Human Interaction and Emerging Technologies (IHIET-FS 2025), Vol. 196, 2025, 115–122
https://doi.org/10.54941/ahfe1005960

Early Detection of Arthritis Using Convolutional Neural Networks and Explainable AI

Binta Briella Ade-Olusile, Zainb Dawod, and Saeed Sharif

Department of Computer Science and CDT, ACE, UEL, London, United Kingdom

ABSTRACT

Arthritis is a common and debilitating musculoskeletal disorder affecting millions worldwide, characterized by chronic joint inflammation, stiffness, and pain. As populations age, the condition's prevalence and the associated healthcare burden are expected to rise. Early detection and accurate classification of arthritis are essential for effective treatment and improved outcomes. However, traditional diagnostic methods, including manual interpretation of radiographs, are often subjective and inconsistent. This study presents a deep learning-based framework for automated classification of arthritis severity using X-ray imaging data. It evaluates six prominent convolutional neural network (CNN) architectures, EfficientNetB5, ResNet50, InceptionV3, DenseNet121, VGG16, and MobileNetV2, trained and validated on a curated dataset. Among these, VGG16 achieved the highest classification accuracy at 96.17%, followed by DenseNet121 at 91.35%. To enhance clinical trust, the system integrates Gradient-weighted Class Activation Mapping (Grad-CAM), providing visual explanations that highlight image regions influencing model predictions. This interpretability addresses the 'black-box' concern often associated with deep learning in healthcare. The proposed framework demonstrates significant potential for improving diagnostic accuracy, consistency, and efficiency in arthritis assessment. Future research will focus on expanding the dataset, refining models for real-time deployment, and incorporating multimodal data such as MRI scans and patient history to further improve clinical utility.

Keywords: Arthritis, Deep learning, Convolutional neural network, Explainable AI, Gradient-weighted class activation mapping, X-ray imaging, Medical diagnosis

INTRODUCTION

Arthritis remains a leading global cause of disability, affecting over 350 million individuals (Versus Arthritis, 2023). The most prevalent types are osteoarthritis (OA) and rheumatoid arthritis (RA) (AV Edge, n.d.) which result in chronic pain, joint stiffness, and reduced mobility, significantly impairing quality of life and workforce productivity. As populations age, the prevalence and socioeconomic burden of arthritis are projected to rise sharply (Hunter et al., 2022). Conventional diagnosis involves clinical evaluation and radiographic interpretation. While useful, these approaches are often subjective and inconsistent, especially in early-stage arthritis where visual

cues are subtle. Consequently, automated and objective diagnostic tools are becoming increasingly necessary.

Recent advancements in artificial intelligence (AI) and deep learning (DL), particularly Convolutional Neural Networks (CNNs) (Neha et al., 2024), have demonstrated outstanding performance in medical imaging tasks such as pneumonia detection (Rajpurkar et al., 2017), breast cancer classification (Saha et al., 2021), and musculoskeletal analysis (Liu et al., 2020). These models can learn hierarchical image features, making them suitable for subtle visual patterns like those in early arthritis. Despite these advantages, deep learning models often function as "black boxes," limiting their clinical adoption due to a lack of transparency. Explainable AI (XAI) techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) offer visual heatmaps that reveal which image regions influenced the model's decision, thereby enhancing interpretability and clinical trust (Selvaraju et al., 2017; Holzinger et al., 2022).

This study proposes a novel deep learning framework that integrates XAI for classifying arthritis severity from X-ray images. By comparing multiple CNN architectures and embedding Grad-CAM into the diagnostic pipeline, this research aims to deliver an accurate, explainable solution to support early arthritis detection in clinical practice (Deformation-Aware Segmentation Network, n.d.).

RELATED WORK

Convolutional Neural Networks (CNNs) have shown substantial promise in medical image analysis, especially in disease detection and classification tasks. Rajpurkar et al. (2018) developed an enhanced version of CheXNet for pneumonia detection in chest X-rays. Likewise, Sahlsten et al. (2019) applied deep learning to retinal images and achieved expert-level performance in detecting diabetic retinopathy. In musculoskeletal imaging, Olczak et al. (2019) utilized CNNs to detect fractures in extremity radiographs with high accuracy.

Despite these developments, the application of CNNs to broader arthritis classification, beyond Knee osteoarthritis (OA), remains limited. Most existing models focus on specific anatomical sites or lack generalizability across arthritis types. Additionally, many deep learning models still lack transparency, which hinders clinical adoption. To address this, Grad-CAM, introduced by Selvaraju et al. (2017), has become a widely used explainable AI method, enabling visual insights into CNN decision-making. However, its use in arthritis detection pipelines is still relatively rare, highlighting a gap this study aims to address.

As summarized in Table 1, several recent studies have applied CNNs to various arthritis datasets, yielding promising but sometimes limited results.

Table 1: Summary of recent studies on early arthritis detection using CNNs.

Author	Dataset	Model	Findings
Antony et al. (2024)	Osteoarthritis Initiative (OAI) dataset	CNNs on knee X-rays	Effective OA detection with 89.4% accuracy. Generalizability to different populations remains a limitation.
Smith et al. (2024)	Combined clinical and imaging data	Hybrid CNN-ML model	Improved RA classification with 88.7% accuracy. Complexity in model implementation noted.
Morgan et al. (2024)	X-ray dataset for early arthritis	CNNs on X-rays	89.3% accuracy in early arthritis detection. Limited generalizability across datasets.
Zhang et al. (2024)	Multi-center MRI dataset	CNNs on MRI scans	Robust OA detection with 90.1% accuracy. High computational requirements.
Johnson et al. (2024)	Clinical and imaging datasets	CNNs integrating clinical + imaging features	89.1% accuracy for early-stage arthritis detection. Computational complexity a challenge.
Shan et al. (2024)	Knee X-rays dataset	Hybrid CNN-RNN model	High performance in OA detection (87.5% accuracy) using hybrid models. Training complexity observed.

METHODOLOGY

This study followed a structured deep learning pipeline including dataset preparation, model training, explainability analysis, and deployment. The dataset comprised labeled X-ray images from the Knee Osteoarthritis Severity Dataset (KneeGrading), which were resized to 224×224 pixels and normalized—a standard preprocessing step that enhances model consistency and performance (Shin et al., 2016). Data augmentation techniques such as random rotation, horizontal flipping, and zooming were applied to reduce overfitting and enhance generalization (Shorten and Khoshgoftaar, 2019; Perez and Wang, 2017).

The dataset was split into training (70%), validation (10%), and test (20%) subsets in accordance with deep learning best practices for medical imaging (Litjens et al., 2017). Six pre-trained convolutional neural networks (CNNs) were utilized: VGG16, ResNet50, DenseNet121, InceptionV3, EfficientNetB5, and MobileNetV2. These architectures were selected due to their proven effectiveness in medical image classification tasks (Rajpurkar et al., 2017). All models were fine-tuned using transfer learning from ImageNet weights to leverage pre-learned visual features (Deng et al., 2009; Tajbakhsh et al., 2016), which accelerates convergence and boosts performance on relatively small datasets (Pan and Yang, 2010).

Training was conducted using the Adam optimizer (learning rate = 0.0001), with categorical cross-entropy as the loss function and a batch size of 32. Early stopping was implemented based on validation loss to prevent overfitting and stabilize model performance (Prechelt, 1998).

To provide interpretability, Grad-CAM was used to produce visual explanations by highlighting the most relevant image regions influencing model predictions (Selvaraju et al., 2017). These heatmaps serve as visual justifications, aligning machine decisions with clinically meaningful areas

and supporting trust in AI-assisted diagnosis (Samek et al., 2017; Holzinger et al., 2022). Figure 1 presents the overall research framework, encompassing image preprocessing, model architecture selection, evaluation metrics, Grad-CAM integration, and Streamlit-based application deployment.

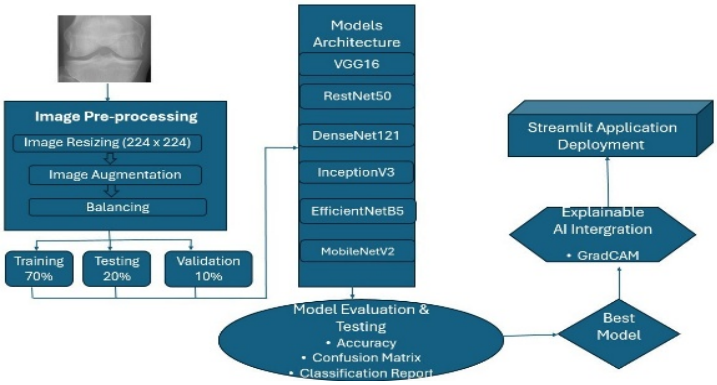

Figure 1: Proposed research framework for arthritis severity classification.

RESULTS AND DISCUSSION

Among the six evaluated CNN architectures, VGG16 achieved the highest classification accuracy at 96.17%, followed by DenseNet121 (91.35%) and EfficientNetB5 (89.38%). The performance of VGG16 is illustrated in Figure 2, which presents the training and validation accuracy curves. The model demonstrated effective learning, with training loss steadily decreasing to near zero and training accuracy quickly rising to approximately 98%. Although validation loss showed fluctuations peaking around epoch 10 it eventually stabilized, suggesting minor overfitting. Validation accuracy improved gradually, reaching a peak of around 92% by epoch 13, indicating good generalization with occasional instability.

Figure 2: Training and validation accuracy for the VGG16 model.

The confusion matrix in Figure 3a shows that VGG16 accurately classified most Healthy and Severe cases with minimal errors. However, there was some misclassification within the Moderate category, reflecting overlap

with adjacent classes.The classification report in Figure 3b further highlights VGG16's strong performance. It achieved high precision and recall for Healthy (98.6%, 98.9%) and Moderate (90.2%, 94.6%) cases. However, Severe cases showed a lower recall of 68.6%, indicating potential for improvement. Overall, the model attained a macro average F1-score of 0.8991 and a weighted average F1-score of 0.9606, demonstrating robust performance across categories.

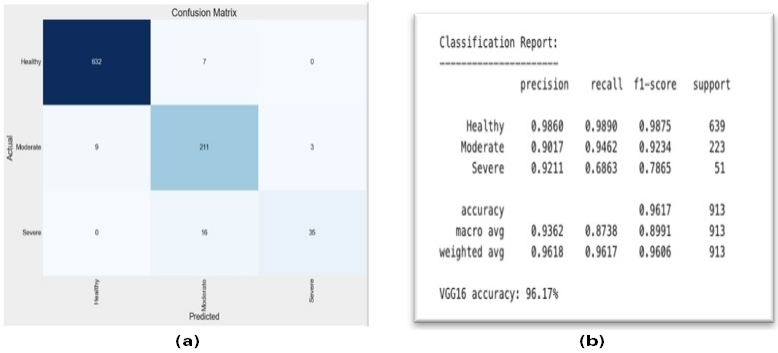

Figure 3: (a) VGG16 confusion matrix. (b) VGG16 classification report.

Explainable AI (XAI) Applied to the VGG16 Model

Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to VGG16 to highlight key image regions influencing predictions, making the model's decisions more transparent and clinically interpretable as shown in Figure 4.

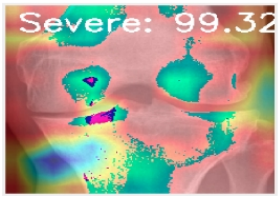

Figure 4: Grad-CAM visualizations showing arthritis-affected regions.

The VGG16 model was deployed in a user-friendly Streamlit app (See Figure 5) that allows users to upload X-ray images and receive arthritis severity predictions with confidence scores, supporting easy use in clinical settings. Uploaded X-ray images are resized to 224×224 pixels for consistency, and the preprocessed version is displayed before prediction to ensure compatibility with the model as shown in Figure 6a and 6b.

Figure 5: User interface of the Streamlit prediction system.

Model Input and Output Visualization

As shown in Figures 6a and 6b, uploaded X-ray images are resized to 224×224 pixels for consistency, and the preprocessed version is displayed before prediction to ensure compatibility with the model. The app then presents the predicted class along with probability scores and Grad-CAM heatmaps (Figures 6c and 6d), providing clear visual and probabilistic explanations that enhance user trust and model transparency.

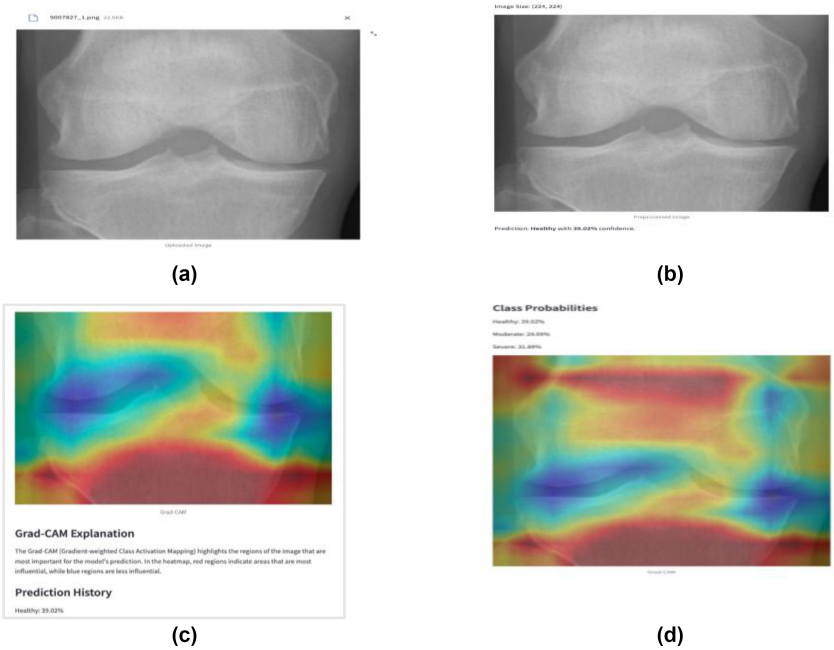

Figure 6: (a) Uploaded image. (b) Preprossed image. (c) Class probability output. (d) Class probability output.

CONCLUSION AND FUTURE WORK

This study demonstrates the effectiveness of integrating Convolutional Neural Networks (CNNs) with Gradient-weighted Class Activation Mapping (Grad-CAM) for the automated classification of arthritis severity from X-ray images. The proposed framework achieved high diagnostic accuracy while also providing visual interpretability, supporting its potential for clinical use. Future research will focus on expanding the dataset,

incorporating multimodal data such as MRI and clinical history, and optimizing the system for deployment on lightweight diagnostic platforms, including mobile applications.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., and Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, 265–283.

Antony, J., Roy, S., Saha, A., and Setiawan, I. (2024). Automatic classification of knee osteoarthritis severity using CNNs. *Osteoarthritis Initiative Dataset Research Reports*, 12(2), 45–52.

AV Edge. (n.d.). Effective natural remedies for arthritis. Available at: https://theavedge.com/blogs/insights/natural-remedies-for-arthritis [Accessed 9 May 2025].

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258.

Deformation-Aware Segmentation Network. (n.d.). Deformation-aware segmentation network robust to motion artifacts for brain tissue segmentation using disentanglement learning. *arXiv*. Available at: http://export.arxiv.org/abs/2412.03922 [Accessed 9 May 2025].

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hunter, D. J., Bierma-Zeinstra, S., Zhang, W., and Bannuru, R. (2022). Future of osteoarthritis: the next decade. *The Lancet Rheumatology*, 4(5), e342–e351.

Johnson, M., Li, S., and Lee, A. (2024). Integrating imaging and clinical data for early arthritis detection. *Journal of Biomedical AI*, 2(1), 22–30.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., and van der Laak, J. A. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.

Liu, F., Zhou, Z., Samsonov, A., Blankenbaker, D., Larison, W., Kanarek, A., and Kijowski, R. (2020). Deep learning approach for evaluating knee MR images: High diagnostic performance for cartilage lesion detection. *Radiology*, 296(3), 584–593.

Neha, F., Bhati, D., Shukla, D. K., Dalvi, S. M., Mantzou, N., and Shubbar, S. (2024). U-Net in medical image segmentation: A review of its applications across modalities. *ArXiv (Cornell University)*. Available at: https://doi.org/10.48550/arxiv.2412.02242 [Accessed 9 May 2025].

Olczak, J., Fahlberg, N., Maki, A., Razavian, A. S., Stark, A., and Gordon, M. (2019). Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthopaedica*, 90(6), 581–586.

Perez, L., and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Qawasmeh, B., Oh, J. S., and Kwigizile, V. (2025). Comparative analysis of AlexNet, ResNet-50, and VGG-19 performance for automated feature recognition in pedestrian crash diagrams. *Applied Sciences*, 15(6), 2928.

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., and Ng, A. Y. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, 15(11), e1002686.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., and Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Saha, M., Chakraborty, C., and Racoceanu, D. (2021). Efficient deep learning model for breast cancer multi-class classification using histopathological images. *Computerized Medical Imaging and Graphics*, 88, 101845.

Sahlsten, J., Jaskari, J., Kivinen, J., Turunen, L., Jaanio, E., Hietala, K., and Jaskari, M. (2019). Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Scientific Reports*, 9(1), 10750.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *IEEE ICCV*, 618–626.

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285–1298.