ImmunoTargets and Therapy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Deep Learning Classification of Rheumatoid Arthritis in Hand Radiographs Interpretability Insights and Web Application

Kanglin Cai[1,2,*], Dengfeng Dou[3,*], Guibing Deng[4,*], Yunzhen Zhan[5], Huilian Huang[1], Zhitao Feng[2,6]

[1]The Second People's Hospital Affiliated to Three Gorges University / Yichang Second People's Hospital, Yichang, Hubei, 443000, People's Republic of China; [2]Third-Grade Pharmacological Laboratory on Chinese Medicine Approved by State Administration of Traditional Chinese Medicine, College of Medicine and Health Science, China Three Gorges University, Yichang, Hubei, 443002, People's Republic of China; [3]Independent Cardiovascular Research Lab, Chinese Institutes for Medical Research, Capital Medical University, Beijing, 100069, People's Republic of China; [4]The First College of Clinical Medical Sciences, China Three Gorges University, Yichang, Hubei, 443003, People's Republic of China; [5]College of Engineering, China Pharmaceutical University, Nanjing, Jiangsu, 211198, People's Republic of China; [6]Institute of Rheumatology, the First College of Clinical Medical Sciences, China Three Gorges University, Yichang, Hubei, 443003, People's Republic of China

*These authors contributed equally to this work

Correspondence: Zhitao Feng, College of Medicine and Health Science, China Three Gorges University, Yichang, Hubei, 443002, People's Republic of China, Tel +86-0717-6396558, Email fengzhitao2008@126.com; Huilian Huang, The Second People's Hospital Affiliated to Three Gorges University, Yichang Second People's Hospital, Yichang, Hubei, 443000, People's Republic of China, Tel +86-13607207438, Email ann19830418@163.com

Purpose: To establish an interpretable deep learning framework for automated classification of rheumatoid arthritis (RA) in hand radiographs, with emphasis on elucidating model decision-making patterns and enabling clinical translation through web-based deployment.

Patients and Methods: A retrospective multicenter study analyzed 1,655 hand radiographs (809 RA patients, including early RA cases, and 846 healthy controls). Enhanced data (random rotation, brightness/contrast adjustment) was applied to the collected X-ray images to improve the model's generalization ability and performance. Subsequently, A lightweight Visual Geometry Group (VGG)-8 convolutional neural network was trained and validated using processed hand X-ray images. This model has the ability to distinguish RA patients from healthy controls. The interpretability of the model was systematically evaluated using both Gradient-weighted Class Activation Mapping (Grad-CAM) and Shapley Additive Explanations (SHAP). Finally, a web application was developed using Streamlit that supports JPEG input, helps to address the clinical practicality of the model.

Results: For distinguishing RA patients from healthy individuals, the classifier achieved excellent training performance (AUC=0.99, accuracy=0.94) and generalizable testing metrics (AUC=0.81, accuracy=0.74). Specifically, the model was successfully constructed and demonstrated good performance in external validation. Interpretability analysis revealed areas of pathological significance, with Grad CAM heatmaps highlighting structural abnormalities (joint space stenosis, bone erosion, trabecular structural changes), and SHAP values analysis identifying metacarpophalangeal and wrist joints as key predictive features. A web application developed using Python and Streamlit framework can assist in the diagnosis of RA hand X-ray images in clinical practice.

Conclusion: This work advances clinical diagnosis, including early RA patients, by integrating deep learning with interpretable decision paths in hand radiographic analysis, while helping clinicians to use the model more proficiently. The framework provides both diagnostic assistance and educational insights into RA radiographic markers.

Keywords: radiograph interpretation, rheumatoid arthritis, visual geometry group, interpretability analysis, translational medicine

Introduction

Rheumatoid Arthritis (RA) is a chronic systemic autoimmune disease characterized by synovitis as its pathological basis, with predominant clinical manifestations of pain, swelling, and morning stiffness in small joints, particularly the proximal interphalangeal joints, metacarpophalangeal joints, wrists, and knees.[1] Delayed diagnosis and intervention in RA can result in progressive joint cartilage and bone destruction, leading to deformities and functional impairment, thereby adversely

Received: 14 June 2025
Accepted: 30 October 2025
Published: 19 November 2025

affecting long-term prognosis and quality of life.[2] Medical imaging plays an indispensable role in rheumatoid arthritis (RA) diagnosis, enabling assessment of synovitis, cartilage degradation, and bone erosion. Despite advancements in and accessibility of various imaging modalities, including ultrasound, magnetic resonance imaging (MRI), and conventional radiography, plain radiography remains essential for RA evaluation.[3,4] Particularly, conventional radiography of the hands and wrists remains the most widely utilized imaging modality for assessing structural joint damage in RA, owing to its operational simplicity, cost-effectiveness, and widespread availability.[5] RA diagnosis depends on synthesizing multiple parameters: clinician expertise, patient symptomatology, imaging findings, and laboratory data. This integrative process is inherently complex and time-intensive, with diagnostic accuracy vulnerable to subjective interpretation and incomplete data synthesis. Such limitations may result in diagnostic errors or oversight. Consequently, developing efficient and objective strategies, particularly those based on the widely available hand radiographs, to enhance RA diagnostic accuracy represents a crucial clinical priority.

The limitations of conventional diagnosis have spurred interest in artificial intelligence (AI)-assisted solutions. Deep learning, particularly convolutional neural networks (CNNs), has emerged as a powerful tool for automating image analysis tasks. CNNs excel at extracting discriminative features from medical images, enabling objective and reproducible disease classification and prediction.[6–8] This potential has been increasingly explored in the context of RA. Several pioneering studies have demonstrated the efficacy of deep learning models in automating the detection of radiographic joint damage and predicting RA progression from hand radiographs.[9–13] For instance, Peng et al conducted a comprehensive evaluation of five mainstream CNN architectures, demonstrating that models like GoogLeNet could achieve an exceptional AUC of 97.80% and a sensitivity of 100.0% for RA recognition from hand radiographs.[10] Beyond utilizing existing architectures, Kesavapillai et al developed a specialized CNN model (RA-XTNet), which outperformed standard pre-trained networks, achieving a classification accuracy of 90% for hand radiographs.[11] These studies underscore the feasibility and evolving efficacy of using deep learning to augment RA diagnosis based on hand radiographs.

However, the adoption of "black-box" deep learning models in clinical practice is often hindered by their lack of interpretability. Clinicians need to understand the rationale behind a model's prediction to trust and act upon it. Techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) address this by generating visual explanations that highlight the image regions most influential to the model's decision.[14–17] While Grad-CAM has proven invaluable in explaining CNN-based models in other radiological domains, such as identifying abnormalities in wrist and elbow radiographs,[18] localizing rib fractures in computed tomography (CT) scans,[19] and interpreting breast cancer detection in mammograms,[20] its application specifically within RA research, particularly for interpreting models trained on hand radiographs, remains relatively unexplored. Similarly, SHapley Additive exPlanations (SHAP) analysis, which quantifies feature contribution and has gained traction in interpreting complex machine learning models in healthcare,[21] has seen limited use in this niche. Moreover, the entire process from model construction to clinical application is of great significance. Therefore, there is an urgent need to develop an accurate RA diagnostic model, integrate interpretable methods, and build a work platform based on this model to validate its clinical knowledge-based decision-making process.

Accordingly, this study investigates the feasibility of applying VGG networks to radiographic RA diagnosis, intending to facilitate accurate and efficient clinical detection. This study has three characteristics: First, a CNN-based automatic diagnosis model is proposed to achieve accurate and fast diagnosis, including early RA cases. Second, to enhance the model's reliability and interpretability, Grad CAM and Shapley value analysis were used for comprehensive visualization to demonstrate whether samples from the same category can be effectively clustered. Third, an integrated data model application analysis framework has been established to allow clinical doctors to simply input hand radiographs and immediately obtain RA diagnostic predictions. This streamlined approach not only facilitates clinical implementation but also demonstrates the potential for AI to provide more accessible and efficient services in rheumatology practice and related medical fields.

This study aims to develop and validate a deep learning-based system for automatically distinguishing RA patients from normal adults using hand X-ray images. By integrating Grad CAM for visual localization and SHAP values analysis for feature importance quantification, the "black box" of this model attempts to be revealed. By building an online APP based on Streamlit, it provides transparent and reliable predictive explanations for clinical doctors to ensure that the model's decisions are based on radiological-related features to promote clinical adoption.

Method

Study Population

A retrospective analysis was conducted on 1655 hand X-ray images collected from patients with rheumatoid arthritis (RA) and healthy controls at Yichang Central People's Hospital and Yichang Second People's Hospital between January 2017 and April 2024. All images were randomized and divided into training and testing sets by institution. The planned work has been approved by the Ethics Committee (Human Research) of the First Clinical Medical College of China Three Gorges University and the Second affiliated hospital of China Three Gorges University (Approval No.2024–070-01 and 202433). The study is retrospective in design and uses only anonymized clinical data; patients cannot be contacted; the research is essential for the public interest of medical science; the waiver does not compromise patients' rights or welfare. All patient data have been anonymized, with identifying information removed, and will be used solely for the purposes of this study. The research strictly adheres to the principles of the Declaration of Helsinki, ensuring data security and protection of patient privacy.

The inclusion criteria comprised: RA patients meeting the 2010 ACR/EULAR classification criteria,[22,23] with completed hand radiographs and complete clinical data; and healthy controls exhibiting neither joint pain nor swelling persisting beyond one week; while exclusion criteria included: RA patients with incomplete or suboptimal quality radiographs; and non-RA individuals lacking wrist/hand joint pain, persistent swelling (>1 week), or history of trauma.

Equipment and Methods

All hand radiographs were obtained by certified radiology technicians using a Wandong VX3733 X-ray scanner under standardized protocols with 60 kV tube voltage, 3 mAs exposure dose, and 100 cm film focus distance, capturing both posteroanterior and lateral views in JPEG format.

X-Ray Image Model Construction

Data Collection

Hand radiographs of RA patients and healthy controls were collected at a 1:1 ratio. To validate model generalizability, we selected 50 normal and 50 RA hand radiographs from these two sources as an independent test set (Figure 1).

Image Preprocessing

All collected images underwent standardized preprocessing, including resizing to 256×256 pixels, followed by center-cropping to 224×224 pixels, vectorization, and sym8 wavelet transformation, with global normalization to maintain optimal pixel resolution while preserving diagnostically relevant information. To address common radiographic artifacts (including variable data types, excessive unfilled regions, abnormal angulations, and image darkening), we implemented

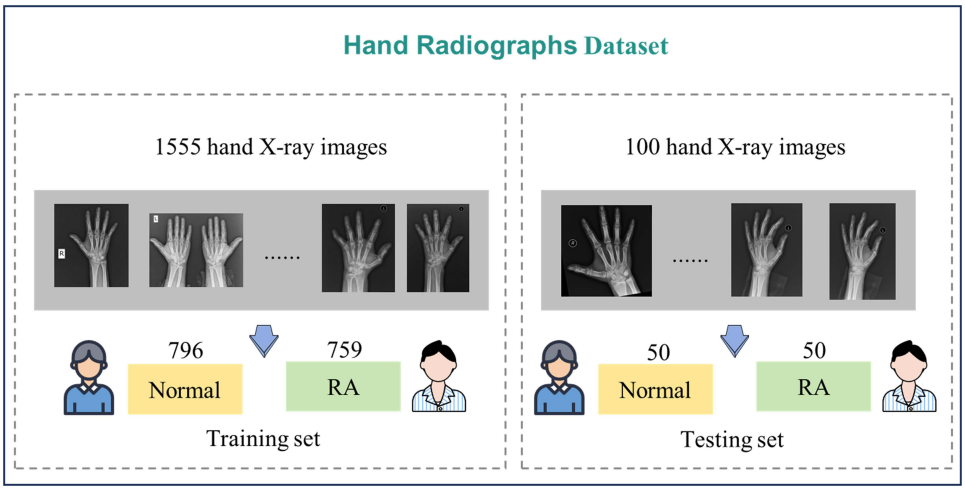

Figure 1 Schematic diagram of the training and testing set construction framework.
Abbreviations: RA, Rheumatoid Arthritis; Normal, healthy controls.

https://doi.org/10.2147/ITT.S547159

quality enhancement techniques through strategic cropping and brightness adjustment. The standardized processing pipeline consisted of: conversion to PyTorch tensors, pixel value normalization to [0,1], and channel-wise standardization using mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] values. Data augmentation employed three strategies: randomized rotation only (±45°), randomized brightness adjustment only (±40%, scaling pixel values 0.6–1.4 with 0–255 clipping), and their combination (±45° rotation followed by ±40% brightness adjustment) (Figure 2).

Model Development

A simplified VGG8 architecture has been implemented, which maintains the iconic design principles of VGG, namely the use of small convolution kernels (3x3) to achieve detailed feature extraction with reduced parameters, ultimately achieving a fully connected layer for binary classification. The architecture consists of: five convolutional blocks containing eight convolutional layers for efficient computation while maintaining accuracy, 2×2 max-pooling layers with stride 2 for spatial reduction, and a classifier comprising a 25,088-dimensional input layer followed by two 512-dimensional hidden layers with ReLU activation and dropout (p=0.5), terminating in a 2-node output layer (RA vs non-RA) with Softmax activation (Figure 3).

Interpretability Analysis

To enhance the transparency of our deep learning model's decision-making process, we implemented a dual interpretability framework combining: Gradient-weighted Class Activation Mapping (Grad-CAM) for spatial visualization of critical image regions influencing predictions, and SHAP analysis to quantitatively assess feature contributions. This

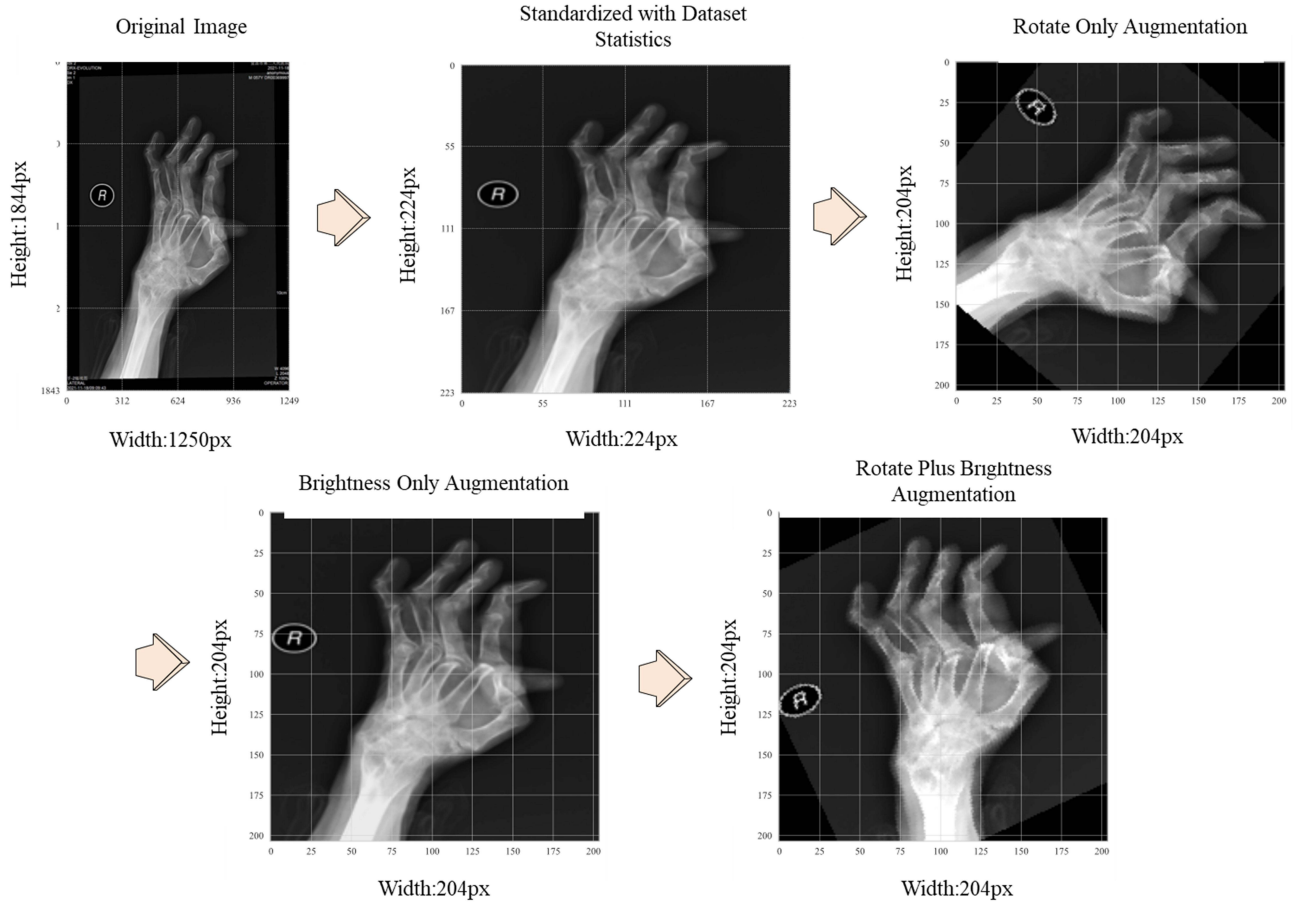

Figure 2 Image processing pipeline. Data Processing. Resize: Resize all images to 256×256 pixels. CenterCrop: Crop a 224 × 224-pixel center patch from each. ToTensor: Convert images to PyTorch tensors and scale pixel values to [0, 1]. Normalize: Standardize each channel with mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Data Augmentation. Rotation only: rotate within ±45. Brightness only: adjust brightness by ±40% (scaled by np.random.uniform (0.6, 1.4)). Rotation + Brightness: combine the two operations above for cumulative augmentation. "R" indicates the right side.

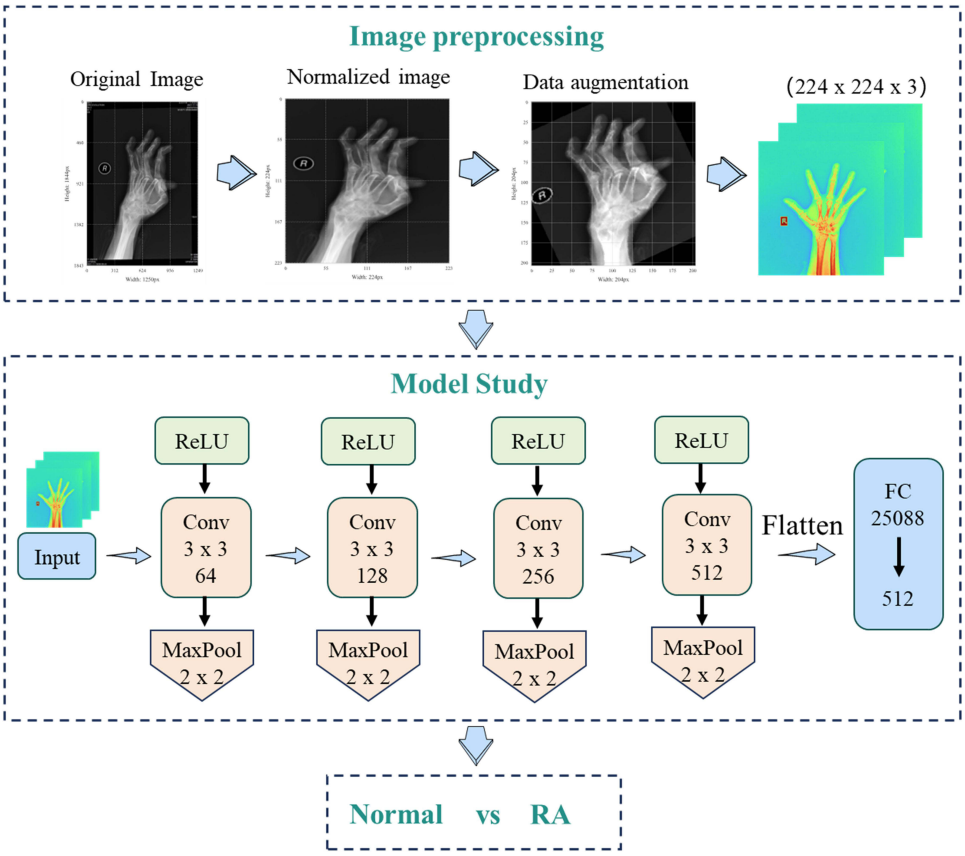

Figure 3 Schematic diagram of the model architecture.
Abbreviations: RA, Rheumatoid Arthritis; Normal, healthy controls.

multimodal approach enables both heatmap-based localization of pathological features and mathematical quantification of input feature importance through additive feature attribution methods.

Web Application Development Based on Streamlit Framework

Based on the VGG model, X-ray classification tasks were carried out, and a Python and Streamlit framework was used to build a web app to make the model more convenient and clinically applicable.

Statistical Analysis

Model performance was rigorously evaluated using standard classification metrics calculated from confusion matrices of both training and test sets: accuracy = (TP+TN)/(TP+TN+FP+FN), sensitivity (recall/true positive rate) = TP/(TP+FN), specificity = TN/(TN+FP), precision = TP/(TP+FP), and area under the receiver operating characteristic curve (AUC), where TP=true positives, FP=false positives, TN=true negatives, and FN=false negatives.

Results

Training Set Modeling

The training dataset comprised 796 hand radiographs from healthy individuals and 759 from RA patients (n=1,555), with an additional independent test set of 100 images (Table 1). Following comprehensive data augmentation and preprocessing, the model demonstrated robust performance with AUC values of 0.99 (training) and 0.81 (test), alongside accuracy rates of 0.94 (training) and 0.74 (test) (Figures 4 and 5 and Table 2). These results indicate successful feature extraction from hand radiographs, with strong discriminative capability maintained across both training and validation phases. The model was trained for 100 epochs using an Adam SGD optimizer (learning rate=0.05) to optimize training data fitting (Figure 5C).

https://doi.org/10.2147/ITT.S547159

Table 1 Data Distribution of Training and Testing Sets

	Training Set	Testing Set	Total
RA	759	50	809
Normal	796	50	846
Total	1555	100	1655

Notes: The dataset comprises 1,655 hand radiographs (training set: 1,555 images; testing set: 100 images).
Abbreviations: RA, Rheumatoid Arthritis; Normal, healthy controls.

Model Interpretability Analysis

The interpretability analysis combines Grad-CAM (Figure 6A) and Shapley value visualization (Figure 6B) to elucidate the model's decision-making process. As shown in Figure 6A, the heatmap overlaid on RA patient radiographs revealed that the model primarily focused on joint spaces and bony articulations (red/yellow regions, particularly in the wrist joints) - anatomical areas clinically known to demonstrate characteristic RA changes. Conversely, blue/green regions indicated minimal diagnostic contribution.

Shapley analysis (Figure 6B) quantitatively confirmed these findings, with red-highlighted areas (potential pathological regions) showing strong positive contributions to "RA" classification, while blue regions (normal bone structures) supported "NORMAL" classification. Notably, the model's attention patterns demonstrate remarkable concordance with established clinical diagnostic priorities, validating its learning of medically relevant features.

Web Application Development Based on Streamlit Framework

Building upon the VGG-based X-ray classification model, Python and the Streamlit framework were used to develop an accessible web application, establishing a comprehensive pipeline from data processing to clinical implementation. As illustrated in Figure 7, the application features three functional modules: user login interface, sample image gallery, and diagnostic prediction interface with integrated Shapley value visualization. This implementation not only operationalizes

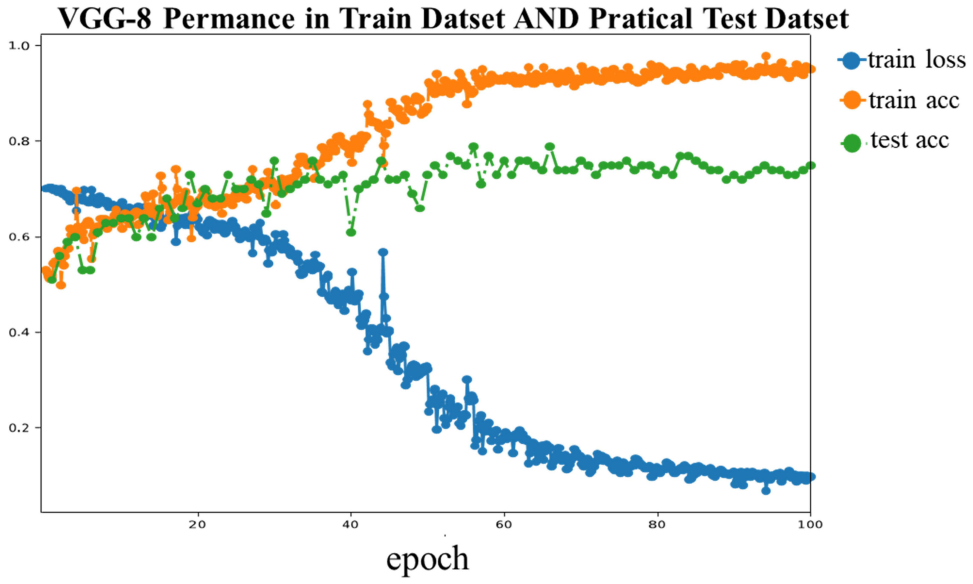

Figure 4 Model learning and test performance curve.

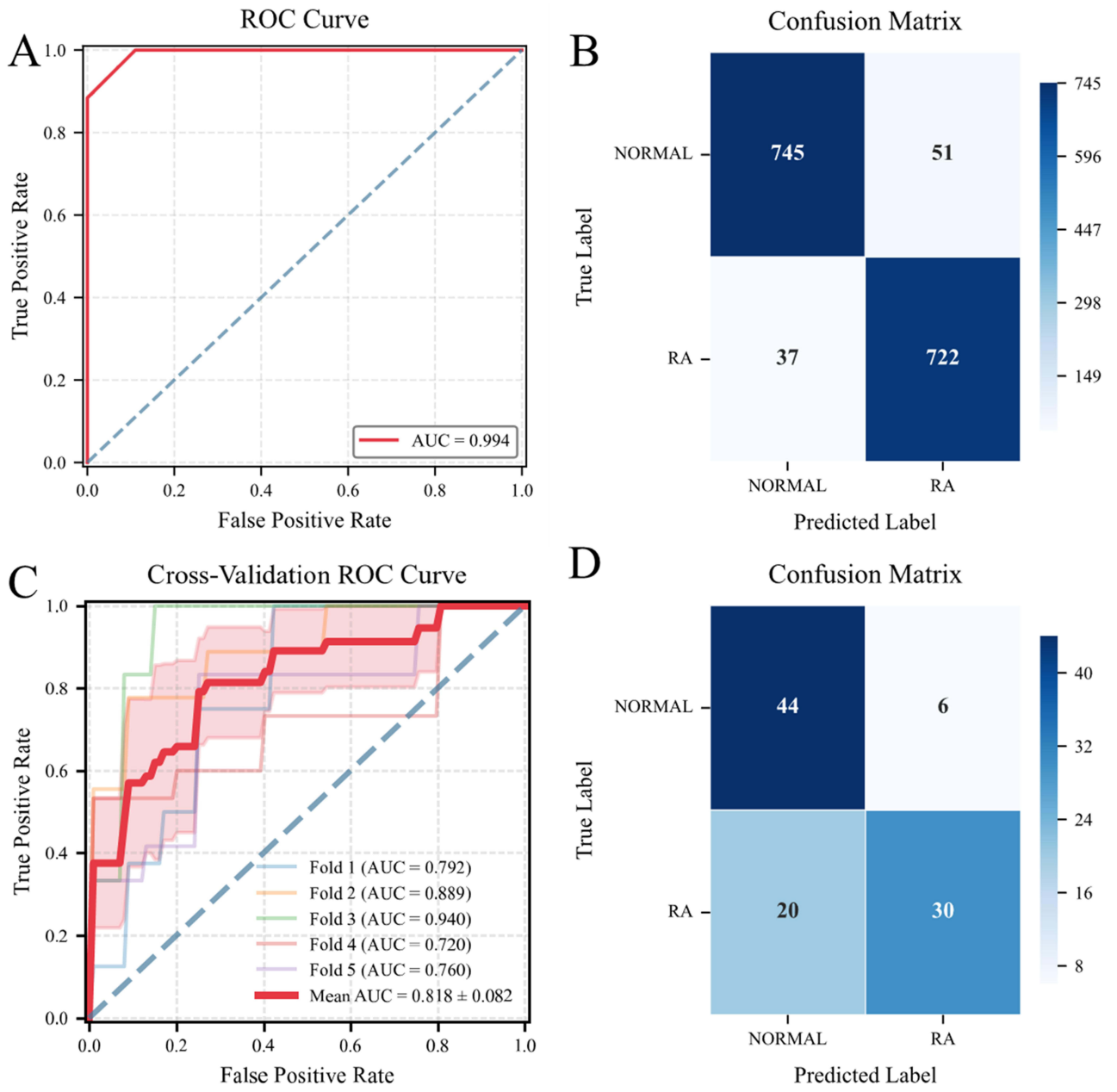

Figure 5 Model evaluation metrics: (**A**) Training set ROC curve, (**B**) Training set confusion matrix, (**C**) Testing set ROC curve, (**D**) Testing set confusion matrix.

our interpretable AI system but also significantly reduces the technical barrier for clinical use. The developed framework represents a complete workflow encompassing data preprocessing, model development, application deployment, and performance evaluation, serving as a replicable template for future medical AI research and translation.

Table 2 Performance Metrics of the Model on Training and Testing Sets

	Accuracy	Sensitivity	Specificity	Precision	AUC
Training set	0.94	0.95	0.94	0.93	0.99
Testing set	0.74	0.6	0.88	0.83	0.81

Notes: Model performance was evaluated using standard metrics calculated from confusion matrices: accuracy = (TP+TN)/total, sensitivity = TP/(TP+FN), specificity = TN/(TN+FP), precision = TP/(TP+FP), and AUC (area under the ROC curve).
Abbreviations: AUC, area under the curve; TP, true positive; FP, false positive; TN, true negative; FN, false negative.

https://doi.org/10.2147/ITT.S547159

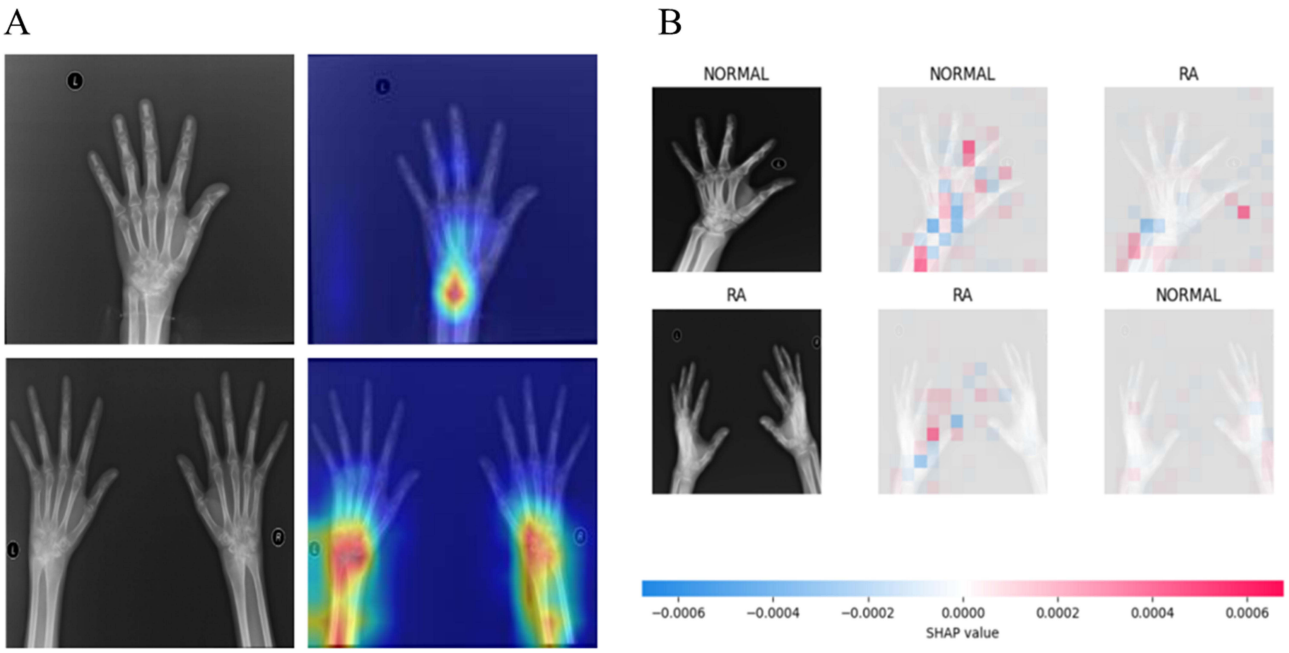

Figure 6 Interpretability analysis: (**A**) Grad-CAM visualization showing regions with significant positive contributions to RA classification (red/yellow areas); (**B**) SHAP analysis demonstrating feature importance (darker colors indicate higher contribution).
Abbreviations: RA, rheumatoid arthritis; NORMAL, healthy controls.

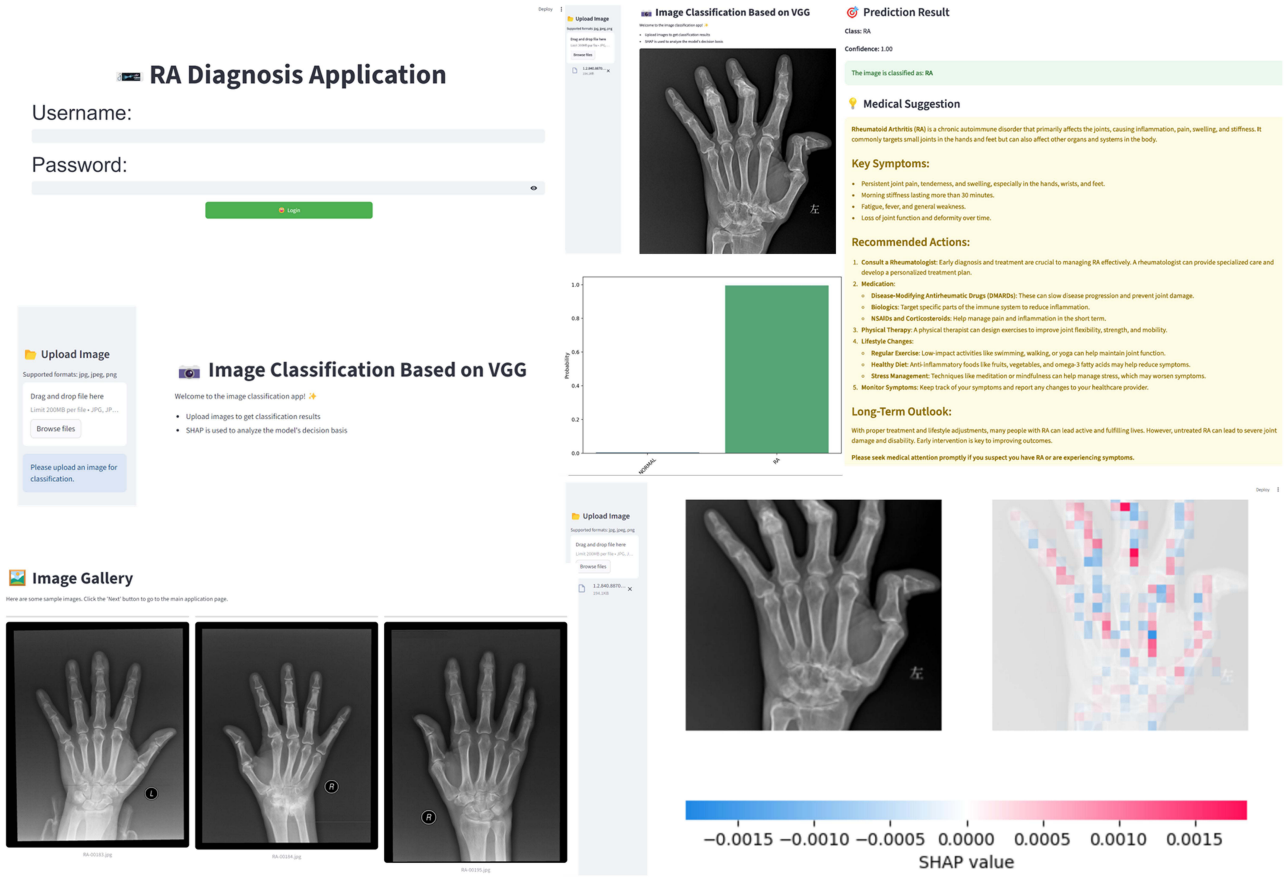

Figure 7 Demonstration of the web-based diagnostic application.

Discussion

In rheumatology, AI models have been predominantly applied to X-ray image analysis, with convolutional neural networks (CNNs) establishing themselves as the leading architecture for AI-based radiological image interpretation.[24] While CNNs have demonstrated robust capability in identifying key pathological features of RA in hand radiographs, including joint space narrowing, bone erosions, and osteoporosis.[25–27] The VGG architecture, as a classical CNN variant, offers unique advantages for medical image analysis due to its deep convolutional design and exceptional compatibility with imaging data characteristics.[28] This study was designed to evaluate the feasibility and applicability of VGG models for computer-aided diagnosis of RA using hand radiographs. A VGG-based convolutional neural network was developed and systematically validated its performance in RA image classification tasks. The experimental results demonstrated robust diagnostic capability, with area under the curve (AUC) values of 0.99 (training set) and 0.81 (test set), accompanied by accuracy rates of 0.94 and 0.74, respectively. These findings substantiate that the VGG architecture achieves both high diagnostic precision (training AUC 0.99) and clinically meaningful generalizability (test AUC 0.81) in discriminating between normal and abnormal hand radiographs in RA, providing empirical evidence for developing reliable RA diagnostic support systems.

Moreover, Grad CAM and Shapley value decomposition were used for in-depth analysis of the decision-making process of the model. The Grad-CAM technique visualizes gradient-weighted feature maps to intuitively reveal the model's focus areas on pathological regions (Figure 6A), while Shapley value analysis quantifies the contribution of each input feature to the model output from a game-theoretic perspective (Figure 6B). Notably, the model's attention patterns demonstrated strong concordance with clinically relevant regions traditionally emphasized by rheumatologists. This dual interpretability approach not only significantly enhanced model transparency and explainability,[29] but more importantly, uncovered fundamental alignment between the model's learning mechanism and clinical diagnostic reasoning. These findings provide crucial theoretical support for the clinical translation of VGG-based radiographic diagnosis in RA management.

Furthermore, the Python ecosystem and Streamlit framework were utilized to develop a lightweight web application that enables efficient deployment and user-friendly access to the RA diagnostic model. This implementation provides a practical technical solution for personalized precision medicine, facilitating multi-center collaboration and data sharing in rheumatology practice. The integrated data-model-application-analysis framework established in this study combines the powerful pattern recognition capabilities of deep learning with advanced interpretability analysis. This synergistic approach not only identifies critical diagnostic regions through explainable AI techniques but also effectively assists clinicians in rapidly extracting diagnostic patterns and clinical insights. While specifically designed for RA assessment, the framework's modular architecture demonstrates strong potential for rapid adaptation to other musculoskeletal disorders and related disease domains.

Notably, AI applications in radiographic diagnosis of RA have demonstrated promising yet heterogeneous outcomes. Early studies utilizing small datasets reported limited accuracy - for instance, a 2020 CNN-based study achieved 73.33% accuracy using 180 hand radiographs for RA/normal classification.[13] Subsequent improvements through dataset expansion and architectural optimization yielded 94.46% accuracy, with 0.95 sensitivity and 0.82 specificity from 290 radiographs.[30] A pivotal advancement came from Üreten et al,[31] who employed transfer learning with pretrained VGG-16 to circumvent data limitations, achieving 90.7% accuracy, 92.6% sensitivity, 88.7% specificity, 89.3% precision, and 97% AUC. Ma et al[32] further escalated the benchmark using a multicenter dataset (9,964 radiographs from 8,533 patients across 7 hospitals), attaining 0.955 AUC for RA detection against normal controls. Peng et al[10] conducted the first comprehensive architecture comparison (AlexNet / VGG / GoogLeNet / ResNet / EfficientNet), reporting >90% AUC and >98% sensitivity for RA detection, and >77% AUC with >80% sensitivity for staging - albeit limited by a small 240-image dataset. Nevertheless, the majority of studies[10,13,30] rely on relatively small-scale datasets and imbalanced data, which may lead to model overfitting and consequently diminish generalizability in real-world clinical settings. Secondly, while certain models demonstrate exceptional performance metrics,[31,32] their clinical utility and physician trustworthiness are substantially constrained by the absence of integrated application deployment frameworks and comprehensive interpretability analyses. Furthermore, although recent comparative studies of multiple architectures[10] report outstanding performance (eg, GoogLeNet achieving 97.80% AUC and 100.0% sensitivity), these results still require validation through larger and more diverse datasets encompassing varied demographic populations and clinical scenarios to ensure robust external validity.

In this study, the lightweight VGG-8 architecture was chosen for several practical and clinical considerations. Due to its interpretability and compatibility, the order and transparent structure of VGG are very suitable for interpretability techniques such as Grad-CAM and SHAP, which are the core of our research objectives. Additionally, considering clinical deployment goals, lighter models can facilitate easier integration into web-based applications and reduce inference time, which is crucial for providing real-time diagnostic support. Furthermore, as shown in our results, the VGG-8 model achieved strong performance on a medium-sized dataset (training AUC=0.99, testing AUC=0.81), demonstrating its adequacy for the task. Compared to several popular models currently, although EfficientNet-B0 performs better in light to medium models, its network structure is complex, and interpreting the results requires more careful verification, indicating that the model may require longer computation time.[33] In addition, there is a strong dependence on pre-trained weights, which requires fine-tuning during domain transfer, which is not conducive to clinical deployment. Similar to the EfficientNet-B0 model, the Vision Transformer (ViT) model requires longer training time due to its transformer-based architecture.[34] Moreover, due to the lack of some key "inductive biases" in computer vision tasks, the Vision Transformer model requires massive amounts of data for learning the model.[35] Therefore, the VGG-8 model is more suitable for conducting this study.

In CNN model construction, both data quantity and quality are paramount for predictive performance. Adequate dataset size ensures proper model training, enhances performance,[36] mitigates overfitting, and improves robustness.[37,38] Additionally, class imbalance—a prevalent challenge in deep learning where certain classes dominate the training set—can introduce prediction bias toward majority classes and compromise accuracy.[39,40] Addressing this imbalance is critical, as it directly enhances model generalizability.[41,42] To overcome these limitations and address prior shortcomings in RA classification research, this study established a multicenter-derived, large-scale dataset with balanced class distribution. This approach minimizes performance evaluation bias caused by data imbalance, ensuring reliable and generalizable classification outcomes. Notably, to address the diagnostic challenge of early-stage RA cases that often show negative findings on X-ray examinations, this study innovatively incorporated clinically confirmed early RA cases with non-characteristic radiographic changes and developed a diagnostic model demonstrating high sensitivity (Table 2) for detecting subtle pathological features in the training set (0.94 accuracy) while maintaining preliminary generalizability in the test set (0.6 sensitivity and 0.74 accuracy), indicating robust feature extraction capability for clinically challenging cases with ambiguous imaging findings. Future studies should focus on expanding the sample dataset and further investigating these subtle imaging features to systematically identify quantitative imaging biomarkers for early RA progression,[43] thereby providing critical evidence for clinical early intervention strategies. Moreover, deep learning models are often regarded as "black boxes"[44,45] due to their inherent lack of interpretability, which significantly limits their clinical adoption in medical image analysis. To address this critical limitation, our study implemented comprehensive interpretability analyses using both Grad-CAM and Shapley value methods. These approaches not only enhanced decision transparency but also revealed remarkable alignment between the model's focus areas and clinically relevant regions, demonstrating that the model effectively learned medically meaningful patterns consistent with established diagnostic principles.

While AI-driven diagnostic support tools show growing potential in rheumatology, their clinical integration remains limited by issues of accuracy, interpretability, and workflow compatibility.[46] Existing patient-facing tools, such as symptom checkers (eg, Ada) and referral systems (eg, Rheport), rely primarily on symptomatic input and demonstrate modest diagnostic accuracy (52–63%) even in high-prevalence settings.[47] In contrast, our Streamlit-based web application leverages radiographic data to provide objective, image-based decision support specifically for RA hand X-ray interpretation.

The developed application offers several distinct advantages: it integrates directly with image-based inputs, supporting standard medical imaging formats and providing real-time predictions with interpretable outputs (SHAP and Grad-CAM visualizations); it achieves clinically relevant performance (test AUC: 0.81, accuracy: 74%) with a focus on structural abnormalities such as joint space narrowing and erosions; and it is designed for seamless integration into clinical workflows via an intuitive, clinician-oriented interface that requires no specialized computational skills.

Unlike general-purpose symptom checkers, our tool provides specialized, reproducible, and transparent decision support for radiographic assessment—a critical component of RA diagnosis and monitoring. This represents a meaningful step toward bridging the gap between AI innovation and clinical practice in rheumatology.

Although this study has achieved promising results, some limitations should be acknowledged. Although the multi-center dataset in this study includes a certain number of hand X-rays from RA patients, its size is still insufficient to meet deep learning standards. There is no better validation method, such as k-fold cross-validation, and it is limited to one region geographically/racially, which may limit the universality of the model.[48] Second, the current binary classification framework (RA vs normal) has not been fully validated for differential diagnosis against other arthritic conditions like osteoarthritis (OA) or ankylosing spondylitis (AS). Most importantly, consistent with the 2010 ACR/EULAR RA classification criteria,[22,23] radiographic findings alone are insufficient for definitive diagnosis - clinical confirmation requires comprehensive evaluation of: joint involvement patterns, serological abnormalities, acute-phase reactants, and symptom duration. Future web-based systems must therefore integrate multimodal data, including advanced imaging analyses, genetic profiles, clinical histories, biomarker patterns, and treatment response predictions, to achieve clinically reliable diagnoses.

Future studies should further expand the scale of datasets and consider cross-institutional data collection and integration to enhance model robust and generalization capability. At the model optimization and extension level, structural and parameter tuning will be performed based on the VGG architecture to improve its diagnostic efficacy for RA hand radiographs. Specifically, we propose to introduce attention mechanisms[49] to strengthen the model's ability to extract and classify complex imaging features, while continuing to develop interpretable AI frameworks such as attention heatmaps and saliency maps. These approaches will systematically improve algorithmic transparency and gradually bridge the gap between model performance and clinical acceptability. Inspired by the residual learning framework,[50] subsequent studies could incorporate identity shortcut connections to mitigate degradation problems encountered when scaling to deeper architectures. In clinical practice, multimodal datasets incorporating hand X-ray, ultrasound, and MRI for diverse diseases such as RA, OA, and AS can be constructed utilizing databases like the China Rheumatism Data Center (CRDC). By developing multi-task classification models[51] and exploring transfer learning-based multimodal fusion diagnostic approaches, complementary information from different imaging modalities can be leveraged to enhance diagnostic accuracy. Meanwhile, transfer learning techniques enable the reuse of pretrained model parameters, accelerating model optimization for RA diagnosis and treatment scenarios, thereby systematically improving differential diagnostic precision in complex clinical settings. Further, we will integrate clinical indicators, genetic data, laboratory test results, and radiomics features to construct a multi-source heterogeneous data fusion model. This approach aims to elucidate the underlying disease mechanisms and enhance predictive performance, thereby supporting personalized precision diagnosis for RA. Concurrently, we plan to conduct large-scale, multicenter prospective clinical trials for external validation of the model, systematically evaluating its real-time diagnostic accuracy and integration efficacy within clinical workflows. These efforts will facilitate the translation of deep learning models from experimental tools into reliable clinical decision-support systems. Ultimately, the fused model will be embedded into a mobile application (APP) framework, establishing an intelligent diagnostic closed-loop system to optimize healthcare delivery efficiency and quality.

Conclusion

In summary, this study demonstrated moderate diagnostic performance by using the VGG8 model to classify and diagnose X-ray images of the hands, including early RA patients. Through Grad-CAM and SHAP analysis, the "black box" of the model was revealed, and an online APP was built based on Streamlit to assist clinical doctors in decision-making, demonstrating the enormous potential of deep learning in medical imaging diagnosis.

Acknowledgments

We would like to express our sincere gratitude to Yunzhen Zhan from China Pharmaceutical University for his invaluable assistance and guidance in data analysis, which significantly contributed to the success of this study.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This project was supported by grants from the National Natural Science Foundation of China (No.81703783 and 82274333); the Hubei Provincial Natural Science Foundation (No. 2017CFB126); the fund for Hubei Provincial Medical Youth TAop-notch Talents Project (EWT-2023) and Hubei Provincial Natural Science Foundation Innovation and Development Joint Fund (Grant No. 2025AFD308).

Disclosure

The authors declare no conflict of interest.

References

1. Di Matteo A, Bathon JM, Emery P. Rheumatoid arthritis. *Lancet*. 2023;402(10416):2019–2033. doi:10.1016/S0140-6736(23)01525-8
2. Brown P, Pratt AG, Hyrich KL. Therapeutic advances in rheumatoid arthritis. *BMJ*. 2024;384:e070856. doi:10.1136/bmj-2022-070856
3. Gilvaz VJ, Reginato AM. Artificial intelligence in rheumatoid arthritis: potential applications and future implications. *Front Med Lausanne*. 2023;10:1280312. doi:10.3389/fmed.2023.1280312
4. Bohndorf K, Schalm J. Diagnostic radiography in rheumatoid arthritis: benefits and limitations. *Baillieres Clin Rheumatol*. 1996;10(3):399–407. doi:10.1016/S0950-3579(96)80038-0
5. Vyas S, Bhalla AS, Ranjan P, Kumar S, Kumar U, Gupta AK. Rheumatoid arthritis revisited - advanced imaging review. *Pol J Radiol*. 2016;81:629–635. doi:10.12659/PJR.899317
6. McMaster C, Bird A, Liew DFL, et al. Artificial intelligence and deep learning for rheumatologists. *Arthritis Rheumatol*. 2022;74(12):1893–1905. doi:10.1002/art.42296
7. Bird A, Oakden-Rayner L, McMaster C, et al. Artificial intelligence and the future of radiographic scoring in rheumatoid arthritis: a viewpoint. *Arthritis Res Ther*. 2022;24(1):268. doi:10.1186/s13075-022-02972-x
8. Momtazmanesh S, Nowroozi A, Rezaei N. Artificial intelligence in rheumatoid arthritis: current status and future perspectives: a state-of-the-art review. *Rheumatol Ther*. 2022;9(5):1249–1304. doi:10.1007/s40744-022-00475-4
9. Wang HJ, Su CP, Lai CC, et al. Deep learning-based computer-aided diagnosis of rheumatoid arthritis with hand X-ray images conforming to modified total sharp/van der heijde score. *Biomedicines*. 2022;10(6):1355. doi:10.3390/biomedicines10061355
10. Peng Y, Huang X, Gan M, Zhang K, Chen Y. Radiograph-based rheumatoid arthritis diagnosis via convolutional neural network. *BMC Med Imaging*. 2024;24(1):180.
11. Kesavapillai AR, Aslam SM, Umapathy S, Almutairi F. RA-XTNet: a novel CNN model to predict rheumatoid arthritis from hand radiographs and thermal images: a comparison with CNN transformer and quantum computing. *Diagnostics*. 2024;14(17):1911. doi:10.3390/diagnostics14171911
12. Izumi K, Suzuki K, Hashimoto M, et al. Ensemble detection of hand joint ankylosis and subluxation in radiographic images using deep neural networks. *Sci Rep*. 2024;14(1):7696. doi:10.1038/s41598-024-58242-0
13. Üreten K, Erbay H, Maraş HH. Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network. *Clin Rheumatol*. 2020;39(4):969–974. doi:10.1007/s10067-019-04487-4
14. Karim MR, Islam T, Shajalal M, et al. Explainable AI for bioinformatics: methods, tools and applications. *Brief Bioinform*. 2023;24(5):1–22. doi:10.1093/bib/bbad236
15. Sheu RK, Pardeshi MS. A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system. *Sensors*. 2022;22(20):8068. doi:10.3390/s22208068
16. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners - saliency-based XAI approaches. *Eur J Radiol*. 2023;162:110787. doi:10.1016/j.ejrad.2023.110787
17. Nazir S, Dickson DM, Akram MU. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Comput Biol Med*. 2023;156:106668. doi:10.1016/j.compbiomed.2023.106668
18. Lysdahlgaard S. Utilizing heat maps as explainable artificial intelligence for detecting abnormalities on wrist and elbow radiographs. *Radiography*. 2023;29(6):1132–1138. doi:10.1016/j.radi.2023.09.012
19. Wu L, Chen H, Li P, Yang K. A novel ensemble approach for rib fracture detection and visualization using CNNs and Grad-CAM. *Ann Ital Chir*. 2025;96(1):86–97. doi:10.62713/aic.3666
20. Cerekci E, Alis D, Denizoglu N, et al. Quantitative evaluation of saliency-based explainable artificial intelligence (XAI) methods in deep learning-based mammogram analysis. *Eur J Radiol*. 2024;173:111356. doi:10.1016/j.ejrad.2024.111356
21. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. *Comput Biol Med*. 2023;166:107555. doi:10.1016/j.compbiomed.2023.107555
22. Neogi T, Aletaha D, Silman AJ, et al. The 2010 American college of rheumatology/European league against rheumatism classification criteria for rheumatoid arthritis: phase 2 methodological report. *Arthritis Rheum*. 2010;62(9):2582–2591. doi:10.1002/art.27580
23. Funovits J, Aletaha D, Bykerk V, et al. The 2010 American college of rheumatology/European league against rheumatism classification criteria for rheumatoid arthritis: methodological report Phase I. *Ann Rheum Dis*. 2010;69(9):1589–1595. doi:10.1136/ard.2010.130310
24. Perronne L, Binvignat M, Foulquier N, et al. Algorithmic approaches in hand imaging for rheumatic musculoskeletal diseases: a systematic literature review. *Semin Arthritis Rheum*. 2025;73:152750. doi:10.1016/j.semarthrit.2025.152750
25. Ichikawa S, Kamishima T, Sutherland K, et al. Computer-based radiographic quantification of joint space narrowing progression using sequential hand radiographs: validation study in rheumatoid arthritis patients from multiple institutions. *J Digit Imaging*. 2017;30(5):648–656. doi:10.1007/s10278-017-9970-9
26. Langs G, Peloschek P, Bischof H, Kainberger F. Automatic quantification of joint space narrowing and erosions in rheumatoid arthritis. *IEEE Trans Med Imaging*. 2009;28(1):151–164. doi:10.1109/TMI.2008.2004401

27. Amani F, Amanzadeh M, Hamedan M, Amani P. Diagnostic accuracy of deep learning in prediction of osteoporosis: a systematic review and meta-analysis. *BMC Musculoskelet Disord*. 2024;25(1):991. doi:10.1186/s12891-024-08120-7
28. Rohrbach J, Reinhard T, Sick B, Duerr O. Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks. *Comput Electr Eng*. 2019;78:472–481. doi:10.1016/j.compeleceng.2019.08.003
29. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst*. 2021;32 (11):4793–4813. doi:10.1109/TNNLS.2020.3027314
30. Mate GS, Kureshi AK, Singh BK. An efficient CNN for Hand X-Ray classification of rheumatoid arthritis. *J Healthc Eng*. 2021;2021:6712785. doi:10.1155/2021/6712785
31. Ureten K, Maras HH. Automated classification of rheumatoid arthritis, osteoarthritis, and normal hand radiographs with deep learning methods. *J Digit Imaging*. 2022;35(2):193–199. doi:10.1007/s10278-021-00564-w
32. Ma Y, Pan I, Kim SY, Wieschhoff GG, Andriole KP, Mandell JC. Deep learning discrimination of rheumatoid arthritis from osteoarthritis on hand radiography. *Skeletal Radiol*. 2024;53(2):377–383. doi:10.1007/s00256-023-04408-2
33. Ulubaba HE, İ A, Çiftçi R, Ö E, Aldhahi MI. Deep learning for gender estimation using hand radiographs: a comparative evaluation of CNN models. *BMC Med Imaging*. 2025;25(1):260. doi:10.1186/s12880-025-01809-8
34. Rautaray J, Ali ABM, Kandpal M, et al. Leveraging FastViT based knowledge distillation with efficient net-B0 for diabetic retinopathy severity classification. *SLAS Technol*. 2025;33:100325. doi:10.1016/j.slast.2025.100325
35. Suman S, Tiwari AK, Sachan S, Singh K, Meena S, Kumar S. Severity grading of hypertensive retinopathy using hybrid deep learning architecture. *Comput Methods Programs Biomed*. 2025;261:108585. doi:10.1016/j.cmpb.2025.108585
36. Fang Y, Wang J, Ou X, et al. The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. *Phys Med Biol*. 2021;66(18):185012. doi:10.1088/1361-6560/ac2206
37. Rossi L, Fiorentino MC, Mancini A, Paolanti M, Rosati R, Zingaretti P. Generalizability and robustness evaluation of attribute-based zero-shot learning. *Neural Netwk*. 2024;175:106278. doi:10.1016/j.neunet.2024.106278
38. Sanaat A, Shiri I, Ferdowsi S, Arabi H, Zaidi H. Robust-deep: a method for increasing brain imaging datasets to improve deep learning models' performance and robustness. *J Digital Imaging*. 2022;35(3):469–481. doi:10.1007/s10278-021-00536-0
39. Zhou Y, Wu J, Xu X, Shi G, Liu P, Jiang L. Investigating perioperative pressure injuries and factors influencing them with imbalanced samples using a synthetic minority over-sampling technique. *Biosci Trends*. 2025;19(2):173–188. doi:10.5582/bst.2025.01013
40. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netwk*. 2018;106:249–259. doi:10.1016/j.neunet.2018.07.011
41. Huang Z, MacLachlan S, Yu L, et al. Generalization challenges in electrocardiogram deep learning: insights from dataset characteristics and attention mechanism. *Future Cardiol*. 2024;20(4):209–220. doi:10.1080/14796678.2024.2354082
42. Du J, Zhang X, Liu P, Vong CM, Wang T. An adaptive deep metric learning loss function for class-imbalance learning via intraclass diversity and interclass distillation. *IEEE Trans Neural Netw Learn Syst*. 2024;35(11):15372–15386. doi:10.1109/TNNLS.2023.3286484
43. Kim M, Yun J, Cho Y, et al. Deep learning in medical imaging. *Neurospine*. 2019;16(4):657–668. doi:10.14245/ns.1938396.198
44. Lee T, Natalwala J, Chapple V, Liu Y. A brief history of artificial intelligence embryo selection: from black-box to glass-box. *Hum Reprod*. 2024;39 (2):285–292. doi:10.1093/humrep/dead254
45. Singla S, Eslami M, Pollack B, Wallace S, Batmanghelich K. Explaining the black-box smoothly-A counterfactual approach. *Med Image Anal*. 2023;84:102721. doi:10.1016/j.media.2022.102721
46. Gilvaz VJ, Sudheer A, Reginato AM. Emerging artificial intelligence innovations in rheumatoid arthritis and challenges to clinical adoption. *Curr Rheumatol Rep*. 2025;27(1):28. doi:10.1007/s11926-025-01193-w
47. Knitza J, Tascilar K, Fuchs F, et al. Diagnostic accuracy of a mobile ai-based symptom checker and a web-based self-referral tool in rheumatology: multicenter randomized controlled trial. *J Med Internet Res*. 2024;26:e55542. doi:10.2196/55542
48. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health*. 2022;4 (6):e406–e14. doi:10.1016/S2589-7500(22)00063-2
49. Brauwers G, Frasincar F. A general survey on attention mechanisms in deep learning. *IEEE Trans Knowledge Data Eng*. 2023;35(4):3279–3298. doi:10.1109/TKDE.2021.3126456
50. Wu W, Liu S, Xia Y, Zhang Y. Dual residual attention network for image denoising. *Pattern Recogn*. 2024;149:110291. doi:10.1016/j.patcog.2024.110291
51. Zhang Y, Yang Q. A survey on multi-task learning. *IEEE Trans Knowledge Data Eng*. 2022;34(12):5586–5609. doi:10.1109/TKDE.2021.3070203