Ahead of print

Med Ultrason 2025:0, 1-8 Online first
DOI: 10.11152/mu-4495

Deep learning-based automated detection and diagnosis of gouty arthritis in ultrasound images of the first metatarsophalangeal joint

Lishan Xiao[1,*], Yizhe Zhao[2,3,*], Yuchen Li[1], Mengmeng Yan[1], Manhua Liu[2,3], Chunping Ning[1]

** the authors share the first authorship*

[1]Department of Ultrasound, the Affiliated Hospital of Qingdao University, Qingdao, [2]The School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, [3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China

Abstract

Aim: This study aimed to develop a deep learning (DL) model for automatic detection and diagnosis of gouty arthritis (GA) in the first metatarsophalangeal joint (MTPJ) using ultrasound (US) images. **Materials and methods:** A retrospective study included individuals who underwent first MTPJ ultrasonography between February and July 2023. A five-fold cross-validation method (training set = 4:1) was employed. A deep residual convolutional neural network (CNN) was trained, and Gradient-weighted Class Activation Mapping (Grad-CAM) was used for visualization. Different ResNet18 models with varying residual blocks (2, 3, 4, 6) were compared to select the optimal model for image classification. Diagnostic decisions were based on a threshold proportion of abnormal images, determined from the training set. **Results:** A total of 2401 US images from 260 patients (149 gout, 111 control) were analyzed. The model with 3 residual blocks performed best, achieving an AUC of 0.904 (95% CI: 0.887~0.927). Visualization results aligned with radiologist opinions in 2000 images. The diagnostic model attained an accuracy of 91.1% (95% CI: 90.4%~91.8%) on the testing set, with a diagnostic threshold of 0.328. **Conclusion:** The DL model demonstrated excellent performance in automatically detecting and diagnosing GA in the first MTPJ.

Keywords: convolutional neural network; deep learning; gout; ultrasound; first metatarsophalangeal joint

Introduction

Gout is characterized by the deposition of monosodium urate crystals in joints, leading to acute and chronic inflammation. Early and accurate diagnosis of gout is crucial for effective management and prevention of disease progression [1]. Currently, the gold standard for diagnosing gout involves joint aspiration and identification of urate crystals under a microscope [2]. However, this invasive procedure may not always be readily available or feasible in clinical practice [3].

In recent years, advancements in medical imaging technologies have opened up a new avenue for non-invasive diagnosis of gout. Ultrasound (US), in particular, has emerged as a promising modality due to its accessibility, cost-effectiveness, and real-time imaging capabilities [4,5]. Two-dimensional US has been widely used to assess joint inflammation and vascularity in various rheumatic diseases [6]. The presence of urate crystals in the joints gives rise to specific US features in gout, including the double contour sign (DCs), tophus, and aggregates [2,7,8]. These distinctive features not only assist in diagnosing gout but also distinguish it from other arthritic conditions [6]. Despite the potential of US in the diagnosis of gout, its interpretation remains subjective and operator-dependent. Variability in image acquisition, personal error and the requirement for technical expertise may impede the widespread adoption and accuracy of US diagnosis.

Deep learning (DL) technology, a branch of artificial intelligence, has achieved remarkable success in various medical image analysis tasks, including disease detection and classification [9]. A typical DL model, convolutional neural networks (CNNs) were particularly well-suited for disease detection and classification tasks, demonstrat-

Received 15.11.2024 Accepted 08.02.2025
Med Ultrason
2025:0 Online first, 1-8
Corresponding author: Chunping Ning
Department of Ultrasound,
the Affiliated Hospital of Qingdao University,
Qingdao, China
E-mail: xls715@outlook.com

ing superior performance in image recognition and feature extraction [10-13]. So, we assumed that, by training a CNNs model on a large, annotated US images dataset, it would be possible to develop a robust and automated system for gout diagnosis and interpretation.

This study aims to explore the feasibility and effectiveness of automatic gout diagnosis using DL on B-mode US images.

Materials and methods

Patient data sets

This retrospective study was approved by the Institutional Review Board of the Affiliated Hospital of Qingdao University, and informed consent from patients was waived. Patient recruitment processes were shown in figure 1.

Patients with tophus or any two or more of the DCs, bone erosion and aggregates in the US images were included in the "Gout group", otherwise they were enrolled in the "Control group". To minimize potential sources of bias in this study, we employed a random sampling method to allocate patients to the training set and testing set, ensuring that each patient had an equal probability of selection across the different data sets. This approach reduces selection bias and enhances the representativeness of both the training and testing groups with respect to the overall population, thereby increasing the external validity of the results. Furthermore, by considering the patients' baseline characteristics during the allocation process, we ensured balance among the groups with respect to key variables, thereby improving the reliability of the study.

Ultrasound protocol to capture the images

In this study, the image scanning protocol complied with the 2017 EULAR standardized procedures for US imaging in rheumatology [14] and US imaging acquisition procedures for evaluating the first metatarsophalangeal joint (MTPJ) published by Molyneux et al in 2021 [15]. US examinations were conducted using a high-resolution US system (ARIETTA 70 US diagnostic instrument) equipped with a linear array transducer (frequency range: 9~14 MHz). Patients were comfortably positioned in a sitting posture, with fully exposure of the first MTPJ. To ensure acoustic coupling, a generous amount of gel was applied to the skin, and the transducer was disinfected with antiseptic wipes before each examination.

To assess the joint, the transducer was gently positioned longitudinally, parallel to the long axis of the first MTPJ. The probe was slowly moved from the dorsal side to the inner side and finally to the sole of the toe, maintaining constant contact with the skin. The depth, focus and gain settings were adjusted to optimize image quality ensuring that the regions of interest (ROI) consistently remain within the two-thirds of the near field. The probe was tilted and rotated as required to acquire the satisfied US view. A satisfied US view should illustrate the joint clearly, including the joint fibrous capsule, synovial membrane, tendons, ligaments, bone, and surrounding soft tissue. At least 10 static images were captured from each patient to ensure sufficient quality and adequately demonstrated anatomical and pathological information. All the US images were stored in DICOM format for subsequent analysis.

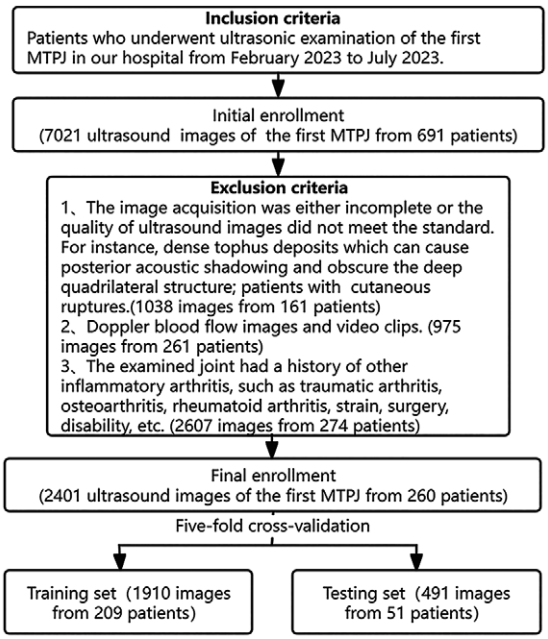

Fig 1. Flowchart for the study population. Three steps were conducted for the study: inclusion and exclusion of images, gout diagnosis by sonographers, and five-fold cross-validation. MTPJ, metatarsophalangeal joint; DCs, Double contour sign

All US examinations were performed by an experienced musculoskeletal US (MSUS) physician. To minimize the bias, the radiologist was blinded to the clinical information and final diagnosis.

Data annotation

This study annotated a total of 2401 US images of the first MTPJ from 260 patients. Two radiologists used MicroDicom Viewer (32-bit) software to read and record the imaging features of each patient, creating an index table as the foundational data for training and evaluating the DL model. To ensure accuracy, any uncertainties in the annotations were discussed and agreed upon with a MSUS physician with over 10 years of experience. The specific signs for gout may be observed in US images of the dorsal, medial, and plantar aspects of the toe: DCs, tophus, aggregates and bone erosion (fig 2).

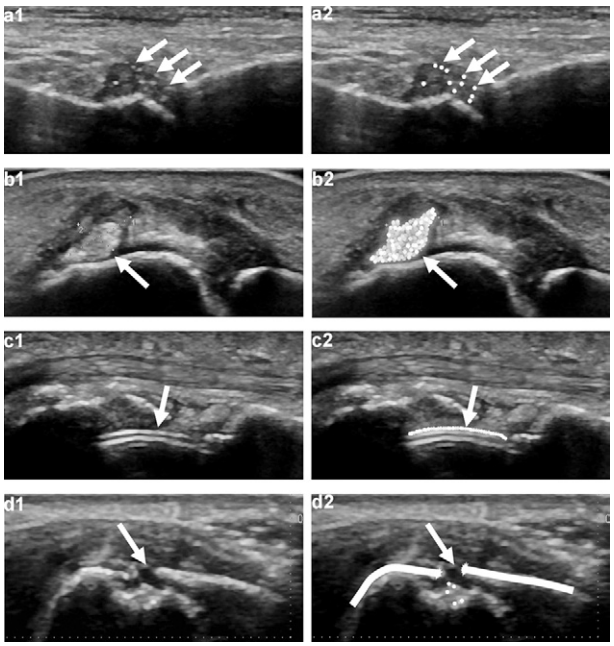

Fig 2. Longitudinal plane of B-mode ultrasound images features in the first metatarsophalangeal joint: original ultrasound images (a1, b1, c1, d1) and pattern diagrams (a2, b2, c2, d2): a) aggregates (arrows), b) tophus (arrows), c) double contour sign (arrows), d) bone erosion (arrows).

Ultrasound data preprocessing

To facilitate the image classification based on the DL model, we performed US data preprocessing as follows. First, we employed thresholding and morphological segmentation techniques to extract the significant regions from the images [16]. Then, to mitigate potential classification errors arising from variations in pixel value distributions within the images, we applied a normalization method to standardize the pixel value distribution [17]. Lastly, a normalized rectangular frame measuring 700×320 pixels was cropped, effectively encompassing all pertinent regions while minimizing extraneous areas.

Deep Learning models for gout lesion detection and gouty arthritis diagnosis

We have developed DL-based classification and diagnostic models, specifically for detecting gout lesions and diagnosing GA. Motivated by the ResNet18 model [17] in the success of image classification in computer vision, we utilized Residual blocks to construct a deep Residual CNN model for the classification of B-mode US images. Our approach in this study involved leveraging three residual blocks to establish the DL network architecture. After the feature extraction process by these residual blocks, a fully-connected layer was utilized to amalgamate all extracted features and transform them into a two-dimensional vector for gout classification. The algorithms were developed by using the training set via cross-validation and were subsequently evaluated by using the testing set. Patients from "Gout group" and "Control group" were randomly divided into 5 subgroups of approximately equal size. Each subgroup served as a validation set, while the images of the remaining 4/5 of patients were used as a training set. This process was iterated five times to obtain average results. We applied translations to the images in the training set to enhance the model's ability to better adapt to the dataset with increased data.

Based on the classification model, we further developed a diagnostic model. For each patient fold, the training set and testing set are based on the classification results obtained from the above classifier. Input the training set data for training and obtain the proportion of abnormal images when the accuracy is highest (i.e., threshold). Set the threshold from 0 to 1 (interval of 0.02). Input the testing set data, calculate the proportion of abnormal images for each case, and compare it with the threshold. If it exceeds the threshold, diagnose it as GA; otherwise, diagnose it as non-gout. Furthermore, for interpretation of the discriminative regions, we used the Grad-CAMs function with the trained DL model and the extracted feature maps were inputted to obtain the Grad-CAMs map [18]. Schematic overview of the DL model development process is visualized in figure 3. Details are presented in the Supplement materials.

Statistical analysis

Continuous variables were expressed as mean ± standard deviation, while categorical variables were presented as percentages. Model performance was assessed through the area under the receiver operating characteristic curve (AUC), and the 95% confidence intervals (CIs) of sensitivity, specificity, accuracy. The Delong test [19] and McNemar's test were employed to assess statistical significance in the predictive power of different models. The calculation formulas for statistical metrics are provided in the additional file. Statistical analyses and graphical presentation were conducted by Python v3.8 and IBM SPSS statistics 25. Two-sided $p<0.05$ was considered indicative of statistically significant difference.

Results

Participant characteristics

A total of 2401 images from 260 eligible patients were included in the analysis. The "Gout group" consisted of 1389 images from 149 patients, while the "Control group" comprised 1021 images from 111 patients. Following five-fold cross-validation, the training set includes data from 209 patients, and the testing set includes data from 51 patients.

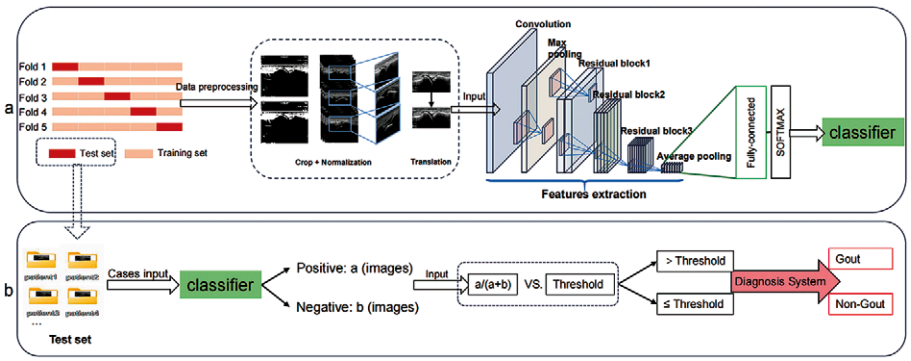

Fig 3. Schematic overview of the DL-model development process for classification and diagnosis of ultrasound images: a) Classifier: The training and testing data are randomly divided into five folds. After preprocessing the data (including cropping, normalization, and translation), they are input into a CNN architecture to extract image features. The convolutional output of the two-dimensional feature maps is transformed into a one-dimensional vector through fully connected layers. Finally, a Softmax classifier is used to classify the integrated features: b) Diagnostic System: The training and testing sets are input to the classifier on a per-case basis. The classifier categorizes a group of objects for each case, resulting in a positive classification of 'a' images and negative classification of 'b' images. It calculates and outputs the proportion of abnormal images. This ratio is compared with a threshold value. If the ratio exceeds the threshold, the diagnosis is gout; otherwise, it is non-gout. The threshold is the proportion of abnormal images at the highest accuracy in the dataset. DL, Deep Learning; CNN, Convolutional Neural Network

Deep residual CNN ablation study and diagnostic model evaluation

In this experiment, we first conducted an ablation study on the deep residual CNN architecture to assess the impact of the number of residual blocks on image classification performance. Subsequently, we developed a diagnostic model based on the optimal classifier's classification results. Specifically, we evaluated configurations with 2, 3, 4, and 6 residual blocks to build DL models. The comparison of classification performance based on different residual block configurations is shown in Table I. Furthermore, we used a confusion matrix to describe the distribution of predictions made by the DL model (fig 4). The results indicate that all four models performed well, with the DL model containing 3 residual blocks showing the best performance: in the testing set, both the AUC value and ACC were higher than the other three models (statistically significant difference compared to the 6-block model, no statistically significant difference compared to the 2 and 4-block models), as detailed in Table I. In terms of model complexity, the 3-block model consumed the least computational resources and had a parameter count slightly higher than the 2-block model, while the 2-block model, despite having the lowest parameter count, had the highest computational load (Table II). In the diagnostic model, we took the proportion of abnormal images with the highest accuracy in the training set as the diagnostic threshold, with thresholds for the five folds of patients being 0.28, 0.30, 0.36, 0.34, and 0.34, with an average of 0.33 (Supplement materials figure S1). The average diagnostic accuracy for the five-fold patients was 91.10% (46/51, 95% CI: 90.4%, 91.8%), sensitivity was 93.91% (95% CI: 89.2%, 98.6%), and specificity was 86.84% (95% CI: 79.9%, 93.8%) (Table III).

Model visualization and interpretation

A visualization technique known as Grad-CAMs, which provides a visual representation of the model's focus on specific regions within the input B-mode images were used to enhance our understanding of the diagnosis process and facilitate clinical interpretation. Visual results demonstrated that the predictions for 2000

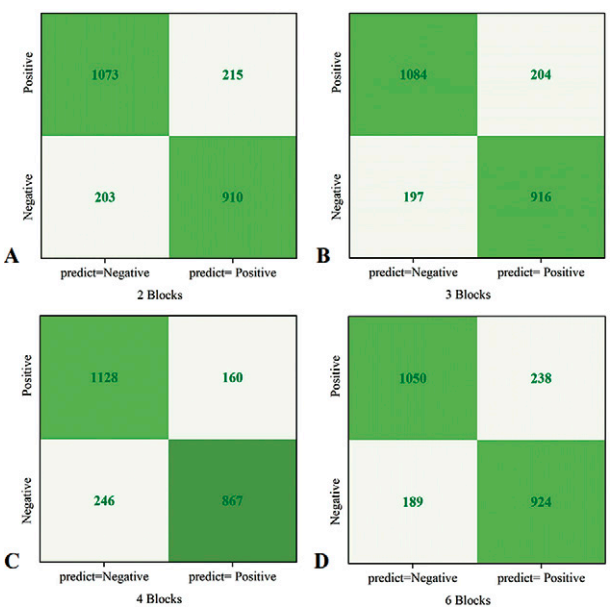

Fig 4. Predicted distributions based on deep learning methods in different residual block models: A) 2 Blocks; B) 3 Blocks; C) 4 Blocks; D) 6 Blocks.

Table I. The comparison between different numbers of Residual blocks in the testing set.

	ACC/%	p value	SEN/%	p value	SPE/%	p value	AUC	p value
3 blocks	83.3(334/401) [81.2, 85.4]	/	82.5(184/223) [78.7, 86.1]	/	84.1(217/258) [78.3, 90.2]	/	0.904 [0.887, 0.927]	/
6 blocks	82.3(330/401) [79.8, 84.7]	.32	83.0(185/223) [79.1, 86.8]	.65	81.4(210/258) [75.0, 88.0]	.08	0.899 [0.887, 0.919]	.14
4 blocks	83.0(333/401) [81.0, 85.2]	.85	77.6(173/223) [70.0, 85.5]	.01*	87.6(226/258) [79.9, 95.6]	.01*	0.901 [0.890, 0.926]	.42
2 blocks	82.6(331/401) [80.5, 84.8]	.52	81.6(182/223) [77.5, 85.8]	.74	83.3(215/258) [77.6, 88.9]	.56	0.901 [0.887, 0.920]	.22

p value: Comparison between the 3blocks and other numbers of blocks. With numbers of images in parentheses and 95% CIs in brackets. ACC, Accuracy; SEN, Sensitivity; SPE, Specificity; AUC, area under the receiver operating characteristic curve.

Table II. Comparison of Computational and Parameter Metrics for 2/3/4/6 Blocks

Model	FLOPs/M	Params/M
2 blocks	7558.27	4.61
3 blocks	2963.73	4.91
4 blocks	4012.33	4.91
6 blocks	6093.39	5.28

FLOPs: Computational workload; Params: Parameter count.

images were in agreement with the opinions of radiologists, while 401 images showed discrepancies. Through the analysis of images with inconsistent predictions, we identified that the lesions more prone to being overlooked by the model were scattered aggregates (35.03%, 69/197), subtle synovial thickening (29.95%, 59/197), and mild DCs (38.58%, 76/197). However, in false-positive images, the primary reason for the model's incorrect predictions often lies in irregularities in bone structure (11.27%, 23/204), anisotropy within ligaments and tendons (23.04%, 47/204), as well as the "cartilage interface sign" [20] formed between cartilage and effusion (4.90%, 10/204). Additionally, we found that the model tends to incorrectly focus on the following areas: in the dorsal region of the toe, the model tends to misdirect its attention to the deeper aspects of bones, where US waves cannot reach (66.67%, 20/30); in the medial region of the toe, the scanning process often requires a substantial amount of coupling agent for clearer and more comprehensive imaging, making the model prone to erroneously focus on the "pseudo-shadow areas" in the upper left/right corners of the image filled with the coupling agent (33.33%, 10/30). The US images and heat maps are shown in Supplementary figure S2.

Discussion

In this study, we conducted research on a DL model based on a deep residual CNN network for the classification and diagnosis of B-mode images of the first MTPJ in gout. To our knowledge, this is the first attempt to apply DL techniques to the diagnosis of gout using US images. We used five-fold cross-validation for internal validation. Our results demonstrate that the DL-based approach shows satisfactory predictive performance in discriminating normal and abnormal images of the first MTPJ and diagnosing GA.

The American College of Rheumatology (ACR) and the EULAR jointly released the first international consensus on ultrasonographic manifestations of GA in 2015 [4], highlighting the crucial role of MSUS in gout diagnosis. However, the complexity of joint anatomical structures poses challenges to the diagnosis of gout, potentially leading to the underestimation of subtle lesions (such as aggregates and minor synovial thickening) and difficulties in identifying subtle disease progression (such as changes in tophus volume). Additionally, the lack of radiological data in early and pre-attack stages hinders progress in early gout prevention studies. As de-

Table III. Five-fold case diagnosis performance in the testing set and threshold

	ACC/%	SEN/%	SPE/%	Threshold
Fold 1	92.31	95.65	87.50	0.28
Fold 2	89.47	95.65	80.00	0.30
Fold 3	94.74	91.30	100.00	0.36
Fold 4	84.21	95.65	66.67	0.34
Fold 5	94.74	91.30	100.00	0.36
average	91.10	93.91	86.84	0.33

ACC, Accuracy; SEN, Sensitivity; SPE, Specificity.

scribed by Yu et al [21], MSUS has many recognized advantages (such as convenience, cost-effectiveness, high spatial resolution, suitability for long-term follow-up of chronic diseases), but it also has limitations in clinical trials. Subjective biases in interpreting results introduce challenges in ensuring diagnostic accuracy, reliability, and result consistency. Furthermore, when using MSUS in clinical research, investigators must establish strict guidelines regarding accuracy, reliability, and consistency. Despite these efforts, scholars remain concerned about the reliability of MSUS and population variability caused by human factors. To address these challenges, advanced technology is needed to support MSUS in gout diagnosis.

In recent years, remarkable strides have been made in the realm of medical imaging, primarily owing to the integration of DL techniques [9,22,23]. By automatically extracting complex features from large amounts of image data, DL has shown promising results in medical imaging tasks [9]. Our study indicates that automated CNN methods have potential effectiveness in identifying gout lesions of the first MTPJ and diagnosing gout. Of note, the application of DL in MSUS was increasingly attracting attention in the medical community, demonstrating potential significance and value [24-26]. Soffer et al [26] pointed out that the application of DL techniques in MSUS are mainly focused on tasks such as bone age assessment, spinal level detection, spinal deformity pathology detection, osteoarthritis detection, and fracture detection. In the past three years, more studies have applied DL techniques to tasks like neural localization and segmentation, myositis classification, synovitis detection, developmental dysplasia of the hip classification, and even joint pathology identification [21,27-31]. However, it has not yet been applied to gout diagnosis of the first MTPJ.

In this study, we applied DL techniques for the first time to the diagnosis of GA of the first MTPJ. We achieved good performance in the test queue. Images of the dorsal, medial, and plantar aspects of the first MTPJ were normatively acquired for each participant. These images were merged and used for training, followed by internal validation and testing using five-fold cross-validation. The DL model was trained to classify images of the first MTPJ and diagnose GA. In the classification task, three residual blocks were identified as the optimal number in this study, which not only aligned with the three down sampling operations in this study but also effectively reduced the risk of overfitting simple models while consuming the least computational resources. In the diagnostic task, the study used the proportion of abnormal images with the highest diagnostic accuracy in the training set as the threshold to diagnose cases based on the classification model, achieving excellent diagnostic performance. It is noteworthy that the standardized image acquisition in this study is an important factor for the threshold diagnosis method. This method is simple and feasible, and when the technology matures, it can also promote the standardization of image acquisition in grassroots hospitals. We chose ResNet18 as the backbone network of our model because it is widely adopted and has proven effective in various medical image analysis tasks [32-34]. We also applied data augmentation techniques such as translation to enhance the diversity and robustness of our training data.

To visually demonstrate the model's reliance on feature extraction for different categories during the prediction process, we introduced the Grad-CAM to visualize the model's attention regions [35]. The integration of Grad-CAMs into DL-based diagnostic framework empowers medical professionals with the ability to discern the focal regions that influence the model's identification of gout-related features. By analyzing the heat maps that were inconsistent with the radiologist's predictions, we found that: first, lesions often overlooked in FN results are typically early and mild lesions [5]; second, several common scenarios in FP images can be easily ruled out by experienced radiologists; finally, for images with errors in identifying ROI, we hypothesize that adjustments may be attempted on non-interest areas, such as cropping, to refine the focal regions and eliminate artifact regions, aiming to reduce misjudgments caused by inaccurate model recognition of ROI, thereby improving the diagnostic accuracy of the model. However, the feasibility of these methods in practical applications still requires further investigation.

Our study has some limitations that need to be addressed in future work. Firstly, the lack of publicly available gout US image databases is a significant challenge, and our relatively small single-center dataset without external validation may limit the model's generalizability. Secondly, we only focused on the first MTPJ, which is the most commonly affected site in gout. Thirdly, we only used B-mode US images to develop the DL model. Integrating multimodal information may further enhance model performance.

In **conclusion**, we developed an innovative DL approach for automatic identification of gout lesions and diagnosis of GA on B-mode US images. Our study results indicate the potential applicability of DL technology in detecting gout lesions and diagnosing GA, potentially alleviating the burden on radiologists. To strengthen the effectiveness and reliability of our approach, future research requires larger and more diverse open-source

datasets, studies on other joints or areas, and comparisons with other methods [25].

Acknowledgments: This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFC2503302 and No. 2022YFC2503305), the Clinical Medicine and X Research Program of Affiliated Hospital of Qingdao University (QDFY+X2024133) and Shanghai Municipal Science and Technology Major Project (Grant 2021SHZ DZX0102).

Conflicts of interest: none

References

1. Mikuls TR. Gout. N Engl J Med 2022;387:1877-1887.
2. Richette P, Doherty M, Pascual E, et al. 2018 updated European League Against Rheumatism evidence-based recommendations for the diagnosis of gout. Ann Rheum Dis 2020;79:31-38.
3. El-Mallah R, Ibrahim RA, El Attar EA. The Role of Ultrasound in Evaluating the Effect of Urate-lowering Drugs in Gout Patients. Curr Rheumatol Rev 2022;18:338-345.
4. Neogi T, Jansen TL, Dalbeth N, et al. 2015 Gout classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Ann Rheum Dis 2015;74:1789-1798.
5. Li S, Xu G, Liang J, Wan L, Cao H, Lin J. The Role of Advanced Imaging in Gout Management. Front Immunol 2021;12:811323.
6. Taljanovic MS, Melville DM, Gimber LH, et al. High-resolution US of rheumatologic diseases. Radiographics 2015;35:2026-2048.
7. Dalbeth N, Gosling AL, Gaffo A, Abhishek A. Gout. Lancet 2021;397:1843-1855.
8. Christiansen SN, Ostergaard M, Slot O, et al. Assessing the sensitivity to change of the OMERACT ultrasound structural gout lesions during urate-lowering therapy. RMD Open 2020;6:e001144.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-444.
10. Zhang Y, Liu YL, Nie K, et al. Deep Learning-based Automatic Diagnosis of Breast Cancer on MRI Using Mask R-CNN for Detection Followed by ResNet50 for Classification. Acad Radiol 2023;30 Suppl 2:S161-S171.
11. Yu TF, He W, Gan CG, et al. Deep learning applied to two-dimensional color Doppler flow imaging ultrasound images significantly improves diagnostic performance in the classification of breast masses: a multicenter study. Chin Med J (Engl) 2021;134:415-424.
12. Al-Antari MA, Al-Masni MA, Kim TS. Deep Learning Computer-Aided Diagnosis for Breast Lesion in Digital Mammogram. Adv Exp Med Biol 2020;1213:59-72.
13. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500-510.
14. Moller I, Janta I, Backhaus M, et al. The 2017 EULAR standardised procedures for ultrasound imaging in rheumatology. Ann Rheum Dis 2017;76:1974-1979.
15. Molyneux P, Bowen C, Ellis R, Rome K, Jackson A, Carroll M. Ultrasound Imaging Acquisition Procedures for Evaluating the First Metatarsophalangeal Joint: A Scoping Review. Ultrasound Med Biol 2022;48:397-405.
16. Wiharto, Palgunadi Y. Blood Vessels Segmentation in Retinal Fundus Image using Hybrid Method of Frangi Filter, Otsu Thresholding and Morphology. IJACSA 2019;10:417-423.
17. Consalvo S, Hinterwimmer F, Neumann J, et al. Two-Phase Deep Learning Algorithm for Detection and Differentiation of Ewing Sarcoma and Acute Osteomyelitis in Paediatric Radiographs. Anticancer Res 2022;42:4371-4380.
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 2020;128:336-359.
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845.
20. Gutierrez M, Schmidt WA, Thiele RG, et al. International Consensus for ultrasound lesions in gout: results of Delphi process and web-reliability exercise. Rheumatology (Oxford) 2015;54:1797-1805.
21. Cheng Y, Jin Z, Zhou X, et al. Diagnosis of Metacarpophalangeal Synovitis with Musculoskeletal Ultrasound Images. Ultrasound Med Biol 2022;48:488-496.
22. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019;69:127-157.
23. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
24. Cronin NJ, Finni T, Seynnes O. Using deep learning to generate synthetic B-mode musculoskeletal ultrasound images. Comput Methods Programs Biomed 2020;196:105583.
25. Shin Y, Yang J, Lee YH, Kim S. Artificial intelligence in musculoskeletal ultrasound imaging. Ultrasonography 2021;40:30-44.
26. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide. Radiology 2019;290:590-606.
27. Lo CM, Lai KL. Deep learning-based assessment of knee septic arthritis using transformer features in sonographic modalities. Comput Methods Programs Biomed 2023;237:107575.
28. Kinugasa M, Inui A, Satsuma S, et al. Diagnosis of Developmental Dysplasia of the Hip by Ultrasound Imaging Using Deep Learning. J Pediatr Orthop 2023;43:e538-e544.
29. Atalar H, Ureten K, Tokdemir G, Tolunay T, Ciceklidag M, Atik OS. The Diagnosis of Developmental Dysplasia of the

Hip From Hip Ultrasonography Images With Deep Learning Methods. J Pediatr Orthop 2023;43:e132-e137.
30. Zhou Z, Zhao C, Qiao H, et al. RATING: Medical knowledge-guided rheumatoid arthritis assessment from multimodal ultrasound images via deep learning. Patterns (N Y) 2022;3:100592.
31. Oelen D, Kaiser P, Baumann T, et al. Accuracy of Trained Physicians is Inferior to Deep Learning-Based Algorithm for Determining Angles in Ultrasound of the Newborn Hip. Ultraschall Med 2022;43:e49-e55.
32. Liu Y, She GR, Chen SX. Magnetic resonance image diagnosis of femoral head necrosis based on ResNet18 network. Comput Methods Programs Biomed 2021;208:106254.
33. Chen Z, Jiang Y, Zhang X, et al. ResNet18DNN: prediction approach of drug-induced liver injury by deep neural network with ResNet18. Brief Bioinform 2022;23:bbab503.
34. Naz J, Sharif MI, Sharif MI, Kadry S, Rauf HT, Ragab AE. A Comparative Analysis of Optimization Algorithms for Gastrointestinal Abnormalities Recognition and Classification Based on Ensemble XcepNet23 and ResNet18 Features. Biomedicines 2023;11:1723.
35. Zhang Y, Hong D, McClement D, Oladosu O, Pridham G, Slaney G. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. J Neurosci Methods 2021;353:109098.