Original Article

Deep learning-assisted classification of calcaneofibular ligament injuries in the ankle joint

Ming Ni[1]^, Yuqing Zhao[1], Xiaoyi Wen[2], Ning Lang[1], Qizheng Wang[1], Wen Chen[1], Xiangzhu Zeng[1], Huishu Yuan[1]

[1]Department of Radiology, Peking University Third Hospital, Beijing, China; [2]Institute of Statistics and Big Data, Renmin University of China, Beijing, China

Contributions: (I) Conception and design: M Ni, Y Zhao, H Yuan; (II) Administrative support: H Yuan, N Lang; (III) Provision of study materials or patients: M Ni, W Chen, Q Wang, X Zeng; (IV) Collection and assembly of data: M Ni, Y Zhao, W Chen, Q Wang, X Zeng; (V) Data analysis and interpretation: M Ni, X Wen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Huishu Yuan, MD. Department of Radiology, Peking University Third Hospital, No.49 Huayuan North Road, Haidian District, Beijing 100191, China. Email: huishuy@bjmu.edu.cn.

Background: The classification of calcaneofibular ligament (CFL) injuries on magnetic resonance imaging (MRI) is time-consuming and subject to substantial interreader variability. This study explores the feasibility of classifying CFL injuries using deep learning methods by comparing them with the classifications of musculoskeletal (MSK) radiologists and further examines image cropping screening and calibration methods.

Methods: The imaging data of 1,074 patients who underwent ankle arthroscopy and MRI examinations in our hospital were retrospectively analyzed. According to the arthroscopic findings, patients were divided into normal (class 0, n=475); degeneration, strain, and partial tear (class 1, n=217); and complete tear (class 2, n=382) groups. All patients were divided into training, validation, and test sets at a ratio of 8:1:1. After preprocessing, the images were cropped using Mask region-based convolutional neural network (R-CNN), followed by the application of an attention algorithm for image screening and calibration and the implementation of LeNet-5 for CFL injury classification. The diagnostic effects of the axial, coronal, and combined models were compared, and the best method was selected for outgroup validation. The diagnostic results of the models in the intragroup and outgroup test sets were compared with those results of 4 MSK radiologists of different seniorities.

Results: The mean average precision (mAP) of the Mask R-CNN using the attention algorithm for the left and right image cropping of axial and coronal sequences was 0.90–0.96. The accuracy of LeNet-5 for classifying classes 0–2 was 0.92, 0.93, and 0.92, respectively, for the axial sequences and 0.89, 0.92, and 0.90, respectively, for the coronal sequences. After sequence combination, the classification accuracy for classes 0–2 was 0.95, 0.97, and 0.96, respectively. The mean accuracies of the 4 MSK radiologists in classifying the intragroup test set as classes 0–2 were 0.94, 0.91, 0.86, and 0.85, all of which were significantly different from the model. The mean accuracies of the MSK radiologists in classifying the outgroup test set as classes 0–2 were 0.92, 0.91, 0.87, and 0.85, with the 2 senior MSK radiologists demonstrating similar diagnostic performance to the model and the junior MSK radiologists demonstrating worse accuracy.

Conclusions: Deep learning can be used to classify CFL injuries at similar levels to those of MSK radiologists. Adding an attention algorithm after cropping is helpful for accurately cropping CFL images.

Keywords: Deep learning; magnetic resonance imaging (MRI); diagnosis; ankle

^ ORCID: 0000-0003-0225-4816.

Submitted May 12, 2022. Accepted for publication Sep 07, 2022. Published online Oct 13, 2022.
doi: 10.21037/qims-22-470
View this article at: https://dx.doi.org/10.21037/qims-22-470

Introduction

An ankle sprain is a common sports injury, and the inversion force of the ankle with the foot in plantar flexion is a common cause of ankle ligament injuries (1). Anterior talofibular ligament (ATFL) and calcaneofibular ligament (CFL) injuries often occur together, with approximately 80% of ankle sprains involving the ATFL and 20% involving the CFL (2). Due to anatomical variability and the thin ligament and complex anatomy of the CFL (3), CFL injuries are prone to missed diagnosis or misdiagnosis (4). Moreover, the need for surgical repair is controversial (5), with an expert consensus stating that 80% of surgeons choose to repair CFL injuries (6). The CFL participates in forming the posterior fibulotalocalcaneal ligament complex (7), and studies have shown that talus and calcaneus varus and calcaneal displacement increase after CFL injuries (5). Therefore, accurate assessment of the severity of CFL injuries is helpful in the formulation of treatment plans. Ankle arthroscopy is the gold standard for diagnosing CFL injuries, but it cannot assess CFL injuries before surgery (8). Magnetic resonance imaging (MRI) is one method for diagnosing CFL injuries preoperatively. Its accuracy has been reported in the literature to be between 0.72 and 0.96, with sensitivities ranging from 0.47 to 0.95 (9-11). Additionally, the diagnosis of CFL injuries is quite time-consuming. Although ultrasonography has also been used to diagnose CFL injuries, its diagnostic effect is inferior to that of MRI, and the examination method is highly controversial (12).

With the development of computer technology and hardware, artificial intelligence (AI) methods based on deep learning have received increasing attention in the medical field (13-16). Thus far, deep learning has been widely used in medical practices related to the musculoskeletal (MSK) system (17,18) and has an excellent performance in segmentation, classification, and angle measurement of different diseases (18-21). Yang *et al.* segmented the ulna and radius in dual-energy X-ray imaging by developing a deep learning network with residual building blocks (ResBlock), and the average dice coefficients were between 0.97 and 0.99 (22). Astuto *et al.* assessed the severity of cartilage, meniscus, and anterior cruciate ligament damage through a hierarchic 3-dimensional convolutional model after segmenting various structures on knee (MRI through V-Net, yielding an area under the receiver operating characteristic (ROC) curve (AUC) of diagnosis between 0.83 and 0.93 (23). Ashkani *et al.* used Inception V3 and Resnet-50 to detect fractures in ankle X-ray images, and the recognition accuracy was between 0.92 and 0.99 (24). However, there has been no deep learning study for grading CFL injuries, and existing deep learning research for ligaments has mainly focused on the anterior cruciate ligament of the knee joint.

A deep learning network consists of multiple nodes, each of which simulates the interconnection of neurons in the human brain. When information received by the node exceeds its threshold value, the information becomes output, and the node is weighted. The weights of different nodes are adjusted repeatedly according to the error, and then an optimal solution is obtained after several iterations (25). Deep learning can reduce the time required for diagnosis and achieve similar diagnostic results as those of radiologists while demonstrating high diagnostic consistency (26). However, due to the complexity of the anatomy of the ankle joint and the slenderness of the CFL, precise locating and cropping of the CFL are challenging, and it is almost impossible to delineate the region of interest (ROI) along the edges of the ligament because the ligament loses its normal shape after tearing. Hence, accurately locating the CFL has become a difficult problem to solve.

This study developed an MRI-based system to classify CFL injuries using deep learning in order to assist radiologists in reporting CFL injuries more quickly and accurately. The study also explored screening and calibration methods for cropping images. We present the following article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-22-470/rc).

Methods

Patients

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The retrospective study was approved by the Ethics Committee of Medical Science Research of Peking University Third Hospital (No. IRB00006761), and individual consent for this retrospective analysis was waived. The clinical and imaging data of patients

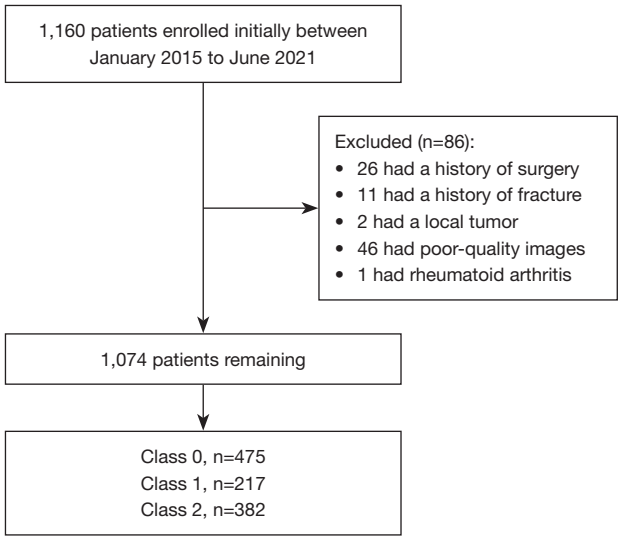

Figure 1 Flowchart showing baseline patient characteristics.

undergoing MRI and ankle arthroscopy in our hospital from January 2015 to June 2021 were retrospectively collected.

The inclusion criteria were the following: MRI examination of the ankle joint performed 3 months before ankle arthroscopy and complete surgical records. The exclusion criteria were the following: history of rheumatoid arthritis (n=1), history of previous ankle surgery (n=26), history of local fractures (n=11), presence of local tumors (n=2), and poor image quality (n=46). A total of 1,160 patients met the inclusion criteria. After 86 patients were excluded according to the exclusion criteria, 1,074 patients were finally included in this study. The details of patient inclusion and exclusion are shown in *Figure 1*.

All patients were divided into 3 groups according to their ankle arthroscopy results: a normal group (class 0, n=475); a degeneration, strain, and partial tear group (class 1, n=217); and a complete tear group (class 2, n=382). All patients were divided into a training set, validation set, and test set at a ratio of 8:1:1. The details of the grouping are shown in *Table 1*. In addition, a data set consisting of 253 patients was used as an outgroup set (46, 93, and 114 patients in classes 0, 1, and 2, respectively) for outgroup validation. The outgroup set was derived from 2 hospitals. The period of data collection and the inclusion and exclusion criteria were the same as those for the intragroup set.

Ankle arthroscopy classification criteria

All ankle arthroscopies in this study were performed at our sports medicine center and were diagnosed and recorded according to the following criteria: normal CFL arthroscopy showing intact ligament alignment, regular morphology, and good tension; CFL degeneration or strain arthroscopy showing continuous ligament fiber bundles but poor tension and ligament laxity with ligament thickening, swelling, or irregular morphology; and acute strains accompanied by surrounding soft tissue edema. Partial tears were characterized by partial discontinuity of the ligament but with a remnant partial connection and local scar formation present in some patients. A complete tear was defined as a complete disruption of the continuity of the ligament with the free ends visible.

MRI scanning

MRI was performed with a 3.0 T MR system (Signa HDxt, GE Medical Systems, Waukesha, WI, USA), and an 8-channel ankle coil was used for the intragroup. Axial and coronal fat-saturated proton density-weighted fast spin-echo (FS-PD-FSE) sequences were obtained in the study. The scan parameters for the axial FS-PD-FSE sequences were as follows: echo time (TE) 36 ms, repetition time (TR) 2,180 ms, field of view (FOV) 14 cm, slice thickness 3.0 mm, slice gap 0.3 mm, and number of excitations (NEX) 2. The scan parameters for the coronal FS-PD-FSE sequences were as follows: TE 35 ms, TR 2,680 ms, FOV 14 cm, slice thickness 3.0 mm, slice gap 0.3 mm, and NEX 2.

The images of the outgroup were obtained from a Discovery MR750 (3.0 T; GE Medical Systems), Discovery MR750WS (3.0 T; GE Medical Systems), MAGNETOM Prisma (3.0 T; Siemens Healthineers, Erlangen, Germany), Signa MRexplorer (1.5 T; GE Medical Systems), UMR770 (3.0 T; United Imaging Healthcare, Shanghai, China), Optima MR430s (1.5 T; GE Medical Systems), or Optima MR360 (1.5 T; GE Medical Systems). All examinations were conducted with axial and coronal FS-PD-FSE sequences. The scan parameters of the axial FS-PD-FSE sequences were as follows: TE 26–45 ms, TR 2,060–3,948 ms, FOV 14–15 cm, slice thickness 2.0–3.0 mm, slice gap 0.2–0.3 mm, and NEX 1–2. The scan parameters of the coronal FS-PD-FSE sequences were as follows: TE 33–72 ms, TR 1,914–3,948 ms, FOV 14–15 cm, slice thickness 2.0–3.0 mm, slice gap 0.2–0.3 mm, and NEX 1–2.

Axial and coronal images of ankle joints obtained by scanning with different devices in the intragroup and outgroup are shown in *Figure 2* and *Figure 3*.

Table 1 Demographic data

Group	Grade	Subjects (n)	Deep learning grouping (n)			Age (Y, Mean ± SD)	Sex (n)
			Training	Validation	Testing		
Intragroup	Class 0	475	380	47	48	32.16±11.79	305 M and 170 F
	Class 1	217	174	21	22	31.53±10.84	129 M and 88 F
	Class 2	382	262	60	60	30.69±10.94	250 M and 132 F
Outgroup	Class 0	46	–	–	–	31.91±11.22	26 M and 20 F
	Class 1	93	–	–	–	33.16±10.91	56 M and 37 F
	Class 2	114	–	–	–	31.66±12.01	73 M and 41 F

–, represents no relevant values.

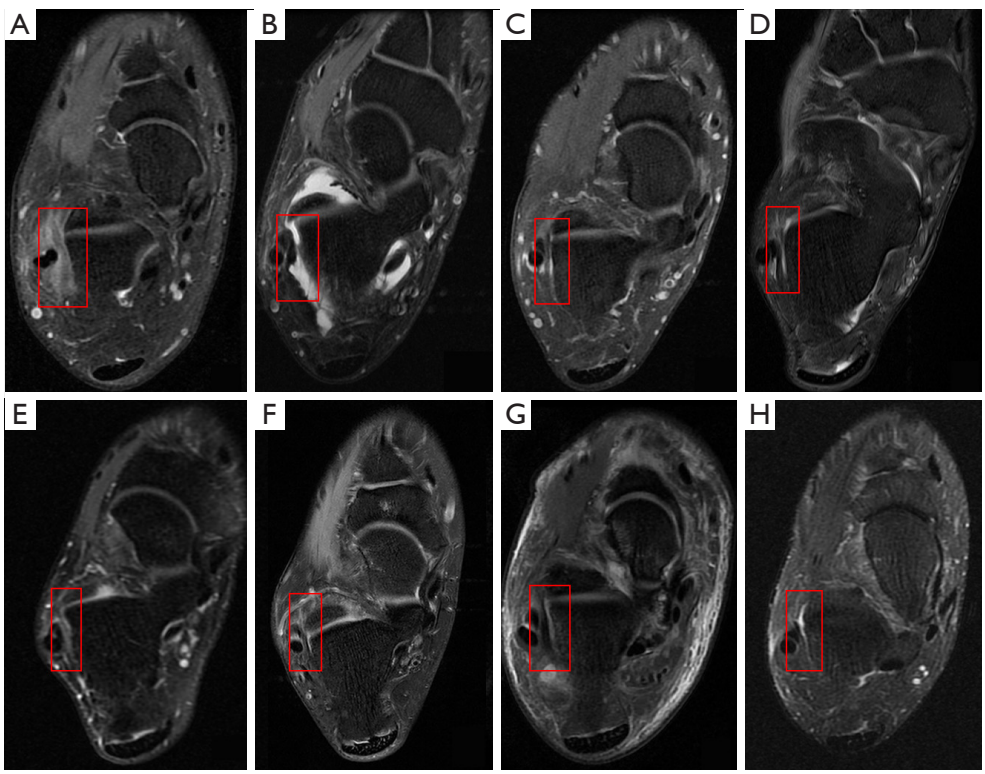

Figure 2 Axial fat-saturation proton density-weighted fast spin-echo images of the ankle joint obtained by different scanning devices in the intragroup and outgroup. The red box in the image represents the region where the calcaneofibular ligament structure is located (not the region of interest). (A) Signa HDxt. (B) Discovery MR750WS. (C) MR750. (D) MAGNETOM Prisma. (E) Signa Mrexplorer. (F) UMR770. (G) Optima MR430s. (H) Optima MR360.

MSK radiologist evaluations

Patients in both the intragroup and outgroup test sets were diagnosed by 4 MSK radiologists with different years of experience: 2 MSK radiologists with approximately 15 years of diagnostic experience (W Chen and YQ Zhao) and 2 MSK radiologists with approximately 5 years of experience (M Ni and QZ Wang). All radiologists were systematically trained at our institution to diagnose patients with axial and coronal FS-PD-FSE images and record the time taken for the diagnosis (specifically, the

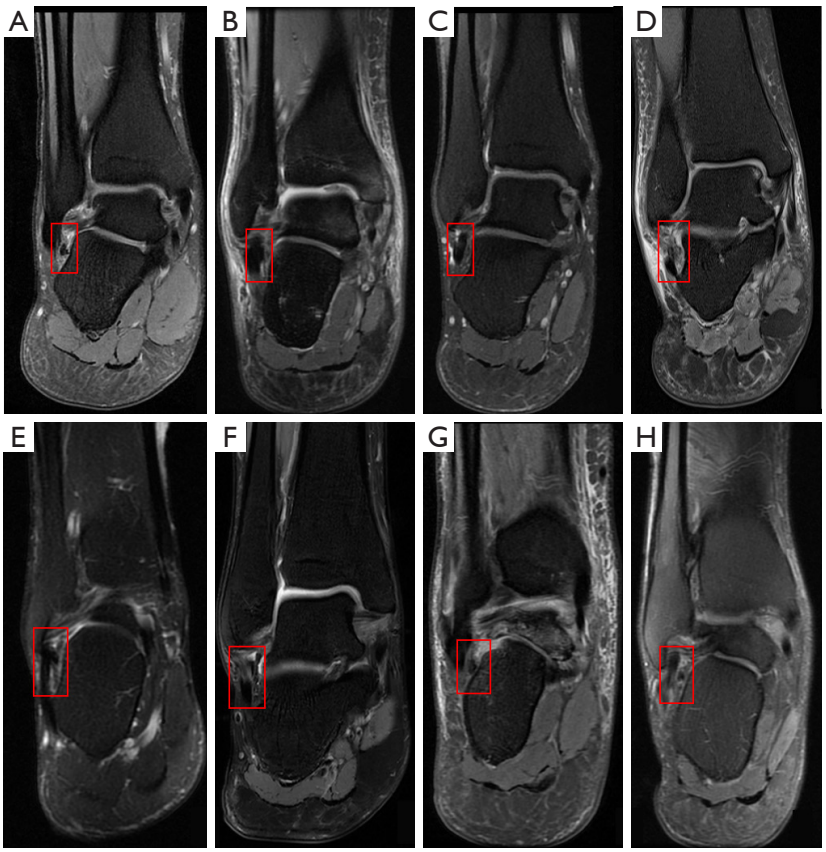

Figure 3 Coronal fat-saturation proton density-weighted fast spin-echo images of the ankle joint obtained by different scanning devices in intragroup and outgroup. The red box in the image represents the region where the calcaneofibular ligament structure is located (not the region of interest). (A) Signa HDxt. (B) Discovery MR750WS. (C) MR750. (D) MAGNETOM Prisma. (E) Signa Mrexplorer. (F) UMR770. (G) Optima MR430s. (H) Optima MR360.

time between opening the patient's image on the computer and obtaining the diagnosis). The diagnostic criteria were as follows: normal ligaments showing good tension and a uniform low signal with clear borders; degeneration and strains showing varying degrees of signal increase within the ligament, and with the ligament potentially being mildly thickened, thinned, or irregular, but continuous and possibly accompanied by surrounding soft tissue edema or effusion; partial ligament tears characterized by local disruption of ligament continuity and a significant high signal, with thickening, laxity, and irregularity of the ligament pattern, but with some ligaments still being connected; and complete ligament tears characterized by a complete disruption of ligament continuity and with free ends and a significant high signal intensity of fluid at the rupture site, potentially accompanied by surrounding soft tissue edema.

The ROI was delineated by 2 MSK radiologists (M Ni and YQ Zhao) with different levels of seniority. The CFL was first outlined by the junior MSK radiologist (M Ni), including the complete CFL structure, with effort made to minimize extraneous structures. The ROI was subsequently revised by the senior MSK radiologist (YQ Zhao) to ensure that it was appropriate. The ROIs were outlined using Python-based LabelImg software (https://github.com/tzutalin/labelImg).

Deep learning workflow

In this study, the images of all patients were first preprocessed and then cropped by a Mask region-based convolutional neural network (R-CNN) to obtain local CFL images, and an attention algorithm was added to screen and calibrate the cropped regions. The cropped images were input into

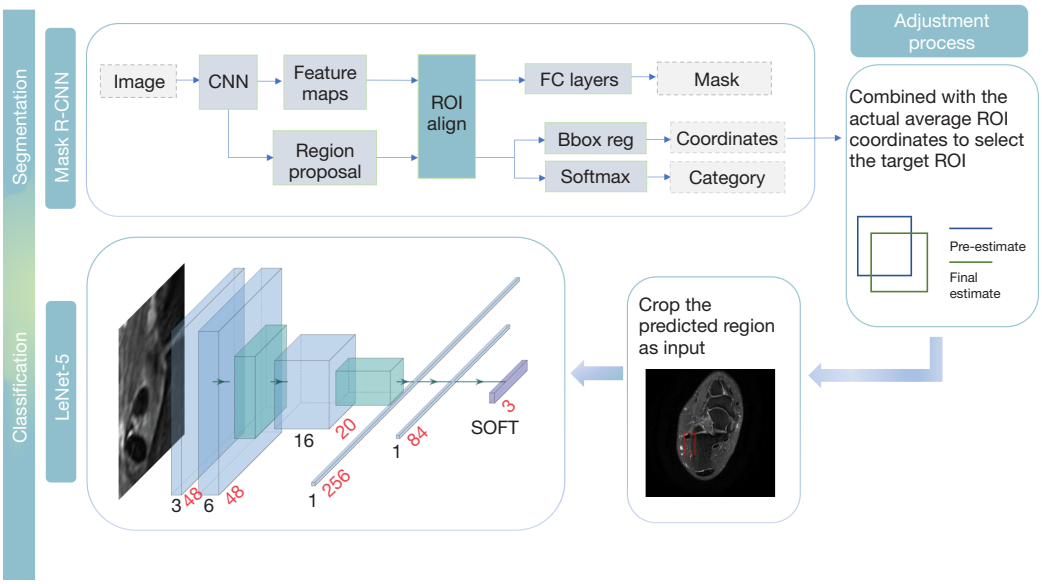

Figure 4 Flowchart of the deep learning process. After image preprocessing, Mask R-CNN was used for cropping, and an attention algorithm was applied to the cropped image for further adjustment. Finally, LeNet-5 was used for classification. R-CNN, region-based convolutional neural network; ROI, region of interest; FC layer, fully connected layer.

LeNet-5 for CFL injury classification model training, and outgroup validation was performed to evaluate model generalizability. All deep learning models were trained on a computer with an Nvidia Tesla V100 [16 GB video random access memory (VRAM)] and Intel Xeon Gold 5215 CPU. The specific process is shown in *Figure 4*.

Preprocessing

Image preprocessing plays a vital role in image analysis; its primary purpose is to eliminate irrelevant information in an image, enhance the detectability of image features of interest, and simplify an image to the greatest extent, thereby improving the reliability of feature extraction, image cropping, and recognition (27). We converted the pixel values in all images into the range of [0, 1] through pixel normalization, which can speed up the convergence of the training network, avoid the exploding gradient problem, and ensure that smaller values in the output data are not lost. Subsequently, we resized all images to 256×256 pixels.

Since there was a data imbalance between the classifications in this study, we augmented and balanced the data by rotating the images (–15°–15°, random) and randomly adding Gaussian noise to the data. After data augmentation, the amount of data between the different groups was balanced, and the total amount of data was expanded, which featured unlearning and overfitting data in the model training. Finally, all data were shuffled before training.

Image detection

The complete image contains a large amount of image information, much of which is irrelevant information that increases the complexity of the input information and can reduce the classification effect of the model. The goal of image cropping is to minimize irrelevant information in the image (28); the features extracted by the classification model are more focused on the target area, thereby improving the classification effect of the model. In this study, Mask R-CNN was used to crop the local CFL image (the left and right sides were cropped separately) to a uniform 48×48 pixels. Mask R-CNN was first proposed in 2018 (29), adding a branch for predicting detection masks based on Faster R-CNN. RoIAlign has replaced the original ROI pooling to improve the accuracy of the ROI layer, and Mask R-CNN is now one of the most eminent models for image detection at PRESENT. The model consists of two parts: one is the backbone for feature extraction, and the other is the head for ROI classification, box regression, and mask prediction. Mask R-CNN is widely used in many deep-learning studies and has achieved excellent results (30,31).

We input the axial and coronal images (including ROIs) into Mask R-CNN for model training, respectively. Due to the complex structure of the ankle joint and the slender ligaments, Mask R-CNN produced many cropped images, most of which did not contain the CFL structure and were thus considered invalid. Therefore, we added an attention algorithm to the Mask R-CNN output. Since the position of the CFL in the same sequence (such as axial FS-PD-FSE) tended to be consistent, we extracted the average coordinates of the ROI in the training set of the same sequence (left and right, separately), which were used to screen and calibrate the image cropping process. The method first extracted all manually outlined ROIs in the training set to calculate and obtain a mean ROI (rectangle). Subsequently, the overlap between the images cropped by Mask R-CNN and the average ROI was detected. When there were multiple overlapping images at the same time, the images with a predicted probability of more than 80% were retained. If there were nonoverlapping cropped images for individual patients, they were directly cropped according to the average ROI. Through this method, cropping results irrelevant to the target area could be excluded, and cropped images related to the target could be retained so that a more accurate cropped image could be input into the classification model to improve the classification effect of the model.

In general, after training Mask R-CNN with the original ROI-containing imaging, we added an attention algorithm to the output of Mask R-CNN to limit the ROI and thus solve the problem of the model not accurately cropping the CFL structure.

Classification

LeNet-5 was trained on CFL injury classification using the cropped images. It separately classified the axial and coronal FS-PD-FSE images, each of which contained multiple images. Different images may show different prediction results; thus, we assigned equal weights to each image of the same sequence. Finally, the classification based on the highest weight was used as the final output prediction result. Additionally, the results for the axial FS-PD-FSE and coronal FS-PD-FSE images were combined to obtain the combined diagnostic results, all images of both sequences were assigned the same weights, and the prediction results were obtained based on the weights. The axial FS-PD-FSE, coronal FS-PD-FSE, and combination model results were then compared, and the model with the best classification effect was selected for the outgroup set. LeNet-5 used the rectified linear activation function (RELU) as the activation function. The batch size was 32, the largest epoch was 500, the learning rate was 0.001, the optimizer was Adam, and the weight decay was 0.0005.

Statistical analysis

This study completed all deep learning model training and statistical analyses using Python (version 3.6.0; Python Software Foundation, Fredericksburg, VA, USA) software, and the TensorFlow (version 2.4.0, open source) framework was used for related data processing. The mAP was used to evaluate the cropping effect for the images. The ROC curve was used to evaluate the effect of model classification. The accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the AUC were used to evaluate the performance of LeNet-5. The intraclass correlation coefficient (ICC) was used to assess the reliability of the diagnoses between the 4 radiologists. McNemar test was used to compare the difference between the diagnostic accuracy of the deep learning model and that of the radiologists. The cropped image was visualized using the Grad-Cam heatmap. A value of $P<0.05$ indicated that the difference was statistically significant.

Results

The total diagnostic time for the test set in the intragroup was 28.6 s, with an average of 220 ms per patient, and the total diagnostic time for the outgroup was 68.31 s, with an average of 270 ms per patient. The 4 MSK radiologists took 4.0, 3.9, 4.6, and 4.4 h, respectively, to diagnose the intragroup test set, averaging 107–128 s per patient, and 7.7, 7.7, 9.4, and 9.0 h for the outgroup, respectively, with an average of 109–134 s per patient.

The mAP of Mask R-CNN with the attention algorithm for axial FS-PD-FSE image cropping was 0.94 (left) and 0.96 (right), and that for coronal FS-PD-FSE image cropping was 0.90 (left) and 0.94 (right). A comparison of the ROI outlined by the radiologist and the cropped image obtained by the Mask R-CNN with the attention mechanism is shown in *Figure 5*. The AUC of LeNet-5 in the classification of the axial FS-PD-FSE sequences into classes 0, 1, and 2 was 0.95, 0.97, and 0.96, and the accuracy was 0.92, 0.93, and 0.92, respectively. The AUC of LeNet-5 in the classification of the coronal FS-PD-FSE sequences into classes 0, 1, and 2 was 0.84, 0.97, and 0.95, and the accuracy was 0.89, 0.92, and 0.90, respectively. The AUC after model

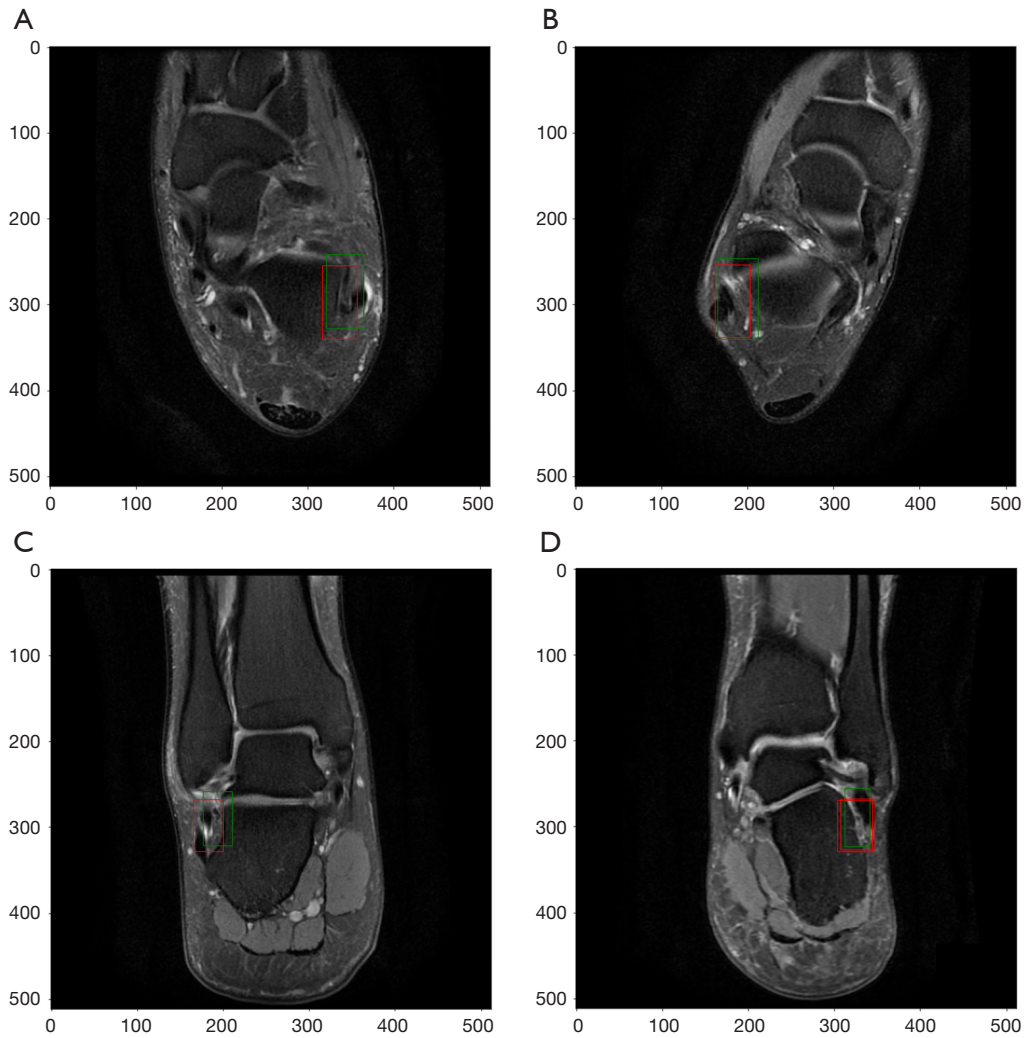

Figure 5 The region of interest outlined by the radiologists compared to the cropping results produced by Mask R-CNN with the attention algorithm. The green box represents the region of interest drawn by the radiologists, and the red box represents the region of interest obtained by Mask R-CNN with the attention algorithm. Images with more than 80% prediction probability were retained. (A,B) Axial images. (C,D) Coronal images, where image D exists with 2 model-cropped regions of interest. R-CNN, region-based convolutional neural network.

combination for classifying classes 0, 1, and 2 was 0.94, 0.97, and 0.96, and the accuracy was 0.95, 0.97, and 0.96, respectively. The detailed classification results of the model are shown in *Table 2*. The confusion matrix of LeNet-5 for coronal images, sagittal images, model combination, and the outgroup data set is shown in *Figure 6*. *Figure 7* shows the Grad-Cam heat map of LeNet-5 classifying CFL injuries under different types of images and classifications. The LeNet-5 pays better attention to CFL.

The ICC of the 4 MSK radiologists for the intragroup test set was 0.98, with mean accuracies of 0.94, 0.91, 0.86, and 0.85. The differences between the diagnostic results of the 4 MSK radiologists and the deep learning results were statistically significant, with the deep learning achieving better classification results than the MSK radiologists. The ICC of the MSK radiologists for the outgroup was 0.98, with mean accuracies of 0.92, 0.91, 0.87, and 0.85. The differences between the diagnostic effect of the deep learning model and that of the 2 senior MSK radiologists were not statistically significant (P>0.99 and P=0.70). The differences between the 2 junior MSK radiologists and the model were statistically significant, with the latter

Table 2 Diagnostic parameters of LeNet-5 for the classifications based on axial, coronal, and combination images

Sequence	Class	AUC	Accuracy*	Sensitivity*	Specificity*	NPV*	PPV*
Axial	0	0.95	0.92 (5,672/6,159)	0.85 (1,624/1,904)	0.95 (4,048/4,255)	0.94 (4,048/4,328)	0.89 (1,624/1,831)
	1	0.97	0.93 (5,755/6,159)	0.92 (1,988/2,150)	0.94 (3,767/4,009)	0.96 (3,767/3,929)	0.89 (1,988/2,230)
	2	0.96	0.92 (5,686/6,159)	0.89 (1,865/2,105)	0.94 (3,821/4,054)	0.94 (3,821/4,061)	0.89 (1,865/2,098)
Coronal	0	0.94	0.89 (7,858/8,858)	0.82 (2,221/2,717)	0.92 (5,637/6,141)	0.92 (5,637/6,133)	0.82 (2,221/2,725)
	1	0.97	0.92 (8,162/8,858)	0.91 (2,835/3,118)	0.93 (5,327/5,740)	0.95 (5,327/5,610)	0.87 (2,835/3,248)
	2	0.95	0.90 (7,932/8,858)	0.82 (2,491/3,023)	0.93 (5,441/5,835)	0.91 (5,441/5,973)	0.86 (2,491/2,885)
Combination	0	0.94	0.95 (1,196/1,255)	0.89 (397/444)	0.99 (802/811)	0.94 (802/852)	0.98 (394/403)
	1	0.97	0.97 (1,216/1,255)	0.97 (403/416)	0.97 (813/839)	0.98 (813/826)	0.94 (403/429)
	2	0.96	0.96 (1,205/1,255)	0.97 (384/395)	0.95 (821/860)	0.99 (821/832)	0.91 (384/423)

*, accuracy is presented as accuracy (TP + TN/TP + FP + FN + TN); Sensitivity is presented as sensitivity (TP/TP + FN); Specificity is presented as specificity (TN/FP + TN); NPV is presented as negative predictive value (TN/FN + TN); PPV is presented as negative predictive value (TP/FP + TP). AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; FP, false positive; FN, false negative; TP, true positive.

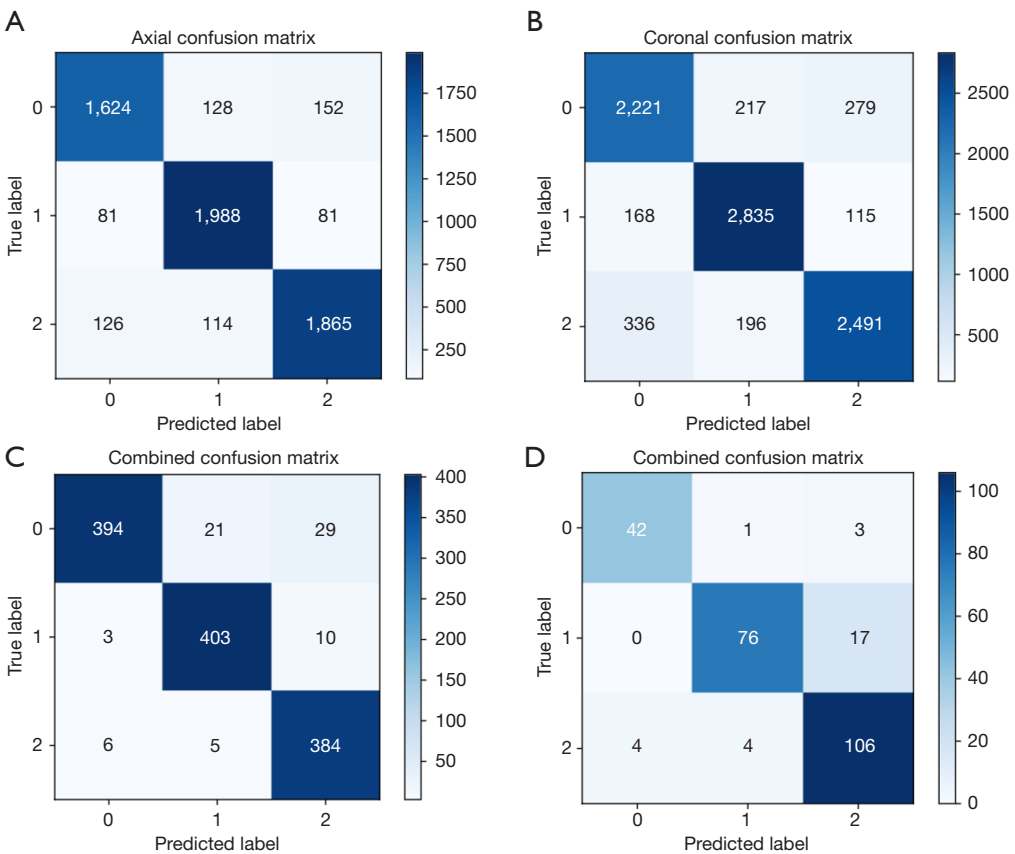

Figure 6 Confusion matrix of LeNet-5 for coronal images, sagittal images, model combination, and outgroup set. (A-C) Intragroup set. (D) Outgroup set. (A) Confusion matrix of axial images. (B) Confusion matrix of coronal images. (C) Confusion matrix of axial + coronal (combination) images. (D) Confusion matrix for outgroup set classification.

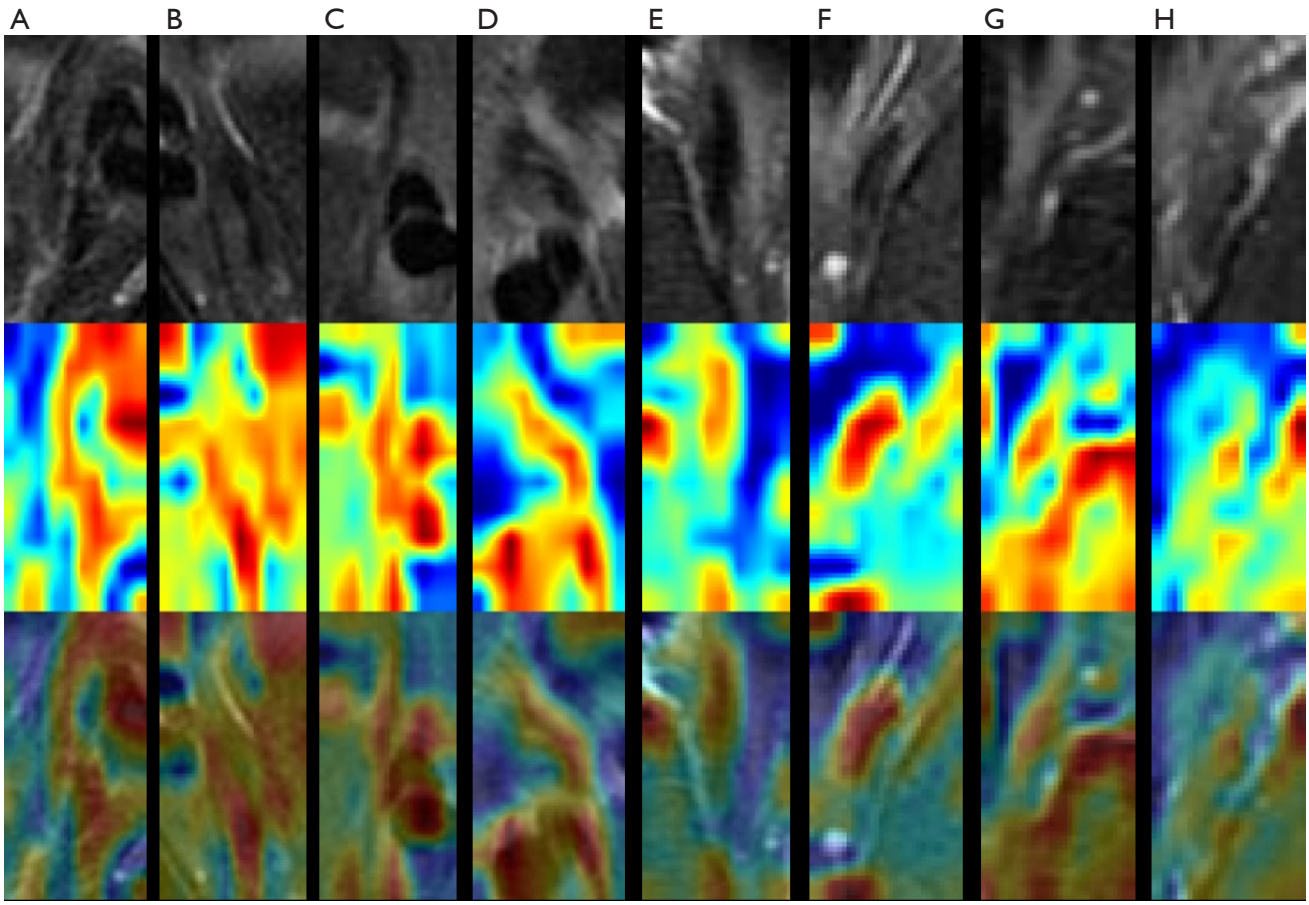

Figure 7 Grad-Cam heatmap of LeNet-5 classifying axial and coronal images. All images are partial images cropped by Mask R-CNN. The colors in the figure, from blue to red, indicate that the model pays less to more attention to the region. The first row of all images is the cropped image output by Mask R-CNN combined with the attention mechanism, the second row represents the model heatmap, and the third row represents the overlapping image of the cropped image and the heatmap. (A-D) The cropped images of the axial images and the arthroscopic results of grade 0, grade 1, and grade 2, respectively. (E-H) The cropped images of coronal images, with arthroscopic results of grade 0, grade 1, and grade 2, respectively. R-CNN, region-based convolutional neural network.

outperforming the former in the outgroup test set. A comparison of the diagnostic performances of the MSK radiologists and the deep learning model is shown in *Table 3*.

Discussion

This study aimed to design and evaluate a deep learning model for CFL injury classification and compare the diagnostic performance with that of MSK radiologists with different levels of experience. The results show that deep learning can be used to classify CFL injuries and achieve similar but faster diagnostic results than can radiologists. Combining the images of multiple sequences can improve the classification effect of the model, and adding an attention algorithm after the cropping model is helpful for accurately cropping the CFL. Additionally, the results for the outgroup test set show that the model trained in this study can obtain excellent diagnostic results with images from different pieces of MR equipment and thus has good generalizability.

In this study, we trained deep learning models using axial and coronal FS-PD-FSE sequences separately and in combination. The results showed that although excellent diagnostic results could be obtained using axial or coronal PD-FSE-FS sequences alone, the accuracy, sensitivity, and specificity of the deep learning models improved after inputting a combination of the 2 sequences; even minor improvements can produce benefits for more patients.

Table 3 Comparison of the diagnostic results of the deep learning model and the 4 musculoskeletal radiologists in the intragroup and outgroup test sets

Group	Reader	Class	Accuracy	Sensitivity	Specificity	χ^2	P value
Intragroup	LeNet-5 (combination)	0	0.95	0.89	0.99	–	–
		1	0.97	0.97	0.97		
		2	0.96	0.97	0.95		
	Radiologist 1	0	0.93	0.9	0.95	41.38	<0.001
		1	0.92	0.86	0.93		
		2	0.95	0.92	0.99		
	Radiologist 2	0	0.92	0.85	0.96	32.88	<0.001
		1	0.88	0.82	0.90		
		2	0.92	0.88	0.94		
	Radiologist 3	0	0.90	0.83	0.94	20.95	<0.001
		1	0.82	0.68	0.85		
		2	0.86	0.80	0.91		
	Radiologist 4	0	0.85	0.77	0.90	17.78	<0.001
		1	0.84	0.68	0.87		
		2	0.85	0.80	0.89		
Outgroup	LeNet-5 (combination)	0	0.97	0.91	0.98	–	–
		1	0.91	0.82	0.97		
		2	0.89	0.93	0.86		
	Radiologist 1	0	0.95	0.83	0.98	0.00	1.0
		1	0.90	0.88	0.91		
		2	0.91	0.90	0.92		
	Radiologist 2	0	0.92	0.80	0.95	0.15	0.70
		1	0.90	0.89	0.90		
		2	0.92	0.88	0.96		
	Radiologist 3	0	0.89	0.74	0.92	5.51	0.02
		1	0.84	0.81	0.86		
		2	0.87	0.82	0.91		
	Radiologist 4	0	0.90	0.70	0.95	7.95	0.005
		1	0.83	0.82	0.83		
		2	0.83	0.78	0.87		

–, represents no relevant values.

Although a CFL tear is not an absolute indication for surgery, a completely torn CFL is often more likely to cause chronic ankle instability (32). Despite the controversy regarding the function of the CFL, studies have shown that CFL tears are associated with subtalar instability and ankle instability (33,34). Therefore, the classification of CFL injuries is valuable in clinical decision-making, and a rapid and consistent diagnosis is important for improving

the efficiency of radiologists and reducing errors. In future studies, this aspect of the model will be used as part of the overall assessment of lateral ankle instability to form a comprehensive diagnostic system.

When we trained Mask R-CNN for image cropping, we found that by using a small amount of medical image data, a slender ligament structure, and a complex image structure, the cropping effect of the model was very poor, and multiple images were not cropped to the region where the CFL was located. Because the CFL loses its normal shape after tearing, the local features of the cropped image were very different. Existing image segmentation technologies require that the part identified for cropping has a specific boundary, so the direct use of the image segmentation algorithm was not effective. To address this, we added an attention algorithm after Mask R-CNN to screen all cropped images by restricting the cropping area to the average ROI and correcting the position by averaging it with the average ROI coordinates to improve the accuracy of the final cropping area. At present, deep learning models still have certain difficulties recognizing small structures within complex structures and remain limited by small amounts of medical data; there is no perfect solution to address these problems, but the method proposed in this study may serve as a helpful approach.

In this study, the performance of the deep learning model was compared with that of radiologists. The differences between the deep learning model and the 4 MSK radiologists of different seniorities for the intragroup were statistically significant; specifically, the diagnostic effect of the deep learning model was better than that of the radiologists, and the speed and consistency of diagnosis were better. The deep learning model performed similarly to the 2 senior MSK radiologists in diagnosing the outgroup but outperformed the 2 junior MSK radiologists. The deep learning model trained in this study had an excellent diagnostic effect when classifying images obtained from different pieces of MR equipment, showing good generalizability. Even the same MSK radiologist showed some intragroup and outgroup differences when diagnosing patients. For radiologists with different levels of experience and from different medical institutions, deep learning can help reduce the impact of a diagnostic gap by producing a diagnosis comparable to that of expert radiologists.

This study had some shortcomings. First, it included only patients who underwent arthroscopic surgery, resulting in a certain bias in the data. Second, the attention algorithm added in this study may need to be restored to the mean ROI when used in different institutions, resulting in some reduction in the generalizability of our method; however, this study also obtained good results with images from multiple pieces of MR equipment (outgroup test set), and the model still has good generalizability in normative scanning conditions. Third, the arthroscopic procedures in this study were performed by multiple surgeons, and there may be some differences in diagnostic results among them. Finally, other structures, such as the ATFL, articular cartilage, and bony structures, were not included in this study; hence we will conduct more studies in the future to incorporate them in the construction of additional models.

Conclusions

This study developed a deep learning model that can classify CFL injuries, compared its performance with the performance of four MSK radiologists using ankle arthroscopy findings as the gold standard and validated it with outgroup data. Better diagnostic results were obtained when multiple sequences were combined, and the model may have potential as an assistant tool. In addition, our proposed attention algorithm can improve the cropping effect and screen cropped images when cropping small structures with complex structures.

Acknowledgments

Funding: This study received funding from the National Natural Science Foundation of China (No. 81871326).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-22-470/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-470/coif). The authors report that this study received funding from the National Natural Science Foundation of China (No. 81871326). The authors have no other 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. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The retrospective study was approved by the Ethics Committee of Medical Science Research of Peking University Third Hospital (No. IRB00006761), and individual consent for this retrospective analysis was waived.

References

1. Fong DT, Ha SC, Mok KM, Chan CW, Chan KM. Kinematics analysis of ankle inversion ligamentous sprain injuries in sports: five cases from televised tennis competitions. Am J Sports Med 2012;40:2627-32.
2. Broström L. Sprained ankles. V. Treatment and prognosis in recent ligament ruptures. Acta Chir Scand 1966;132:537-50.
3. Pereira BS, van Dijk CN, Andrade R, Casaroli-Marano RP, Espregueira-Mendes J, Oliva XM. The calcaneofibular ligament has distinct anatomic morphological variants: an anatomical cadaveric study. Knee Surg Sports Traumatol Arthrosc 2020;28:40-7.
4. Cao S, Wang C, Ma X, Wang X, Huang J, Zhang C. Imaging diagnosis for chronic lateral ankle ligament injury: a systemic review with meta-analysis. J Orthop Surg Res 2018;13:122.
5. Hunt KJ, Pereira H, Kelley J, Anderson N, Fuld R, Baldini T, Kumparatana P, D'Hooghe P. The Role of Calcaneofibular Ligament Injury in Ankle Instability: Implications for Surgical Management. Am J Sports Med 2019;47:431-7.
6. Michels F, Pereira H, Calder J, Matricali G, Glazebrook M, Guillo S, et al. Searching for consensus in the approach to patients with chronic lateral ankle instability: ask the expert. Knee Surg Sports Traumatol Arthrosc 2018;26:2095-102.
7. De Leeuw PAJ, Vega J, Karlsson J, Dalmau-Pastor M. The posterior fibulotalocalcaneal ligament complex: a forgotten ligament. Knee Surg Sports Traumatol Arthrosc 2021;29:1627-34.
8. Shakked R, Sheskier S. Acute and Chronic Lateral Ankle Instability Diagnosis, Management, and New Concepts. Bull Hosp Jt Dis (2013) 2017;75:71-80.
9. Park HJ, Lee SY, Park NH, Kim E, Chung EC, Kook SH, Lee JW. Usefulness of the oblique coronal plane in ankle MRI of the calcaneofibular ligament. Clin Radiol 2015;70:416-23.
10. Park HJ, Cha SD, Kim SS, Rho MH, Kwag HJ, Park NH, Lee SY. Accuracy of MRI findings in chronic lateral ankle ligament injury: comparison with surgical findings. Clin Radiol 2012;67:313-8.
11. Kumar V, Triantafyllopoulos I, Panagopoulos A, Fitzgerald S, Van Niekerk L. Deficiencies of MRI in the diagnosis of chronic symptomatic lateral ankle ligament injuries. Foot Ankle Surg 2007;13:171-6.
12. Hattori S, Nimura A, Koyama M, Tsutsumi M, Amaha K, Ohuchi H, Akita K. Dorsiflexion is more feasible than plantar flexion in ultrasound evaluation of the calcaneofibular ligament: a combination study of ultrasound and cadaver. Knee Surg Sports Traumatol Arthrosc 2020;28:262-9.
13. Schork NJ. Artificial Intelligence and Personalized Medicine. Cancer Treat Res 2019;178:265-83.
14. Currie G. Intelligent Imaging: Anatomy of Machine Learning and Deep Learning. J Nucl Med Technol 2019;47:273-81.
15. Olveres J, González G, Torres F, Moreno-Tagle JC, Carbajal-Degante E, Valencia-Rodríguez A, Méndez-Sánchez N, Escalante-Ramírez B. What is new in computer vision and artificial intelligence in medical image analysis applications. Quant Imaging Med Surg 2021;11:3830-53.
16. Ng D, Du H, Yao MM, Kosik RO, Chan WP, Feng M. Today's radiologists meet tomorrow's AI: the promises, pitfalls, and unbridled potential. Quant Imaging Med Surg 2021;11:2775-9.
17. Leung K, Zhang B, Tan J, Shen Y, Geras KJ, Babb JS, Cho K, Chang G, Deniz CM. Prediction of Total Knee Replacement and Diagnosis of Osteoarthritis by Using Deep Learning on Knee Radiographs: Data from the Osteoarthritis Initiative. Radiology 2020;296:584-93.
18. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skeletal Radiol 2020;49:183-97.
19. 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.
20. Pei Y, Yang W, Wei S, Cai R, Li J, Guo S, Li Q, Wang J, Li X. Automated measurement of hip-knee-ankle angle on the unilateral lower limb X-rays using deep learning. Phys Eng Sci Med 2021;44:53-62.
21. Hesamian MH, Jia W, He X, Kennedy P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. J Digit Imaging 2019;32:582-96.
22. Yang F, Weng X, Miao Y, Wu Y, Xie H, Lei P. Deep learning approach for automatic segmentation of ulna and radius in dual-energy X-ray imaging. Insights Imaging 2021;12:191.
23. Astuto B, Flament I, K Namiri N, Shah R, Bharadwaj U, M Link T, D Bucknor M, Pedoia V, Majumdar S. Automatic Deep Learning-assisted Detection and Grading of Abnormalities in Knee MRI Studies. Radiol Artif Intell 2021;3:e200165.
24. Ashkani-Esfahani S, Mojahed Yazdi R, Bhimani R, Kerkhoffs GM, Maas M, DiGiovanni CW, Lubberts B, Guss D. Detection of ankle fractures using deep learning algorithms. Foot Ankle Surg 2022. [Epub ahead of print]. pii: S1268-7731(22)00102-3. doi: 10.1016/j.fas.2022.05.005.
25. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, Chepelev L, Cairns R, Mitchell JR, Cicero MD, Poudrette MG, Jaremko JL, Reinhold C, Gallix B, Gray B, Geis R; Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. Can Assoc Radiol J 2018;69:120-35.
26. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, Kim N. Deep Learning in Medical Imaging: General Overview. Korean J Radiol 2017;18:570-84.
27. Masoudi S, Harmon SA, Mehralivand S, Walker SM, Raviprakash H, Bagci U, Choyke PL, Turkbey B. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. J Med Imaging (Bellingham) 2021;8:010901.
28. Flannery SW, Kiapour AM, Edgar DJ, Murray MM, Fleming BC. Automated magnetic resonance image segmentation of the anterior cruciate ligament. J Orthop Res 2021;39:831-40.
29. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 2020;42:386-97.
30. Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, Villain N, Bloch I, Cotten A, Boussel L. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. Diagn Interv Imaging 2019;100:235-42.
31. Zhang R, Cheng C, Zhao X, Li X. Multiscale Mask R-CNN-Based Lung Tumor Detection Using PET Imaging. Mol Imaging 2019;18:1536012119863531.
32. Vuurberg G, Hoorntje A, Wink LM, van der Doelen BFW, van den Bekerom MP, Dekker R, van Dijk CN, Krips R, Loogman MCM, Ridderikhof ML, Smithuis FF, Stufkens SAS, Verhagen EALM, de Bie RA, Kerkhoffs GMMJ. Diagnosis, treatment and prevention of ankle sprains: update of an evidence-based clinical guideline. Br J Sports Med 2018;52:956.
33. Choisne J, Hoch MC, Bawab S, Alexander I, Ringleb SI. The effects of a semi-rigid ankle brace on a simulated isolated subtalar joint instability. J Orthop Res 2013;31:1869-75.
34. Ringleb SI, Udupa JK, Siegler S, Imhauser CW, Hirsch BE, Liu J, Odhner D, Okereke E, Roach N. The effect of ankle ligament damage and surgical reconstructions on the mechanics of the ankle and subtalar joints revealed by three-dimensional stress MRI. J Orthop Res 2005;23:743-9.

Cite this article as: Ni M, Zhao Y, Wen X, Lang N, Wang Q, Chen W, Zeng X, Yuan H. Deep learning-assisted classification of calcaneofibular ligament injuries in the ankle joint. Quant Imaging Med Surg 2023;13(1):80-93. doi: 10.21037/qims-22-470