Original Article

Artificial intelligence diagnostic model for multi-site fracture X-ray images of extremities based on deep convolutional neural networks

Yanling Xie[1#], Xiaoming Li[1#], Fengxi Chen[1], Ru Wen[1], Yang Jing[2], Chen Liu[1], Jian Wang[1]

[1]Department of Radiology, Southwest Hospital, Army Medical University (Third Military Medical University), Chongqing, China; [2]Huiying Medical Technology Co., Ltd., Beijing, China

Contributions: (I) Conception and design: C Liu, X Li, J Wang; (II) Administrative support: C Liu, J Wang; (III) Provision of study materials or patients: X Li, Y Xie; (IV) Collection and assembly of data: C Liu, Y Xie, X Li; (V) Data analysis and interpretation: C Liu, Y Xie, X Li, F Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Jian Wang, PhD; Chen Liu, PhD. Department of Radiology, Southwest Hospital, Army Medical University (Third Military Medical University), Chongqing 400038, China. Email: wangjian_811@foxmail.com; liuchen@aifmri.com.

Background: The rapid and accurate diagnosis of fractures is crucial for timely treatment of trauma patients. Deep learning, one of the most widely used forms of artificial intelligence (AI), is now commonly employed in medical imaging for fracture detection. This study aimed to construct a deep learning model using big data to recognize multiple-fracture X-ray images of extremity bones.

Methods: Radiographic imaging data of extremities were retrospectively collected from five hospitals between January 2017 and September 2020. The total number of people finally included was 25,635 and the total number of images included was 26,098. After labeling the lesions, the randomized method used 90% of the data as the training set to develop the fracture detection model, and the remaining 10% was used as the validation set to verify the model. The faster region convolutional neural networks (R-CNN) algorithm was adopted to construct diagnostic models for detection. The Dice coefficient was used to evaluate the image segmentation accuracy. The performances of detection models were evaluated with sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).

Results: The free-response receiver operating characteristic (FROC) curve value was 0.886 and 0.843 for the detection of single and multiple fractures, respectively. Additionally, the effective identification AUC for all parts was higher than 0.920. Notably, the AUC for wrist fractures reached 0.952. The average accuracy in detecting bone fracture regions in the extremities was 0.865. When analyzing single and multiple lesions at the patient level, the sensitivity was 0.957 for patients with multiple lesions and 0.852 for those with single lesions. In the segmentation task, the training set (the data set used by the machine learning model to train and learn) and the validation set (the data set used to evaluate the performance of the model) reached 0.996 and 0.975, respectively.

Conclusions: The faster R-CNN training algorithm exhibits excellent performance in simultaneously identifying fractures in the hands, feet, wrists, ankles, radius and ulna, and tibia and fibula on X-ray images. It demonstrates high accuracy, low false-negative rates, and controllable false-positive rates. It can serve as a valuable screening tool.

Keywords: Deep learning; artificial intelligence (AI); X-ray; multi-site fracture of extremities

Submitted Jun 23, 2023. Accepted for publication Nov 24, 2023. Published online Jan 09, 2024.
doi: 10.21037/qims-23-878
View this article at: https://dx.doi.org/10.21037/qims-23-878

Introduction

Fractures are a common problem in trauma cases (1). Studies have shown that the worldwide annual incidence of fractures ranges from 9.0 to 22.8 per 1,000 people (2). There is a growing clinical demand for imaging patients with extremity fractures, including magnetic resonance imaging (MRI), computed tomography (CT), and radiography (3,4). This demand will continue to increase in the coming years, especially for radiography (5). X-rays are the primary means of diagnosing fractures and are widely accepted by most trauma emergency patients due to their speed, convenience, low dose, and affordability. However, the diagnostic error rate still reaches 17.9% (6,7). Misdiagnosis and omission rates have increased due to the growing workloads of physicians (8,9). Limb fractures are the second most commonly missed diagnosis in medical malpractice litigations in radiology departments (10). More experienced clinical and imaging physicians are needed to accurately identify fractures, especially in primary hospitals where clinical and radiology staff may lack experience in X-ray diagnosis. These settings may result in higher rates of diagnostic errors, leading to serious patient consequences (11). Therefore, it is urgent to implement technology that can reduce physician workload and decrease misdiagnosis and missed diagnosis rates effectively.

Deep learning, which is one of the most widely used forms of artificial intelligence (AI), has found extensive application in medical imaging for fracture detection, including fractures of the hip, shoulder, wrist, and ankle (12,13). Several studies have demonstrated the potential benefits of computerized analysis based on deep learning as a diagnostic strategy, and this has recently become feasible (14). The applications and achievements of deep convolutional neural networks (DCNNs) in the medical field are expected to grow rapidly, with several studies offering significant opportunities to apply deep learning to trauma (15-17). DCNNs have demonstrated their proficiency in accurately classifying skeletal structures and pinpointing site-specific fractures with expert-level precision (18-20). Cheng *et al.* (21) demonstrated the performance of DCNNs to help junior physicians achieve urgent screening and assessment of hip fractures with 91% accuracy, 98% sensitivity, 2% false negative rate, and 0.98 area under the receiver operating characteristic curve (AUC), based on 29,210 X-ray images. Adams *et al.* (22) used DCNNs to detect the accuracy of femoral neck fractures on radiographs and compared them with subjective human perceptual judgments, showing that DCNNs can perform similarly to radiologists. Kim *et al.* (3) implemented an accurate classification model for wrist fractures based on X-ray lateral wrist films using transfer learning techniques, with an AUC of 0.954, a sensitivity of 0.90, and a specificity of 0.88. However, in the emergency department, a single examination often includes radiography from multiple sites to observe the presence of multisite fractures, and a deep learning model using only a single site is likely to miss the diagnosis (23). Single-region deep-learning models may fail to detect multiple fractures. Moreover, there is a lack of adequate reports in the literature on the X-ray AI detection of ulna, radius and tibiofibular fractures.

Hence, in this research, we utilized deep learning algorithms to develop intricate multiple skeletal deep learning models of the limbs, utilizing extensive data collected from various centers, to automatically and precisely identify fractures in the extremities. The aim was to employ these models to detect fractures across multiple regions, enhance the precision of single fracture identification, and address the challenge of accurately pinpointing multiple fractures. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-878/rc).

Methods

Study population

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the review boards of Southwest Hospital of Army Medical University (AMU), Tianjin People's Hospital, Tianjin First Central Hospital, Second Hospital of Tianjin Medical University, and Third Hospital of Nanchang, and informed consent was waived due to the retrospective study design. We retrospectively gathered

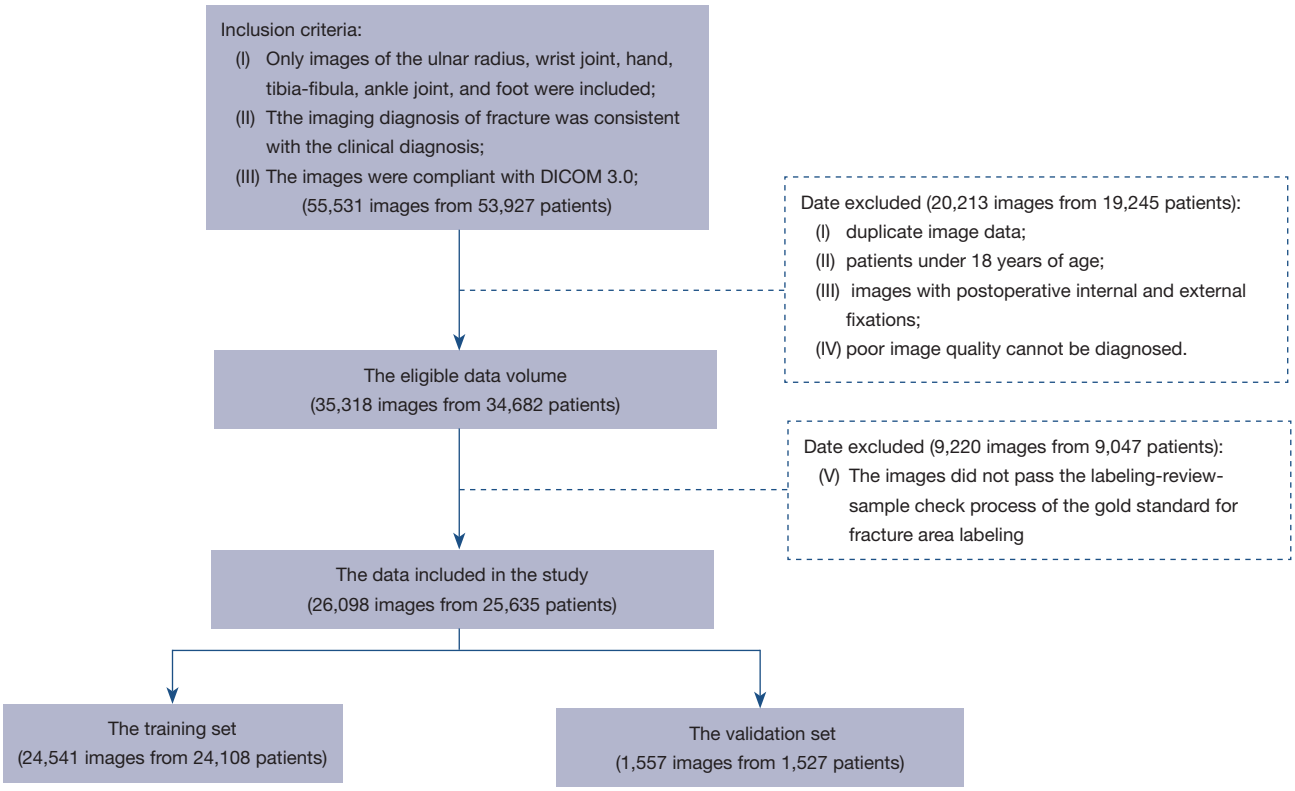

Figure 1 The original image were screened and grouped. DICOM, Digital Imaging and Communications in Medicine.

radiographic imaging data from five hospitals between January 2017 and September 2020. The data were collected from patients diagnosed with fractures of the ulnar radius, wrist, hand, tibia-fibula, ankle, and foot. The hospitals included in the study were Southwest Hospital of AMU, Tianjin People's Hospital, Tianjin First Central Hospital, Second Hospital of Tianjin Medical University, and Third Hospital of Nanchang.

The inclusion criteria for this study were as follows: (I) only images of the ulnar radius, wrist joint, hand, tibia-fibula, ankle joint, and foot were included; (II) the imaging diagnosis of fracture was consistent with the clinical diagnosis; (III) the images were compliant with Digital Imaging and Communications in Medicine (DICOM) 3.0.

The exclusion criteria were as follows: (I) duplicate image data; (II) patients under 18 years of age; (III) images with postoperative internal and external fixations; (IV) poor image quality; (V) the images did not pass the labeling-review-sample check process of the gold standard for fracture area labeling. All images were anonymized, and a total of 26,098 images were included in the follow-up study. The randomized method used 90% of the data as the training set to develop the fracture detection model, and the other 10% was used as the validation set to verify the model (see *Figure 1*).

Scanning parameters

The requirements for tube current, tube voltage, and exposure time of the limb bone are listed in *Table 1*. Any digital radiography equipment that satisfies these requirements can be utilized as acquisition equipment.

The data acquisition equipment selected for this study included Fujifilm (Japan), Simens (Germany), Kodak (USA), United Imaging (China), Cannon (Japan), Philips (Netherlands), GE (USA), Shinva (China), ECOM (China), Samsung (Korea), Orich (China), Orich (USA), Neusoft (China), Angell (China), GMM (China), Wandong (China), and Mindray (China). These devices meet the tube voltage, tube current, and exposure time requirements for photographing limb bones.

Image annotation

In reference to the gold standard fracture detection

Table 1 Scanning parameters of different parts of the extremity bones

Part	KV	mA	mAs	Ms
Ulnar radius	50	100	10	100
Wrist joint	50	100	8	80
Hand	45	100	8	80
Tibiofibular	55	100	12.5	125
Ankle joint	55	100	12.5	125
Foot	50	100	10	100

KV can be increased when the overall color of the image is white; when it is dark, KV is reduced appropriately.

method used by OsteoDetect [a similar Food and Drug Administration (FDA)-approved fracture diagnostic product], the gold standard fracture detection algorithm in this study was developed by four imaging specialists. Two annotating physicians (with 6 and 7 years of experience, respectively) manually marked all images, one physician (with 15 years of experience) reviewed the gold standard, and another physician (with 30 years of experience) spot-checked it; all the three of these duties were not being performed by the same physician. The final gold standard involved marking the location of the fracture lesion with the smallest enclosed box for images determined to be positive for fracture. The process of developing the gold standard involved two annotators independently reviewing the images, annotating the fracture lesions, and providing the annotated results to the reviewer. The reviewer then reviewed and modified the results of both annotators to form a unique final annotation result for each datum. The final annotations were randomly checked by experts to ensure quality.

Accurate annotation is an essential indicator of the absence of mislabeling or omission when annotating a single instance of a data lesion. Annotators must achieve an accuracy rate of at least 85%, while reviewers and spot checkers must achieve an accuracy rate of at least 90%. Only when the annotation accuracy meets these requirements can the data annotation work be formally performed. In this context, accuracy rate was defined as the ratio of correctly labeled images to all labeled images.

Image pre-processing

Image enhancement

We employed contrast enhancement techniques to attain image enhancement, particularly with the aim of ameliorating the gray-scale contrast within the desired gray-scale interval. This was accomplished through the suppression of aberrant pixels in the image and the extension of the gray-scale range for the pixels of significance. Additionally, image enhancement was applied to make the distribution of data more uniform, thereby increasing algorithm stability. The grayscale range stretching (13) method was used for image enhancement, as shown in *Figure 2*. The horizontal axis represents the original image pixel value, while the vertical axis represents the image pixel range mapped after enhancement. The black solid line indicates no enhancement, while the red solid line represents the result of enhancement. As seen in the image, the effective pixel range is stretched and overall contrast is greatly improved. After enhancement, grayscale values were normalized and the pixel range was adjusted to between 0 and 255.

Data amplification

The algorithm model was trained using online data augmentation. The data augmentation techniques employed included random flipping (both horizontal and vertical) and random grayscale transformation (as shown in *Figures 3*,*4*). The augmented data resulted in three times the amount of the original data. The distribution of the amplified data samples in terms of sex, age, device manufacturer, and site remained consistent with that of the original training set.

Unet-based bone segmentation algorithm

The target detection area for fracture detection was the bone region. Often, the images contain a large background area and some clutter, which can cause interference in fracture detection. Accurately localizing the target detection region is important for improving the accuracy of fracture detection. The bone segmentation algorithm located the

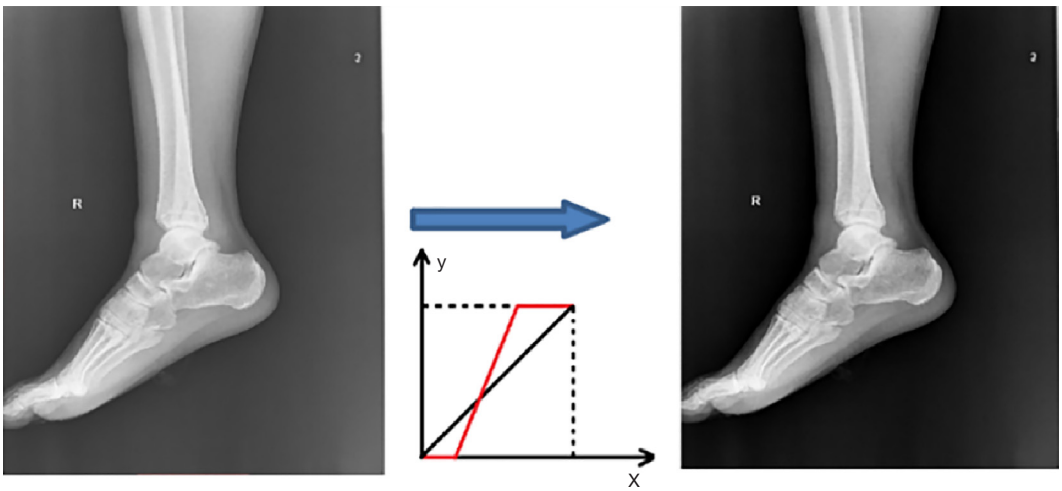

Figure 2 Example image enhancement with gray-scale range stretching. x-axis represents the pixel value of the original image, y-axis represents the mapped image pixel range after image enhancement, where the black solid line represents no image enhancement, and the red solid line represents the result after image enhancement.

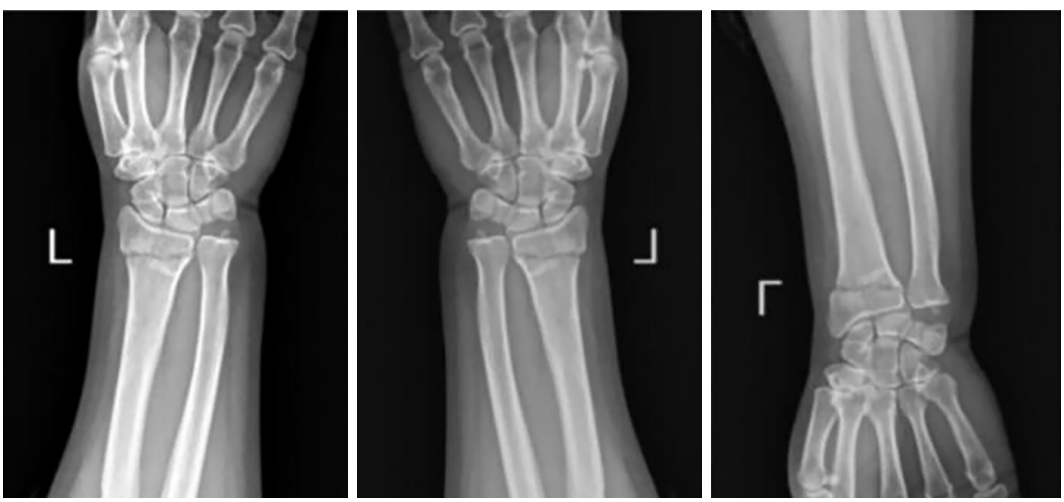

Figure 3 Example diagram of flip transformation.

bone region segmentation of the input image. After locating the bone region, it took the minimum enclosing frame and crops the image to obtain the region of interest (ROI). This region became the input of the fracture detection module.

Unet is a segmentation network proposed based on Fully Convolutional Networks (FCN) and applied to medical influence. Unet consists of an encoding network on the left half and a decoding network on the right half, with feature fusion between the encoding and decoding modules via jump connections (*Figure 5*). The encoder network iteratively consisted of two 3×3 convolutional layers and 2×2 maximum pooling layers (stride =2), with a total of four downsamplings. The number of channels was doubled with each downsampling. The decoder network iteratively consisted of one 2×2 upsampling convolutional layer and two 3×3 convolutional layers, with a total of four upsamplings. The last layer of the network used a 1×1 convolution to turn the number of channels into the desired category number.

The Unet network was 19 layers, 24.44 million parameters (crop size =256×256; batch size =32; learning rate =3E−4; epoch =60), and 31.3 GMAC (Giga Multiply-

Figure 4 Grayscale transformation example diagram (left 1: grayscale 0.8, left 2: grayscale 0.9, left 3: original image, left 4: grayscale 1.1, left 5: grayscale 1.2).

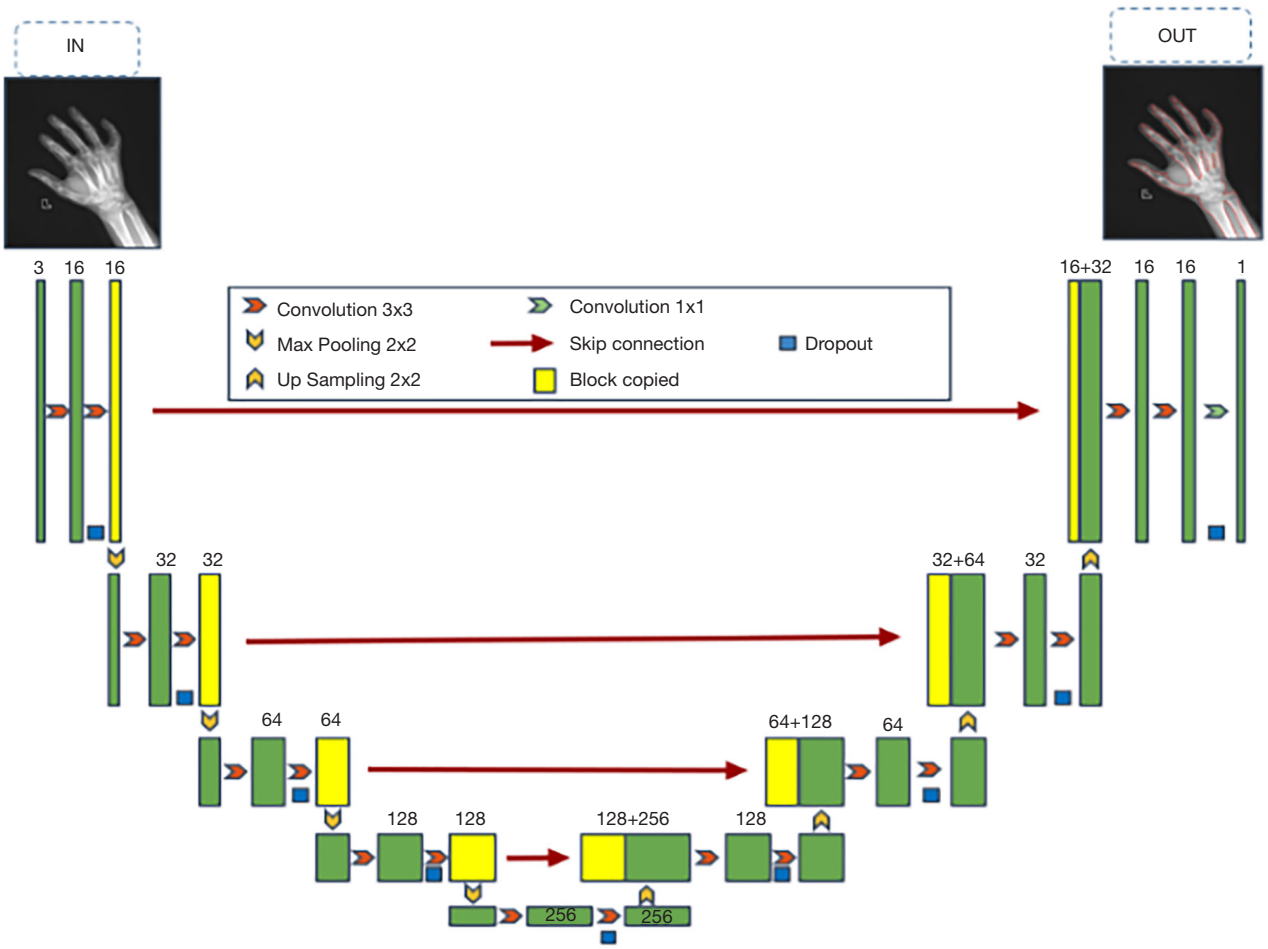

Figure 5 Model architecture of Unet partition network.

Table 2 The specific evaluation parameters

Confidence threshold	Recall rate	Accuracy rate	Average number of false positive detections per image
0.576	0.742	0.862	0.1
0.379	0.785	0.768	0.2
0.262	0.805	0.693	0.3
0.193	0.824	0.634	0.4
0.149	0.840	0.586	0.5

Add Operations per Second) computation.

Post-processing

The predictive tendency of the model can be adjusted by adjusting the thresholds for prediction in practical clinical applications. Our model considered the model sensitivity and accuracy with a threshold of 0.4 false positives per image, with a threshold of 0.193. The specific false positives were 0.4 false positive detections generated per image. The specific evaluation parameters are summarized in the following *Table 2*.

Training and validation dataset

The total number of people finally included was 25,635 and the total number of images included was 26,098. We did image ID number binding for the same patient when grouping the data to ensure that the images of the same patient were grouped into the same group. A total of 24,541 images were utilized as an independent training set to develop the fracture-detection model. The images varied in size from 2,128×2,248 pixels to 2,688×2,688 pixels and were in 8-bit grayscale color. Within the training set, 14,196 images contained fractures while 10,345 images did not. The images were categorized into six groups: hand, foot, wrist, ankle, tibiofibular and ulnar radius, with 6,641, 5,722, 4,324, 3,004, 2,548, and 2,302 images respectively. The validation set consisted of 1,557 samples, of which 842 contained fractures. The images in the validation set were categorized into six groups as well: hand, foot, wrist, ankle, tibiofibular, and ulnar radius, with 358, 299, 219, 214, 249, and 218 images respectively.

Development of faster region convolutional neural networks (R-CNN) detection algorithm

The faster R-CNN model is a classical algorithm for target recognition in computer vision. It integrates feature extraction, candidate frame extraction, rectangular frame regression, and target detection and classification into a single network. The network architecture is illustrated in *Figure 6*.

In the feature extraction stage, Resnet 50 (a pre-trained model) was utilized as the backbone network for feature extraction, while Feature Pyramid Network (FPN) (24) was employed for multi-scale feature extraction. FPN is a feature pyramid structure, as shown in *Figure 7*. This structure leverages both the high-resolution characteristics of low-level features and the high semantic information of high-level features to enhance the detection of targets of varying sizes by fusing the features of these different layers. The region proposal network (RPN) was then utilized to generate candidate target frames, determine their category through classification, and obtain the exact candidate frames through regression network correction. ROI pooling (or ROI alignment) was used to extract proposal (the output box of RPN in the two-stage approach) features from the feature map and scale them to a fixed size for proposal category discrimination. Finally, the proposed feature map was used to calculate the proposed category and further optimize the target frame boundaries.

The loss function (optimization objective function) used for faster R-CNN is the sum of classification loss and regression loss. The categorization loss is the cross-entropy loss and the regression loss is the smooth *L1* loss. The cross-entropy loss is used to determine how close the

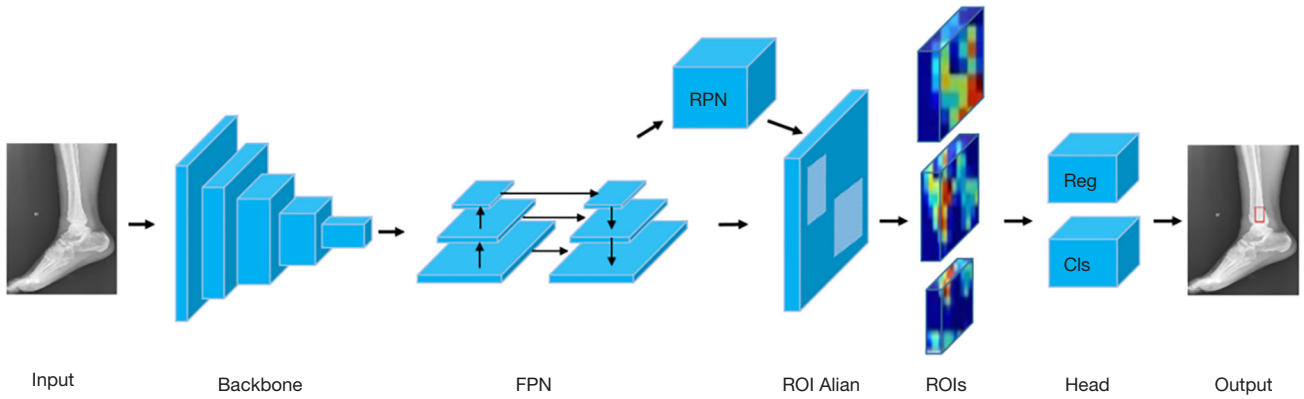

Figure 6 Network structure of fracture detection model. The red box indicates the detected fracture lesion area. FPN, feature pyramid network; RPN, region proposal network; ROI, region of interest; Reg is reg-layer (it predicts the coordinates of the proposal corresponding to the central anchor of the proposal); Cls is cls-layer (it determines whether a proposal is in the foreground or background).

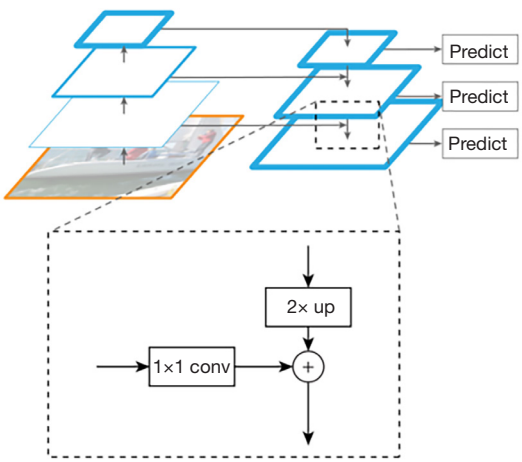

Figure 7 Schematic diagram of the feature pyramid structure. Conv, convolution.

predicted category probabilities are to the true category probabilities, and the smooth *L1* loss is used to determine how close the predicted border coordinates are to the true border coordinates. The loss function is used to measure the difference between the predicted value and the real value of the model. It is a non-negative real value function. The smaller the loss function, the better the robustness of the model. The loss function of the algorithm includes the RPN and fast R-CNN phases, which are defined as follows Eqs. [1–3]:

$$L\left(\{p_i\},\{t_i\}\right)=\frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i,p_i^*\right)+\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}\left(t_i,t_i^*\right) \quad [1]$$

$$L_{reg}\left(t_i,t_i^*\right)=R\left(t_i-t_i^*\right) \quad [2]$$

$$L_{cls}\left(p_i,p_i^*\right)=-\left(p_i^*\log p_i+\left(1-p_i^*\right)\log\left(1-p_i\right)\right) \quad [3]$$

Among them, the P_i, t_i are the predicted target probabilities and coordinate parameters, respectively p_i^*, and t_i^* are the truth target probabilities and coordinate parameters, respectively, and R is the smooth *L1* loss function.

Our model relied on the Detectron 2 framework, utilizing the Pytorch deep learning environment and Python as the programming language.

Statistical analysis and software

The basic data were analyzed using R-4.0.4 and expressed as counts (percentages), while numerical data were expressed as means (standard deviations). The Chi-squared test and *t*-test were utilized to analyze the variability of patients with and without fractures under different indicators. To evaluate the performance of the fracture region segmentation algorithm, this study utilized the Dice coefficient as the metric. The Dice coefficient takes values in the range of [0–1], with higher values indicating greater consistency between the two sets. *A* value of 0 indicates that the two sets have no intersection, while a value of 1 indicates complete consistency. In image segmentation, *A* represents the gold standard segmented image, while *B* represents the model-predicted segmented image.

$$Dice(A,B)=\frac{2|A\cap B|}{|A|+|B|} \quad [4]$$

Table 3 Data distribution of the included cases in the training and validation sets

Part	Dataset	Number of images	Gender			Results		
			Male, n (%)	Female, n (%)	P value	Positives, n (%)	Negatives, n (%)	P value
Hand	Training set	6,641	4,124 (62.1)	2,517 (37.9)	0.472	3,389 (51.03)	3,252 (48.97)	0.127
	Validation set	358	215 (60.06)	143 (39.94)		198 (55.3)	160 (44.7)	
Wrist joint	Training set	4,324	2,059 (47.62)	2,261 (52.29)	0.611	2,515 (58.17)	1,809 (41.83)	0.607
	Validation set	219	100 (45.66)	119 (54.34)		123 (56.01)	96 (43.99)	
Ulnar radius	Training set	2,302	1,290 (56.04)	1,012 (43.96)	0.576	1,440 (62.55)	862 (37.45)	0.002
	Validation set	218	127 (58.26)	91 (41.74)		113 (51.78)	105 (48.22)	
Foot	Training set	5,722	2,802 (48.97)	2,920 (51.03)	0.141	3,106 (54.28)	2,616 (45.72)	0.355
	Validation set	299	160 (53.51)	139 (46.49)		171 (57.27)	128 (42.73)	
Ankle joint	Training set	3,004	1,431 (47.64)	1,571 (52.29)	0.262	1,959 (65.2)	1,045 (34.8)	<0.001
	Validation set	214	93 (43.46)	121 (56.54)		103 (48.19)	111 (51.81)	
Tibiofibular	Training set	2,548	1,236 (48.51)	1,312 (51.49)	0.526	1,787 (70.13)	761 (29.87)	<0.001
	Validation set	249	115 (46.18)	134 (53.82)		139 (55.73)	110 (44.27)	

Columns with no gender information are not displayed in the table.

Among them, the term $|A \cap B|$ denotes the intersection of sets A and B, while $|A|$ and $|B|$ signify the cardinality of sets A and B, respectively. In this context, the coefficient of two accounts for the presence of common elements between A and B in the denominator. The formula can be comprehended as a doubling of the predicted correct outcomes divided by the sum of the real outcomes and the predicted outcomes.

The metrics used for the performance metrics of the faster R-CNN fracture detection algorithm were categorized into the lesion and patient levels. The lesion level: categorized by single lesion and multiple lesions, the assay detection performance evaluation was based on the FROC curve, which is a curve with lesion recall as the vertical coordinate and the number of false-positive lesions averaged over all images as the horizontal coordinate. The patient level: the highest predicted probability of detecting a lesion in an image was taken as the probability that the image was predicted to have a fracture. If no lesion was predicted, the probability of a fracture was predicted to be zero for that image. Thus each image had a probability that could be used to calculate receiver operating characteristic (ROC) classification performance. If the patient had multiple images, the highest probability of a fracture lesion in multiple images was used as the patient's probability of fracture. And if no lesion was predicted in any of the patient's multiple images, the patient's probability of fracture was predicted to be zero. At the lesion-based level, the free-response receiver operating characteristic (FROC) curve, recall rate, and precision rate were utilized as evaluation metrics. At the patient-based level, the ROC curve, AUC, recall rate, and specificity were used as evaluation indices. These metrics were calculated as follows: $Recall\ (Sensitivity)=TP/(TP+FN)$, $Precision=TP/(TP+FP)$, where TP represents the number of targets with positive predicted results and positive true results; and FN represents the number of targets with negative predicted results and positive true results; and FP represents the number of targets with positive predicted results and negative true results.

Results

Study population

Table 3 displays the distribution of data for the cases included in this study, categorized into training and validation sets. A total of 26,098 valid X-ray images were obtained to construct the model, comprising of 6,999 radiographs of the hand, 4,543 of the wrist, 2,520 of the radius-ulna, 6,021 of the foot, 3,218 of the ankle, and 2,797

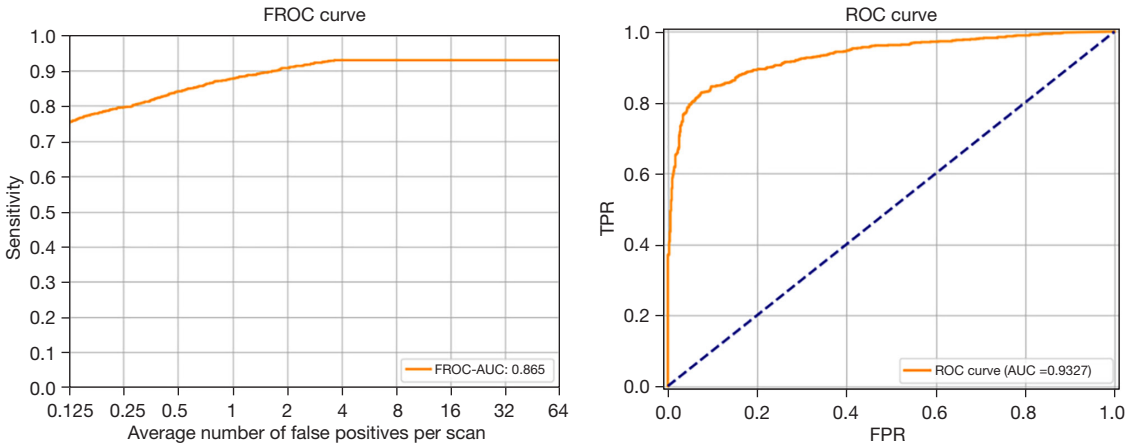

Figure 8 FROC curve and ROC curve. FROC, free-response receiver operating characteristic; AUC, area under the receiver operating characteristic curve; TPR, true positive rate; FPR, false positive rate; ROC, receiver operating characteristic.

of the tibia fibula. The data collected were well balanced in terms of gender and fracture presence, ensuring that the subsequent training of the algorithm generalized well across the training and validation sets.

The impact factors analysis results

The study results revealed that the performance of the algorithm stabilized after the training data volume reached 18,000, as indicated by a FROC-AUC of 0.865. To further investigate, separate analyses were conducted based on gender, age, collection devices, and sites using the validation set. Between-group comparisons were performed using the Chi-squared test for the area under two independent ROC curves. Notably, there were no statistically significant differences in algorithm performance observed across different genders, ages, acquisition devices, or sites ($P>0.05$). This suggests that the aforementioned factors do not exert an influence on the algorithm's performance. Additionally, in analyzing the impact of noise level on algorithm performance, it was observed that when the noise level exceeded 2.4E–02, there was a notable decline in algorithm performance. This finding emphasizes the influence of noise level, or image quality, on the algorithm's effectiveness.

Bone region segmentation results and fracture detection results

The model was trained on a dataset of radiographs of extremity bone fractures. In the segmentation task, the training set (the data set used by the machine learning model to train and learn) and the validation set (the data set used to evaluate the performance of the model) reached 0.996 and 0.975, respectively. The final validation dataset showed that the Dice values for bone region segmentation were 0.978 for the ankle, 0.977 for the hand, 0.969 for the foot, 0.959 for the wrist, 0.973 for the radius-ulna, and 0.992 for the tibia-fibula. The average accuracy in detecting bone fracture regions in the extremities was 0.865. The ulnar flexure had the highest calibration efficacy with an FROC value of 0.922, while the foot was the most difficult to detect with an FROC value of 0.849. Additionally, this study predicted the presence or absence of fractures in patients, with an average predictive AUC of 0.933. *Figure 8* displays the FROC and ROC curves of the validation set for the two tasks of fracture detection and classification of patients with and without fractures. *Table 4* presents the evaluation metrics for fracture detection and prediction in different parts of the extremity bones at the lesion and patient levels.

Single and multiple fracture detection results

In this study, the faster R-CNN algorithm was used to detect single and multiple fracture lesions simultaneously. A single fracture is one image containing one fracture, and multiple fractures are multiple fractures in the same image. FROC-AUC for the evaluation of the accuracy of lesion detection, according to an image containing one fracture or multiple fractures that can be divided into fracture single dataset and fracture multiple datasets, for different datasets were performed FROC calculation respectively. The results indicated that the FROC-AUC values were 0.886

Table 4 Fracture detection and fracture prediction evaluation indexes in different parts of extremity bones based on lesion and based on patient level

Part	Lesions			Patients		
	FROC-AUC	Sensitivity	Precision	AUC (95% CI)	Sensitivity	Precision
Ulnar radius	0.889	0.824	0.651	0.924 (0.879–0.969)	0.839	0.783
Hand	0.851	0.786	0.775	0.934 (0.911–0.958)	0.868	0.840
Wrist	0.878	0.795	0.813	0.952 (0.928–0.977)	0.894	0.846
Tibiofibular	0.922	0.852	0.810	0.939 (0.913–0.965)	0.888	0.902
Foot	0.849	0.792	0.776	0.924 (0.899–0.950)	0.910	0.790
Ankle	0.886	0.804	0.682	0.922 (0.893–0.950)	0.869	0.851

FROC, free-response receiver operating characteristic; AUC, area under the curve; CI, confidence interval.

and 0.843 for single and multiple lesions, respectively. The multiple lesion assay demonstrated a higher precision of 0.892, but a lower sensitivity of 0.796. Furthermore, when analyzing single and multiple lesions at the patient level, the sensitivity was 0.957 for patients with multiple lesions and 0.852 for those with single lesions.

Discussion

The results of the study demonstrate that deep learning can accurately identify fractures in multiple parts of the extremities. We utilized a Unet network-based skeletal region segmentation and fracture region detection task for extremity bones using deep learning algorithms, based on extremity bone X-ray image data. The faster R-CNN algorithm was also effective in identifying fracture regions in the ulnar radius, wrist, hand, tibia, fibula, ankle, and foot in the detection task. In the independent validation set, the FROC values for single and multiple fracture detection were 0.886 and 0.843, respectively. Moreover, detection based on the image level was superior, with an effective identification AUC higher than 0.920 for all sites, particularly for wrist joint fractures, which had an AUC value of 0.952. In the segmentation task, the model demonstrated excellent segmentation performance, with the Dice value in the bone segmentation area of the training set and validation set reached 0.996 and 0.975, respectively.

Deep learning demonstrated outstanding performance in image processing tasks. Numerous studies have reported the effective use of deep learning in identifying fractures in various locations, including the ribs (16), lumbar spine (25), shoulder (26), hip (27), wrist (3,8), and ankle (28), with a diagnostic accuracy of over 0.90. Kalmet *et al.* utilized DCNNs to accurately identify fractures on plain wrist films, and the AUC of the validation set was greater than 0.95, outperforming traditional computational methods such as segmentation and feature extraction (14). Additionally, it has been reported in the literature that deep learning models show superior performance in fracture detection, similar to intermediate or advanced radiologists and superior to junior radiologists (29). Lindsey *et al.* (8) evaluated emergency physicians' ability to diagnose fractures with and without the use of machines and showed an average of 47% reduction in the rate of misdiagnosis by emergency physicians when using machine-assisted detection of fractures in radiographs. In this study, the evaluation metrics of detection and segmentation reached 0.886 and 0.996, respectively, in the simultaneous multiple-fracture detection and segmentation task scenario.

Although our study achieved satisfactory prediction results, there were still limitations to our model. Firstly, the sensitivity is lower for multiple fractures, especially in the more complex anatomy of the hand, wrist and foot, which can lead to missing some lesion in an image. Given that the sensitivity for detecting multiple foci at the patient level reaches an impressive 0.957, and the likelihood of direct missed diagnosis in patients with multiple foci is exceedingly low, our model can serve as a valuable tool in aiding physicians with diagnosis. However, it is important to note that our model cannot entirely substitute the expertise and judgment of a diagnosing physician. Furthermore, it is worth noting that our trained fracture detection model utilizes data from five different centers, encompassing multiple devices and device parameters. This grants our model a certain degree of generalizability when predicting data from other centers and devices. However, we have not

examined whether factors such as racial diversity, domestic or foreign, or geographic variations have an impact on the results. Moreover, it is important to acknowledge that our model is currently only capable of detecting fractures in the extremities and not all fractures visible on radiography. Further optimization is required to expand its capabilities in this regard. In future studies, our plan is to extend the model to encompass other types of fractures. However, we anticipate several challenges in this process, including the need for more comprehensive multicenter data to detect a wider range of fracture types, as well as the requirement for more advanced fracture feature extraction and detection networks. As the volume of data and complexity of the model increase, the hardware demands for training will also escalate. In upcoming research endeavors, we aim to collaborate with additional hospitals to gather data from various fracture locations, enabling us to train, validate, and ultimately implement the model in clinical settings.

Conclusions

In conclusion, the faster R-CNN training algorithm exhibits excellent performance in simultaneously identifying fractures in the hands, feet, wrists, ankles, radius and ulna, and tibia and fibula on X-ray images. It demonstrates high accuracy, low false-negative rates, and controllable false-positive rates. It can serve as a valuable screening tool. Furthermore, our algorithms can accurately localize the fracture site and assist doctors in diagnosis. This will not only improve the diagnostic efficiency of physicians but also reduce the rate of missed fractures.

Acknowledgments

We are grateful to the staff of the Department of Radiology at Southwest Hospital of AMU for their invaluable assistance in this study. At the same time, we would like to thank Tianjin People's Hospital, Tianjin First Central Hospital, Second Hospital of Tianjin Medical University, and Third Hospital of Nanchang for data support.
Funding: This work was supported by the Key Technology Research and Application Demonstration of Deep Intelligent Diagnostic Platform for Medical Imaging (No. cstc2018jszx-cyztzxX0017), Chongqing Young and Middle-aged Medical High-end Talent Program (No.414Z395), and Chongqing Young and Middle-aged High-end Medical Talent Expert Workshop (Precision Imaging and Intelligent Diagnostic Workshop).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-878/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-878/coif). Y.J. is an employee of Huiying Medical Technology Co., Ltd. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Ethics Committee of Southwest Hospital of AMU, Tianjin People's Hospital, Tianjin First Central Hospital, The Second Hospital of Tianjin Medical University and The Third Hospital of Nanchang, and individual consent for this retrospective analysis was waived.

References

1. Perlepe V, Omoumi P, Larbi A, Putineanu D, Dubuc JE, Schubert T, Vande Berg B. Can we assess healing of surgically treated long bone fractures on radiograph? Diagn Interv Imaging 2018;99:381-6.
2. Court-Brown CM, Caesar B. Epidemiology of adult fractures: A review. Injury 2006;37:691-7.
3. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73:439-45.
4. Chantry A, Kazmi M, Barrington S, Goh V, Mulholland N, Streetly M, Lai M, Pratt G; . Guidelines for the use of

imaging in the management of patients with myeloma. Br J Haematol 2017;178:380-93.
5. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, Sköldenberg O, Gordon M. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017;88:581-6.
6. Leeper WR, Leeper TJ, Vogt KN, Charyk-Stewart T, Gray DK, Parry NG. The role of trauma team leaders in missed injuries: does specialty matter? J Trauma Acute Care Surg 2013;75:387-90.
7. Petinaux B, Bhat R, Boniface K, Aristizabal J. Accuracy of radiographic readings in the emergency department. Am J Emerg Med 2011;29:18-25.
8. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, Hanel D, Gardner M, Gupta A, Hotchkiss R, Potter H. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018;115:11591-6.
9. Kachalia A, Gandhi TK, Puopolo AL, Yoon C, Thomas EJ, Griffey R, Brennan TA, Studdert DM. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. Ann Emerg Med 2007;49:196-205.
10. Pinto A, Berritto D, Russo A, Riccitiello F, Caruso M, Belfiore MP, Papapietro VR, Carotti M, Pinto F, Giovagnoni A, Romano L, Grassi R. Traumatic fractures in adults: missed diagnosis on plain radiographs in the Emergency Department. Acta Biomed 2018;89:111-23.
11. Pinto A, Brunese L. Spectrum of diagnostic errors in radiology. World J Radiol 2010;2:377-83.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
13. Kijowski R, Liu F, Caliva F, Pedoia V. Deep Learning for Lesion Detection, Progression, and Prediction of Musculoskeletal Disease. J Magn Reson Imaging 2020;52:1607-19.
14. Kalmet PHS, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, Ibrahim A, Hulst LV, Lambin P, Poeze M. Deep learning in fracture detection: a narrative review. Acta Orthop 2020;91:215-20.
15. Zhang J, Xie Y, Wang Y, Xia Y. Inter-Slice Context Residual Learning for 3D Medical Image Segmentation. IEEE Trans Med Imaging 2021;40:661-72.
16. Zhang B, Jia C, Wu R, Lv B, Li B, Li F, Du G, Sun Z, Li X. Improving rib fracture detection accuracy and reading efficiency with deep learning-based detection software: a clinical evaluation. Br J Radiol 2021;94:20200870.
17. Zhang L, Lang Z, Wei W, Zhang Y. Embarrassingly Simple Binarization for Deep Single Imagery Super-Resolution Networks. IEEE Trans Image Process 2021;30:3934-45.
18. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017;284:574-82.
19. Grossman S, Gaziv G, Yeagle EM, Harel M, Mégevand P, Groppe DM, Khuvis S, Herrero JL, Irani M, Mehta AD, Malach R. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. Nat Commun 2019;10:4934.
20. Schönenberger C, Hejduk P, Ciritsis A, Marcon M, Rossi C, Boss A. Classification of Mammographic Breast Microcalcifications Using a Deep Convolutional Neural Network: A BI-RADS-Based Approach. Invest Radiol 2021;56:224-31.
21. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, Chung IF, Liao CH. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29:5469-77.
22. 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.
23. Langerhuizen DWG, Janssen SJ, Mallee WH, van den Bekerom MPJ, Ring D, Kerkhoffs GMMJ, Jaarsma RL, Doornberg JN. What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review. Clin Orthop Relat Res 2019;477:2482-91.
24. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017;9:936-44.
25. Burns JE, Yao J, Summers RM. Vertebral Body Compression Fractures and Bone Density: Automated Detection and Classification on CT Images. Radiology 2017;284:788-97.
26. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, Kim JY, Moon SH, Kwon J, Lee HJ, Noh YM, Kim Y. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018;89:468-73.
27. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, McConnell MV, Percha B, Snyder TM, Dudley JT. Deep learning predicts hip fracture using

confounding patient and healthcare variables. NPJ Digit Med 2019;2:31.
28. 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.
29. Wang Y, Li Y, Lin G, Zhang Q, Zhong J, Zhang Y, Ma K, Zheng Y, Lu G, Zhang Z. Lower-extremity fatigue fracture detection and grading based on deep learning models of radiographs. Eur Radiol 2023;33:555-65.

Cite this article as: Xie Y, Li X, Chen F, Wen R, Jing Y, Liu C, Wang J. Artificial intelligence diagnostic model for multi-site fracture X-ray images of extremities based on deep convolutional neural networks. Quant Imaging Med Surg 2024;14(2):1930-1943. doi: 10.21037/qims-23-878

