Automated Knee Arthritis Detection Using X-Ray Imaging and Explainable AI

Sujal T S, Satish G T, Sudeep Patil, Rohit S, and Dr.Uday Kulkarni

KLE Technological University, Hubballi, Karnataka, India
{01fe22bci051, 01fe22bci053, 01fe22bci055, 01fe23bci407,
uday_kulkarni}@kletech.ac.in
https://www.kletech.ac.in/

Abstract

Osteoarthritis (OA) is a degenerative joint disease that significantly reduces mobility and quality of life, especially in aging populations. Early and reliable detection is crucial for effective clinical intervention, yet existing diagnostic approaches rely heavily on subjective assessment of radiographic images. In this work, we propose a deep learning-based framework for automated OA detection and grading using knee X-ray images, with a focus on explainability and deployability. Leveraging the publicly available osteoarthritis Initiative (OAI) dataset, we train an Xception-based convolutional neural network to classify images into Kellgren-Lawrence (KL) grades 0–4, achieving a 71% classification accuracy on unseen test data. To enhance clinical interpretability, we integrate Grad-CAM to visualize class-discriminative regions influencing model predictions. The trained model is then optimized and converted to TensorFlow Lite, enabling efficient real-time inference on mobile devices. This enables OA screening in remote or resource-limited settings, where access to radiological expertise is minimal. Our pipeline provides a practical, explainable, and portable solution, advancing the role of edge AI in medical diagnostics.

1 Introduction

Detecting osteoarthritis (OA) from knee X-ray images requires careful identification of subtle anatomical changes such as joint space narrowing, osteophyte development, and subchondral bone alterations. The objective of automated diagnostic systems is not only to classify disease severity with high accuracy but also to ensure that the underlying decision process aligns with clinical reasoning [1],[8]. Figure 1 illustrates the ideal outcome of such a system—predictions that are both quantitatively reliable and supported by interpretable visual cues that highlight relevant pathological regions [6],[10]. In practice, achieving this balance is essential to support clinical adoption, where medical professionals require transparency and trust in model outputs [11]. Automated OA detection plays a crucial role in numerous applications, including early screening, continuous monitoring, and mobile diagnostics in under-resourced settings [12]. However, ensuring consistency across varying image qualities and gaining clinical confidence in AI-driven decisions remain significant challenges, particularly when explainability is absent or insufficiently integrated [13],[14].

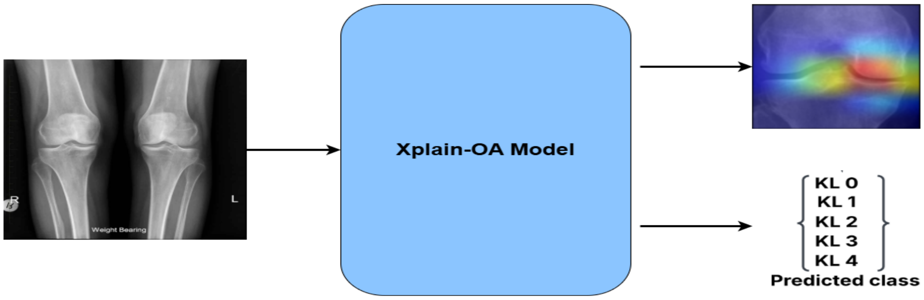


Figure 1. *Output from our model*

To address the challenges in automated osteoarthritis detection, recent approaches leverage deep convolutional neural networks (CNNs) to classify knee X-ray images by disease severity. These models are typically trained on large annotated datasets to capture complex radiographic features relevant to diagnosis [3] ,[15]. However, their performance often degrades in real-world settings due to variations in image quality, patient demographics, and scanner-specific artifacts [1]. While state-of-the-art architectures perform well under controlled conditions, they often lack the transparency needed for clinical trust [6]. Moreover, much of the existing research emphasizes accuracy, with limited focus on interpretability or deployment [11]. A significant gap remains in developing lightweight, explainable models that generalize effectively across diverse clinical environments, especially where expert access and computational resources are limited [16].

We propose a comprehensive, human-centered framework for automated osteoarthritis (OA) detection and grading from knee X-rays to overcome these limitations in conventional deep learning for medical imaging. The design prioritizes diagnostic accuracy, interpretability, and deployability. By integrating Gradient-weighted Class Activation Mapping (Grad-CAM), the model highlights class-discriminative regions within radiographs, offering clinicians meaningful visual explanations that support decision validation and build trust. Trained on the publicly available Osteoarthritis Initiative (OAI) dataset, the model is tailored to capture structural variations relevant to OA. It is further optimized using quantization and converted into a lightweight TensorFlow Lite format to enable efficient real-time inference on mobile devices. The framework combines transparent decision-making, diagnostic insight, and computational efficiency to support scalable deployment in real-world clinical settings, including low-resource environments.

Overall, the main contributions of our work include the development of a deep learning framework for automated detection and grading of knee osteoarthritis using X-ray images, with the trained model made publicly available. Our model achieves a classification accuracy of 71% across Kellgren-Lawrence (KL) grades 0–4 on unseen test data. To enhance clinical interpretability, we integrate visual explanation techniques such as Grad-CAM, allowing medical professionals to better understand and trust the model's decisions. Additionally, we optimize and quantize the model for real-time inference, enabling efficient deployment on mobile devices for use in resource-constrained settings.

Both the trained model and the mobile application are publicly available. Section 2 reviews related work and highlights existing research gaps. Section 3 presents the proposed methodology, including model architecture, training process, interpretability integration, and deployment strategy. Section 4 provides experimental results and evaluation metrics. Section 5 discusses the conclusion and future work.

2 Related Work

Deep learning, especially convolutional neural networks (CNNs), has been widely used for automated osteoarthritis (OA) detection from knee X-rays. Early studies applied architectures like AlexNet and VGGNet for OA severity classification [1],[2]. Later works improved accuracy using deeper models such as ResNet, DenseNet, and Xception, often leveraging transfer learning [3],[4]. While these models capture important spatial features, they typically function as "black boxes," limiting interpretability. Some research has explored multi-task learning to predict OA grades alongside clinical scores [5]. However, most CNN-based methods lack integrated interpretability and are not designed for real-time deployment, making them less suitable for use in clinical settings where responsiveness and clarity are critical.

Gradient-weighted Class Activation Mapping (Grad-CAM) [6] is commonly used to visualize image regions influencing CNN decisions. In OA imaging, Grad-CAM highlights clinically relevant areas like joint space narrowing and osteophytes, improving transparency and clinician trust [7]. Unlike previous work that uses Grad-CAM for post-hoc visualization, our framework integrates it directly into the diagnostic process to provide interpretable outputs alongside predictions. This approach enables real-time decision support that is not only accurate but also explainable, facilitating greater acceptance in healthcare workflows.

Mobile deployment of OA detection models is underexplored. Model compression and quantization have enabled real-time inference on limited devices in other medical fields, such as diabetic retinopathy [12], tuberculosis [13], and

[1] https://huggingface.co/Sujalts/Xplain-OA

skin lesion classification [14]. TensorFlow Lite quantization techniques [16] reduce model size and latency, making edge deployment feasible. Despite these advancements, similar efforts in OA detection remain limited, highlighting the need for lightweight, interpretable models tailored for portable medical applications. However, these advances have not been fully applied to OA diagnosis. Our work compresses an Xception-based model by 71%, achieving a 3.6x speed-up with minimal accuracy loss, enabling practical use in low-resource and point-of-care environments.

2.1 Xception-net

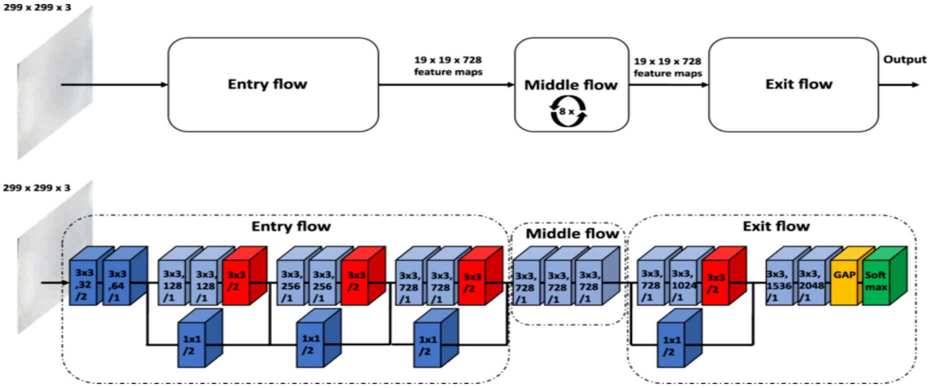


Figure 2. *Architecture of Xception-net*

The Xception network is a deep convolutional neural network architecture based on depthwise separable convolutions, proposed by François Chollet [17]. It extends the Inception architecture by replacing Inception modules with depthwise separable convolution layers.

The network has three parts: entry flow, middle flow, and exit flow. The entry flow extracts low-level features; the middle flow applies repeated depthwise separable convolution blocks that separate spatial filtering from channel-wise feature extraction, reducing parameters while maintaining power. The exit flow refines features for classification. Depthwise separable convolutions split the convolution into a depthwise convolution (one filter per input channel) followed by a pointwise (1x1) convolution combining channels, cutting computational complexity and size.

Xception efficiently learns rich spatial and channel-wise features, outperforming Inception-v3 with fewer parameters, making it well-suited for transfer learning in medical imaging tasks like osteoarthritis detection.

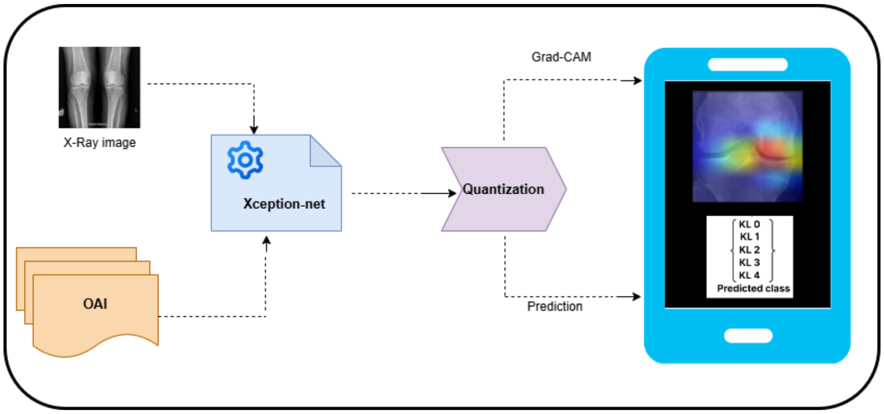


Figure 3. Comprehensive Architecture of the Proposed System

3 Proposed Work

This section details the overall methodology of our osteoarthritis detection framework. We begin by describing the dataset and preprocessing techniques applied to enhance image quality and variability for robust model training, We outline the design and training of a deep convolutional neural network for automated OA classification and grading 3.1. We then explain the integration of visual explanation methods to improve model interpretability and clinical trust 3.2. Finally, we describe the model optimization and deployment process, including quantization and conversion for real-time inference on mobile devices3.3, enabling practical use in low-resource settings

3.1 Fine tuning Xception net

This section details the fine-tuning pipeline used to improve knee osteoarthritis detection from X-ray images. We utilize 9,786 radiographs from the Osteoarthritis Initiative (OAI), each labeled with Kellgren-Lawrence (KL) grades. To boost model robustness and generalization, various data augmentation techniques are applied during training, including brightness adjustment, width shifts, zooming, and horizontal flipping, with pixel fill strategies preserving image integrity. All images are normalized following ResNet50 standards to align with the pretrained model's input distribution, facilitating effective transfer learning and faster convergence (see Algorithm 1 for preprocessing details).

The model was fine-tuned for 100 epochs with a batch size of 256, using the Adam optimizer and categorical cross-entropy loss for multi-class classification. Training batches were augmented, while validation data remained unchanged to ensure unbiased evaluation. EarlyStopping and ReduceLROnPlateau callbacks prevented overfitting and adapted learning rates dynamically. To address class imbalance, class weights were applied, emphasizing underrepresented KL grades.

The final model was selected based on the lowest validation loss, balancing accuracy and generalization for reliable osteoarthritis grading.

Algorithm 1: Training Loop for Fine-Tuning Knee Osteoarthritis Detection

```
Input Dataset $\mathcal{D}$, pretrained model $M$, augmentations $\mathcal{A}$, class weights $w_c$,
  epochs $E$, batch size $B$ Output Fine-tuned model $M^*$
for $e \leftarrow 1$ to $E$ do
    foreach batch $(I, y)$ in training set do
        Preprocess $I$: normalize and apply augmentations $\mathcal{A}$;
        Compute prediction: $\hat{y} \leftarrow M(I)$;
        Compute weighted loss: $\mathcal{L} \leftarrow \text{CrossEntropy}(\hat{y}, y) \times w_y$;
        Backpropagate loss: $\nabla\mathcal{L}$;
        Update model weights via Adam optimizer;
    end
    Evaluate $M$ on validation set; record validation loss;
    Adjust learning rate if plateau detected;
    Apply early stopping if no improvement;
end
return $M^*$ (best model by validation loss)
```

3.2 GradCAM visualization

To improve the interpretability of our model's predictions, we integrated Gradient-weighted Class Activation Mapping (Grad-CAM) into the diagnostic workflow. Grad-CAM highlights the regions in an input X-ray image that most influence the model's decision, making the process more transparent and clinically trustworthy.

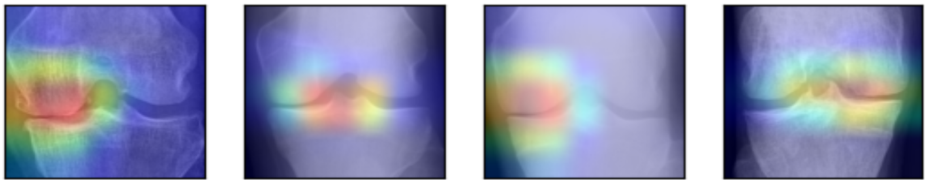


Figure 4. *Grad-Cam demonstration*

Our implementation uses a modified model that outputs both the final convolutional features and class predictions. Using TensorFlow's GradientTape, we compute the gradients of the predicted class score with respect to the feature maps. These gradients are then averaged to derive importance weights, which are combined with the feature maps to generate a class-discriminative heatmap(see Algorithm 2 for the detailed Grad-CAM computation steps).

Algorithm 2: Algorithm for Generating and Superimposing Grad-CAM Heatmap

```
Input Trained Grad-CAM model $M$, input image $I \in \mathbb{R}^{H \times W \times 3}$, optional class
  index $c$, heatmap transparency $\alpha \in [0, 1]$ Output Image $I_{cam}$ with
  Grad-CAM heatmap superimposed
begin
    Preprocess input image:
        $I_{\text{tensor}} = \text{Preprocess}(I)$
    Compute Grad-CAM heatmap:
        $H = \text{ReLU}\left(\sum_k \overline{g}_k \cdot A^k\right) \quad \text{where} \quad \overline{g}_k = \frac{1}{Z}\sum_{i,j}\frac{\partial y^c}{\partial A^k_{i,j}}$
     Here, $A^k$ is the activation map of the $k$-th channel from the last
     convolutional layer, and $y^c$ is the model output score for class $c$.
    Normalize heatmap:
        $\hat{H} = \frac{H}{\max(H)}$
    Convert $\hat{H}$ to a color map $C$ via a colormap function, e.g., Jet:
        $C = \text{Colormap}(\hat{H})$
    Resize $C$ to match input image size $(H \times W)$.
    Superimpose heatmap on original image:
        $I_{cam} = \alpha \times C + (1 - \alpha) \times I$
    return $I_{cam}$
end
```

The heatmap is normalized and overlaid on the original X-ray using a colormap to produce a visual explanation.This allows clinicians to see whether the model focuses on medically relevant regions like joint spaces and bone margins when grading osteoarthritis severity. By aligning AI predictions with clinical reasoning, Grad-CAM supports greater trust and adoption in real-world settings. Figure 4 illustrates sample Grad-CAM outputs from web web-based app, confirming that the model attends to pathologically meaningful regions across different OA grades.

3.3 Quantization & Mobile deployment

To enable real-time, resource-efficient deployment on mobile devices, we applied post-training quantization using TensorFlow Lite. This reduces model weights and activations from 32-bit floating point (FP32) to 8-bit integers (INT8), lowering memory usage and computation without significantly impacting accuracy.

The quantization converts each full-precision tensor $T \in \mathbb{R}^n$ into an 8-bit integer tensor Q, using a scale factor and zero-point offset to map floating-point values to integers, as shown in Equation (1). This enables efficient storage and faster arithmetic on resource-constrained devices.

$$Q = \text{clip}\left(\text{round}\left(\frac{T}{s}\right) + z,\ q_{\min},\ q_{\max}\right) \tag{1}$$

where

$$s = \frac{T_{\max} - T_{\min}}{q_{\max} - q_{\min}}, \quad z = \text{round}\left(q_{\min} - \frac{T_{\min}}{s}\right) \tag{2}$$

The original model, approximately 49 MB with over 23 million parameters, was compressed to 11.5 MB after quantization, a 75% reduction in size. This makes the model suitable for devices with limited storage.

The size reduction from FP32 to INT8 is computed in Equation (3):

$$R = \frac{\text{Size}(M_{\text{FP32}}) - \text{Size}(M_{\text{INT8}})}{\text{Size}(M_{\text{FP32}})} \times 100\% = \left(1 - \frac{8}{32}\right) \times 100\% = 75\% \tag{3}$$

We deployed the Xplain-OA model on Android by integrating it within a Flutter framework, enabling cross-platform compatibility and a smooth user interface. The model was first converted into a TensorFlow Lite format suitable for mobile devices, then embedded into the Flutter app using Flutter's TensorFlow Lite plugin. This approach allowed efficient on-device inference, providing fast and accurate osteoarthritis detection while maintaining a lightweight and responsive application for Android users. The image of the app interface is shown in Figure 6.

4 Results and Evaluation

In this section, we analyze the performance of our Xception-based deep learning model for automated knee osteoarthritis (OA) grading. The model was evaluated on a diverse test set of 1,656 X-ray images, achieving an overall accuracy of 0.70 and a balanced accuracy of 0.74, indicating consistent performance across imbalanced classes.

The detailed classification metrics, including precision, recall, and F1-score for each Kellgren-Lawrence (KL) grade are summarized in Table 1. The model demonstrates strong predictive capabilities for advanced OA stages (grades 3 and 4), with F1-scores of 0.72 for both grades, reflecting effective detection of pronounced radiographic changes. However, the performance on early OA grades, particularly grade 1, is relatively lower (F1-score 0.37), highlighting the inherent challenges of identifying subtle early-stage disease patterns that often exhibit less distinct radiographic features.

Table 1. Classification Report for OA Grading

Grade	Precision	Recall	F1-score	Support
0	0.82	0.60	0.69	639
1	0.31	0.46	0.37	296
2	0.65	0.66	0.66	447
3	0.73	0.71	0.72	223
4	0.59	0.92	0.72	51

Our model demonstrates competitive, and in some cases improved, accuracy and balanced accuracy across multiple OA severity grades compared to previous approaches, confirming our framework's robustness and clinical relevance.

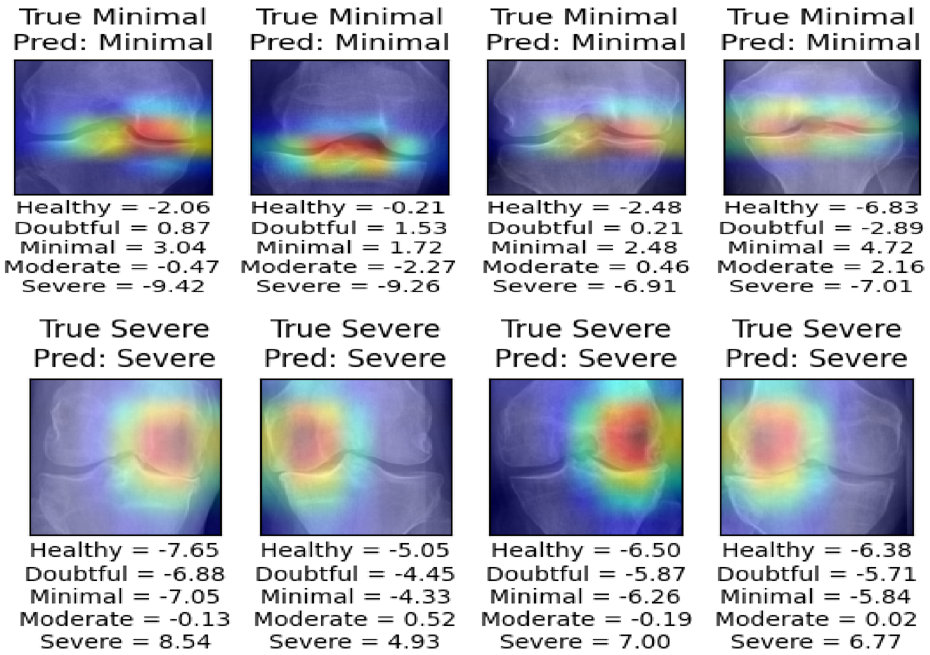


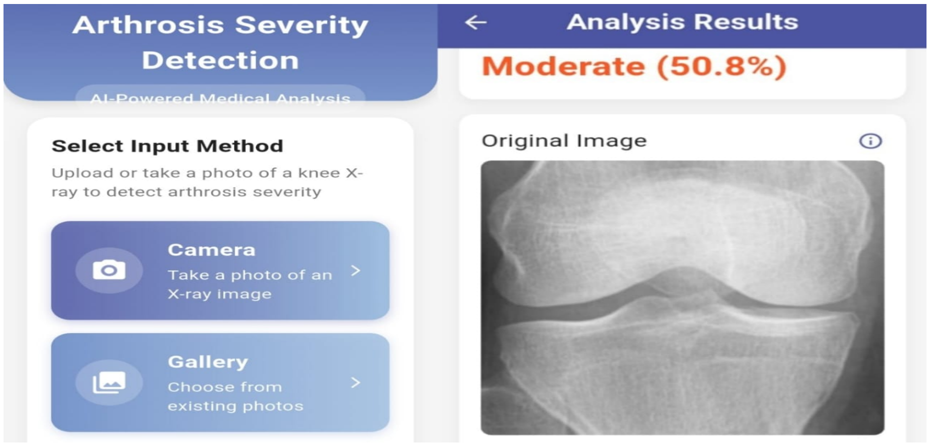
Figure 5. *Grad-Cam Results*

Moreover, the Grad-CAM visualizations presented in Fig 5 provide crucial interpretability by generating class-discriminative heatmaps. These heatmaps highlight clinically significant regions such as joint spaces and bone margins, which correspond closely to the anatomical areas assessed by radiologists for OA grading. This interpretability not only enhances transparency in the model's decision-making but also facilitates clinical validation and acceptance, fostering greater trust in AI-assisted diagnostics.

Table 2. Comparison of model parameters, size, and inference time before and after quantization

Model	**Model Size (MB)**	**Inference Time (ms)**
Xception (full)	48	120
Xception (quantized)	11.5	84

In addition to predictive performance, efficiency is vital for real-world deployment. Quantization significantly optimizes the model, reducing memory footprint and computational demands without notable loss in accuracy. Table 3 compares the model's size, FLOPs, and inference time before and after quantization. Quantization reduces the model size by approximately 75%, lowers FLOPs, and accelerates inference time by 30%, enabling real-time OA grading on mobile and edge devices.



Select Input Method

Upload or take a photo of a knee X-ray to detect arthrosis severity

Camera: Take a photo of an X-ray image

Gallery: Choose from existing photos

Analysis Results

Moderate (50.8%)

Arthrosis Severity Detection
AI-Powered Medical Analysis
Camera
Take a photo of an X-ray image
Gallery
Choose from existing photos
Original Image

Figure 6. *Mobile app interface for the Xplain-OA model*

These improvements not only reduce memory and computational requirements but also enhance inference speed, which is critical for deployment in resource-limited and point-of-care clinical environments. This efficiency, coupled with strong diagnostic accuracy and interpretability, makes our framework highly practical and scalable for widespread adoption in osteoarthritis screening, monitoring, and management.

5 Conclusion & Future Work

We propose a fine-tuned Xception-based deep learning framework for automated knee osteoarthritis grading from X-rays, enhanced with Grad-CAM for improved interpretability. The model is quantized and optimized for real-time inference on mobile devices, enabling deployment in resource-limited clinical settings. Experimental results show robust performance across OA severity grades, bridging AI and healthcare accessibility, especially for underserved populations. This framework offers an interpretable, scalable, and efficient tool for early OA diagnosis, advancing musculoskeletal imaging research.

Future work includes integrating additional imaging modalities and larger datasets, exploring advanced interpretability techniques for deeper clinical insights, and further optimizing the model for real-time OA diagnosis in remote settings..

References

1. Tiulpin, A., Thevenot, J., Rahtu, E., et al. (2018). Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. *Scientific Reports*, 8(1), 1727. https://doi.org/10.1038/s41598-018-19939-7
2. Antony, J., McGuinness, K., O'Connor, N. E., et al. (2016). Automatic Detection of Knee Osteoarthritis from Plain Radiographs: A Deep Learning-Based Approach. *In Proceedings of the International Conference on Image Analysis and Recognition*, 90–97. https://doi.org/10.1007/978-3-319-44470-9_11
3. Shamir, L., Ling, S. M., Scott, W. W., et al. (2020). Deep Learning for Automated Radiographic Knee Osteoarthritis Diagnosis. *Radiology: Artificial Intelligence*, 2(1), e190065. https://doi.org/10.1148/ryai.2019190065
4. Antony, J., McGuinness, K., O'Connor, N. E., et al. (2017). Automatic Classification of Osteoarthritis Severity in Knee Radiographs Using Deep Convolutional Neural Networks. *In Proceedings of the IEEE International Conference on Image Processing*, 2597–2601. https://doi.org/10.1109/ICIP.2017.8296767
5. Zhang, Y., Yang, Q., Xu, Z., et al. (2019). Multi-task Learning for Osteoarthritis Grading and Clinical Score Prediction. *Medical Image Analysis*, 58, 101547. https://doi.org/10.1016/j.media.2019.101547
6. Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74
7. Feng, X., Feng, L., Liu, T., et al. (2020). Grad-CAM-Based Visualization for Knee Osteoarthritis Diagnosis Using Deep Learning. *IEEE Access*, 8, 113760–113771. https://doi.org/10.1109/ACCESS.2020.3006710

8. Antony, J., McGuinness, K., Moran, K., & O'Connor, N. E. (2016). Quantifying Radiographic Knee Osteoarthritis Severity Using Deep Convolutional Neural Networks. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 1413–1417. https://doi.org/10.1109/ICIP.2016.7532641
9. Selvaraju, R. R., Cogswell, M., Das, A., et al. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128, 336–359. https://doi.org/10.1007/s11263-019-01228-7
10. Feng, X., Yang, Z., Zhao, X., et al. (2020). Deep Learning for the Classification of Osteoarthritis Severity on Knee X-rays Using Grad-CAM. *Medical Image Analysis*, 64, 101745. https://doi.org/10.1016/j.media.2020.101745
11. Guan, Q., Xu, Y., Feng, J., et al. (2021). Interpretable Deep Learning for Automated Knee Osteoarthritis Severity Classification Using Grad-CAM and Statistical Analysis. *Computers in Biology and Medicine*, 135, 104548. https://doi.org/10.1016/j.compbiomed.2021.104548
12. Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216
13. Lakhani, P., & Sundaram, B. (2017). Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, 284(2), 574–582. https://doi.org/10.1148/radiol.2017162326
14. Talo, M., Yildirim, Ö., Baloglu, U. B., et al. (2019). Application of Deep Transfer Learning for Automated Detection of COVID-19 Using Chest X-ray Images. *Applied Soft Computing*, 97, 106810. https://doi.org/10.1016/j.asoc.2020.106810
15. Antony, J., McGuinness, K., O'Connor, N. E., & Moran, K. (2017). Automated Detection of Knee Osteoarthritis from Plain Radiographs: A Deep Learning-Based Approach. *Journal of Medical Imaging*, 4(4), 041302. https://doi.org/10.1117/1.JMI.4.4.041302
16. Jacob, B., Kligys, S., Chen, B., et al. (2018). Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2704–2713. https://doi.org/10.1109/CVPR.2018.00286
17. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *ArXiv*, abs/1610.02357. https://arxiv.org/abs/1610.02357