors who lack experience in diagnosing difficult cases.

V. STUDY LIMITATIONS

There are several limitations to this preliminary study. First, although the results of this study are promising, applying this automatic detection algorithm into clinical work to increase the detection rate of femoral neck fractures presents a great challenge. A randomized, prospective study should be conducted to evaluate the clinical impact on the diagnostic accuracy and economic value of DCNN for identifying any other types of fractures on radiographs, in addition to femoral neck fractures. Second, the diagnostic accuracy of clinicians and the model in this study is limited to determining what is visible within a radiograph. In future studies, clinical studies with a full range of clinical information such as medical history and physical examination can be included. Third, the subtypes of femoral neck fractures are not explored in this study. Future studies will introduce multi-task learning to further distinguish the subtypes of femoral neck fractures.

VI. CONCLUSION

In conclusion, the object detection DCNN that is fine-tuned with a small dataset has high sensitivity and specificity and the universal ability to detect and locate femoral neck fractures on pelvic radiographs. Our fracture detection system can assist doctors who lack work experience, especially in evaluating difficult cases.

ACKNOWLEDGMENT

The authors are grateful to the Deepwise Corporation for their help with the deep learning algorithm used in this research. *(Lin Mu and Taiping Qu contributed equally to this work.)*

REFERENCES

[1] M. Kukar, I. Kononenko, and T. Silvester, "Machine learning in prognosis of the femoral neck fracture recovery," *Artif Intell Med*, vol. 8, no. 5, pp. 431–451, Oct. 1996.
[2] K. K. Kani, J. A. Porrino, H. Mulcahy, and F. S. Chew, "Fragility fractures of the proximal femur: Review and update for radiologists," *Skeletal Radiol.*, vol. 48, no. 1, pp. 29–45, Jan. 2019.
[3] T. Sözen, L. Özışık, and N. Başaran, "An overview and management of osteoporosis," *Eur. J. Rheumatol.*, vol. 4, no. 1, pp. 46–56, Mar. 2017.
[4] S. M. Dyer, M. Crotty, N. Fairhall, J. Magaziner, L. A. Beaupre, I. D. Cameron, and C. Sherrington, "A critical review of the long-term disability outcomes following hip fracture," *BMC Geriatrics*, vol. 16, no. 1, p. 158, Sep. 2016.
[5] D. Ryan, H. Yoshihara, D. Yoneoka, K. Egol, and J. D. Zuckerman, "Delay in hip fracture surgery: An analysis of patient-specific and hospital-specific risk factors," *J. Orthopaedic Trauma*, vol. 29, no. 8, pp. 343–348, 2015.
[6] W. B. Chellam, "Missed subtle fractures on the trauma-meeting digital projector," *Injury*, vol. 47, no. 3, pp. 674–676, Mar. 2016.
[7] D. K. Hakkarinen, K. V. Banh, and G. W. Hendey, "Magnetic resonance imaging identifies occult hip fractures missed by 64-slice computed tomography," *J. Emerg. Med.*, vol. 43, no. 2, pp. 303–307, Aug. 2012.
[8] S. M. Tarrant, B. M. Hardy, P. L. Byth, T. L. Brown, J. Attia, and Z. J. Balogh, "Preventable mortality in geriatric hip fracture inpatients," *Bone Joint J.*, vol. 96, no. 9, pp. 1178–1184, Sep. 2014.
[9] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Rev. Cancer*, vol. 18, no. 8, pp. 500–510, 2018.
[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[11] C.-T. Cheng, T.-Y. Ho, T.-Y. Lee, C.-C. Chang, C.-C. Chou, C.-C. Chen, I.-F. Chung, and C.-H. Liao, "Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs," *Eur. Radiol.*, vol. 29, no. 10, pp. 5469–5477, Oct. 2019.
[12] S. Derkatch, C. Kirby, D. Kimelman, M. J. Jozani, J. M. Davidson, and W. D. Leslie, "Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: A registry-based cohort study of dual X-ray absorptiometry," *Radiology*, vol. 293, no. 2, pp. 405–411, Nov. 2019.
[13] J. Olczak, N. Fahlberg, A. Maki, A. S. Razavian, A. Jilert, A. Stark, O. Sköldenberg, and M. Gordon, "Artificial intelligence for analyzing orthopedic trauma radiographs," *Acta Orthopaedica*, vol. 88, no. 6, pp. 581–586, Nov. 2017.
[14] M. Adams, W. Chen, D. Holcdorf, M. W. McCusker, P. D. Howe, and F. Gaillard, "Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures," *J. Med. Imag. Radiat. Oncol.*, vol. 63, no. 1, pp. 27–32, Feb. 2019.
[15] R. Lindsey, A. Daluiski, S. Chopra, A. Lachapelle, M. Mozer, S. Sicular, D. Hanel, M. Gardner, A. Gupta, R. Hotchkiss, and H. Potter, "Deep neural network improves fracture detection by clinicians," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 45, pp. 11591–11596, Nov. 2018.
[16] J. S. Yu, S. M. Yu, B. S. Erdal, M. Demirer, V. Gupta, M. Bigelow, A. Salvador, T. Rink, S. S. Lenobel, L. M. Prevedello, and R. D. White, "Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: Proof of concept," *Clin Radiol*, vol. 75, no. 3, pp. 237.e1-237.e9, Mar. 2020.
[17] Y. L. Thian, Y. Li, P. Jagmohan, D. Sia, V. E. Y. Chan, and R. T. Tan, "Convolutional neural networks for automated fracture detection and localization on wrist radiographs," *Radiol., Artif. Intell.*, vol. 1, no. 1, Jan. 2019, Art. no. e180001.
[18] K. Gan, D. Xu, Y. Lin, Y. Shen, T. Zhang, K. Hu, K. Zhou, M. Bi, L. Pan, W. Wu, and Y. Liu, "Artificial intelligence detection of distal radius fractures: A comparison between the convolutional neural network and professional assessments," *Acta Orthopaedica*, vol. 90, no. 4, pp. 394–400, Jul. 2019.
[19] S. Wu, L. Yan, X. Liu, Y. Yu, and S. Zhang, "An end-to-end network for detecting multi-domain fractures on X-ray images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 448–452.
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[22] S. W. Chung, S. S. Han, J. W. Lee, K.-S. Oh, N. R. Kim, J. P. Yoon, J. Y. Kim, S. H. Moon, J. Kwon, H.-J. Lee, Y.-M. Noh, and Y. Kim, "Automated detection and classification of the proximal humerus fracture by using deep learning algorithm," *Acta Orthopaedica*, vol. 89, no. 4, pp. 468–473, Jul. 2018.
[23] S. Mutasa, S. Varada, A. Goel, T. T. Wong, and M. J. Rasiej, "Advanced deep learning techniques applied to automated femoral neck fracture detection and classification," *J. Digit. Imag.*, vol. 33, no. 5, pp. 1209–1217, Oct. 2020.
[24] J. D. Krogue, K. V. Cheng, K. M. Hwang, P. Toogood, and V. Pedoia, "Automatic hip fracture identification and functional subclassification with deep learning," *Radiol., Artif. Intell.*, vol. 2, no. 2, 2020, Art. no. e190023.

LIN MU received the B.S. degree in clinical medical course and the M.S. degree in medical imaging and nuclear medicine from the School of Medicine, Jilin University, Changchun, China, in 2010 and 2013, respectively. She currently works as an Attending Doctor with the Department of Diagnostic Radiology, The First Hospital of Jilin University, Changchun. Her research interests include body imaging and medical data analysis.

TAIPING QU received the B.S. degree from the Changchun University of Technology, Changchun, China, in 2017, and the M.S. degree from Jilin University, Changchun, in 2020. His main research interests include medical image analysis, computer vision, machine learning, and artificial intelligence.

DONG DONG received the B.S. degree in clinical medical course and the M.S. degree in medical imaging and nuclear medicine from the School of Medicine, Jilin University, Changchun, China, in 2003 and 2005, respectively. She currently works as an Attending Doctor with the Department of Diagnostic Radiology, The First Hospital of Jilin University, Changchun. Her research interests include body imaging and musculoskeletal imaging medical image data analysis

XIULI LI received the Ph.D. degree in medical image analysis from the Institute of Automation, Chinese Academy of Sciences, in 2014. She has been a Senior Researcher Staff of medical image analysis with the Cognitive Medical Group, IBM China Research Laboratory, and an Algorithm Engineer with the GE Medical Research Center, China. Her primary research interests include medical image analysis, computer vision, machine learning, and artificial intelligence.

YUN PEI received the B.E. degree in electronic information engineering from Jilin University, Changchun, China, in 2019, where he is currently pursuing the M.E. degree in circuit and system. From 2017 to 2018, he worked as an Intern with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Since 2019, he has been an Intern with the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China. His current research interests include the next-generation integrating driverless systems and the medical image analysis using deep learning.

YUCHONG WANG received the B.S. degree in clinical medical course from the Liaoning Medical College, Jinzhou, China, in 2009, and the M.S. degree in medical imaging and nuclear medicine from the School of Medicine, Jilin University, Changchun, China, in 2013. He currently works as an Attending Doctor with the Department of Diagnostic Radiology, The Jilin Province FAW General Hospital, Changchun. His research interest includes musculoskeletal imaging.

GUANGYAO SHI received the B.S. degree in clinical medical course from Beihua University, Jilin, China, in 2002, and the M.S. degree in medical imaging and nuclear medicine from the School of Medicine, Jilin University, Changchun, China, in 2011. He currently works as an Attending Doctor with the Department of Diagnostic Radiology, The Jilin People's Hospital, Jilin. His research interests include body imaging and medical image data analysis.

YONGRUI LI received the B.S. degree in clinical medical course and the M.S. degree in medical imaging and nuclear medicine from the School of Medicine, Jilin University, Changchun, China, in 2012 and 2015, respectively. He currently works as a Resident with the Department of Diagnostic Radiology, The First Hospital of Jilin University, Changchun. His research interest includes the musculoskeletal imaging.

FUJIN HE received the B.S. degree from Zhejiang Normal University, Zhejiang, China, in 2016, and the M.S. degree from Xiamen University, Xiamen, China, in 2019. His main research interests include machine learning and image processing.

HUIMAO ZHANG received the B.S. degree in Japanese medicine from the Norman Bethune Health Science Center, Jilin University, Changchun, China, in 1994, and the M.S. and Ph.D. degrees in medical imaging and nuclear medicine from the School of Medicine, Jilin University, in 2001 and 2006, respectively.

In 2012, she visited the Molecular Imaging Research Center, Stanford University. She has been a Chief Physician with the Department of Diagnostic Radiology, The First Hospital of Jilin University, Changchun. Her research interests include body imaging, tumor molecular imaging diagnosis, and medical data analysis. From 2003 to 2004, she received the Ministry of Health Shikawa Medical Scholarship for visiting scholar qualification to visit the National Cancer Center in Japan for exchange.