International Journal of
Applied Data Science in Engineering and Health
https://ijadseh.com

Automated Knee Osteoarthritis Severity Grading in X-ray Images Using Transfer Learning

Mohammad Mojtaba Rohani

Department of Radiology, Poursina Hospital, Guilan University of Medical Sciences, Rasht, Iran

(Email: rohani.mmojtaba@gmail.com)

Received date: February 11, 2025; Accepted date: March 18, 2025

Abstract

Knee osteoarthritis (OA) is a chronic degenerative joint disease that leads to cartilage breakdown, joint space narrowing, and osteophyte formation, significantly impacting mobility and quality of life. Early and accurate grading of OA severity is crucial for effective treatment and disease management. X-ray imaging is the most widely used diagnostic tool for knee OA, providing a cost-effective and non-invasive method to assess structural changes in the knee joint. However, manual interpretation of X-rays is subjective and prone to inter-observer variability, necessitating the development of automated AI-based diagnostic systems. This study proposes a deep learning-based approach for knee OA severity grading using transfer learning with MobileNet and ResNet models. The results indicate that ResNet significantly outperformed MobileNet, achieving 83.5% accuracy, 83.0% precision, and 83.1% recall, compared to 74.1%, 71.5%, and 71.0%, respectively, for MobileNet. Statistical analysis using ANOVA confirmed significant differences ($p < 0.05$) in classification performance between the models.

Keywords*:* Knee Osteoarthritis; X-ray; MobileNet; ResNet.

Introduction

Knee osteoarthritis (OA) is a degenerative joint disease that affects millions of people worldwide. It is characterized by the gradual breakdown of articular cartilage, leading to joint pain, stiffness, and loss of mobility. As the disease progresses, it also causes changes in the underlying bone and surrounding soft tissues, further exacerbating symptoms [1]. Knee OA is considered the most common form of arthritis responsible for knee pain and disability, significantly impacting the quality of life of affected individuals. The condition is primarily associated with aging but can also be influenced by factors such as obesity, previous joint injuries, and genetic predisposition. Given its high prevalence and impact on daily life, accurate diagnosis and grading of knee OA are essential for effective disease management and treatment planning [2].

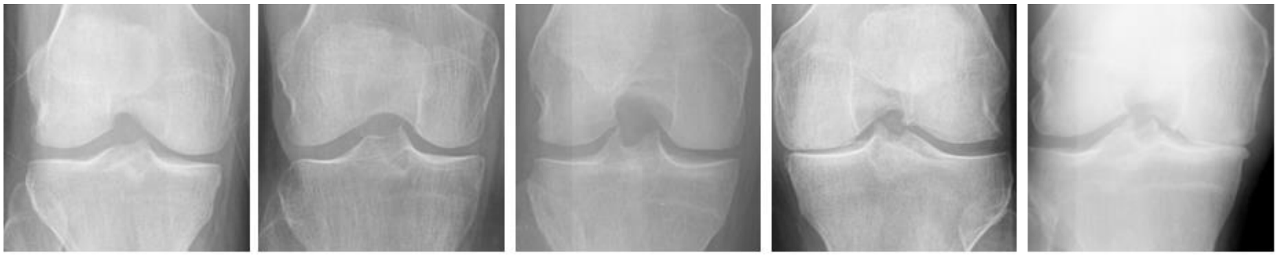

Figure 1 X-ray images of Knee Osteoarthritis Left to Right: Grade 0: Healthy knee image, Grade 1 (Doubtful): Doubtful joint narrowing with possible osteophytic lipping, Grade 2 (Minimal): Definite presence of osteophytes and possible joint space narrowing, Grade 3 (Moderate): Multiple osteophytes, definite joint space narrowing, with mild sclerosis, Grade 4 (Severe): Large osteophytes, significant joint narrowing, and severe sclerosis [3].

To assess the severity of knee OA, the Kellgren-Lawrence (KL) grading system is widely used in clinical practice [3]. This system classifies knee OA into five grades based on radiographic features observed in X-ray images. Grade 0 represents a normal, healthy knee with no signs of osteoarthritis. Grade 1 is characterized by doubtful joint space narrowing and possible osteophytic lipping. Grade 2 indicates the presence of definite osteophytes and possible joint space narrowing, while Grade 3 shows multiple osteophytes, definite joint space narrowing, and mild sclerosis. Finally, Grade 4 represents the most severe stage, with large osteophytes, significant joint space narrowing, and severe sclerosis [3]. Accurate classification of these grades is crucial for monitoring disease progression and determining appropriate interventions, including non-surgical and surgical treatments. However, manual grading is subjective, time-consuming, and prone to inter-observer variability, which can lead to inconsistencies in diagnosis.

X-ray imaging is a widely used and cost-effective modality for diagnosing various medical conditions [4-6]. It provides detailed structural information about bones and joints, making it a fundamental tool in evaluating knee OA. Despite its advantages, X-ray interpretation requires expert knowledge and experience, as radiologists must carefully examine subtle changes in joint structure, bone density, and osteophyte formation [7,8]. Inconsistent assessments between different radiologists and the labor-intensive nature of manual evaluation have led to a growing interest in automated diagnostic systems that can improve efficiency and accuracy.

In recent years, deep learning techniques have transformed medical image analysis by enabling automated feature extraction and classification [9]. Convolutional Neural Networks (CNNs) have shown remarkable success in identifying patterns in medical images, often achieving diagnostic performance comparable to, or even exceeding, human experts. In the context of knee OA diagnosis, deep learning models can learn to detect key radiographic features associated with different severity levels, providing a more objective and scalable approach [10-12]. Transfer learning, which involves fine-tuning pre-trained models on domain-specific data, has been particularly effective in medical imaging tasks, allowing for robust feature extraction even with relatively small datasets.

In this study, we propose an automated knee osteoarthritis severity grading system using transfer learning with MobileNet and ResNet models. Our approach leverages pre-trained deep learning architectures to classify knee X-ray images into different KL grades. By utilizing transfer learning, we aim to enhance classification performance while reducing training time and computational costs. We compare the performance of MobileNet and ResNet in terms of accuracy, precision, recall, and F1-score to determine the most effective model for this task. The findings of this study contribute to the growing field of AI-driven medicalKnee osteoarthritis (OA) is a degenerative joint disease that affects millions of people globally. It is characterized by the gradual breakdown of articular cartilage, resulting in joint pain, stiffness, and reduced mobility. As the disease advances, it also leads to changes in the underlying bone and surrounding soft tissues, further worsening symptoms. Knee OA is recognized as the most common type of arthritis responsible for knee pain and disability, significantly affecting the quality of life for those impacted. While the condition is primarily linked to aging, it can also be influenced by factors such as obesity, previous joint injuries, and genetic predisposition. Given its high prevalence and effect on daily life, accurate diagnosis and grading of knee OA are vital for effective disease management and treatment planning.

Methods

In this study, we propose an automated Knee Osteoarthritis (OA) severity grading system using transfer learning with MobileNet and ResNet models. Our approach involves multiple stages, including data preprocessing, model selection, training, and evaluation, to develop a deep learning-based system capable of classifying knee X-ray images into different grades based on the Kellgren-Lawrence (KL) grading system. By leveraging pre-trained convolutional neural networks (CNNs), we aim to enhance the efficiency and accuracy of knee OA classification while reducing the reliance on manual radiographic interpretation.

Dataset

The dataset used in this study consists of knee X-ray images labeled according to the Kellgren-Lawrence (KL) grading system, which is widely used for assessing the severity of knee osteoarthritis (OA) [3]. This dataset provides essential information for both knee joint detection and OA severity grading, making it suitable for deep learning-based classification tasks. The dataset includes X-ray images categorized into five different KL grades, each representing a distinct level of OA progression.

The grade descriptions in the dataset are as follows:

Grade 0 (Healthy): No radiographic evidence of osteoarthritis.
Grade 1 (Doubtful): Possible osteophytic lipping and minor joint space narrowing.
Grade 2 (Minimal): Definite presence of osteophytes and possible joint space narrowing.
Grade 3 (Moderate): Multiple osteophytes, definite joint space narrowing, and mild sclerosis.
Grade 4 (Severe): Large osteophytes, severe joint space narrowing, and subchondral sclerosis.

This dataset provides a structured and standardized representation of knee OA severity, allowing deep-learning models to learn the distinguishing radiographic features of each grade. The X-ray images vary in contrast, brightness, and anatomical positioning, introducing real-world challenges that make automated classification a valuable tool for clinical applications. To prepare the dataset for model training, preprocessing techniques such as image resizing, normalization, and augmentation were applied. These steps help enhance model generalization and improve performance by reducing overfitting. The dataset is split into training, validation, and test sets to ensure robust evaluation of the proposed models. The preprocessed images serve as input to the transfer learning models, facilitating efficient feature extraction and classification. By leveraging this dataset, our study aims to develop an automated deep learning-based system for knee OA severity grading, reducing the subjectivity and inconsistency associated with manual grading by radiologists. The structured nature of the dataset ensures that the proposed models can be trained effectively and evaluated comprehensively to assess their classification accuracy and clinical applicability.

Data Preprocessing

To prepare the dataset for deep learning model training, several preprocessing steps were applied to standardize and enhance the input images. First, all X-ray images were resized to 224×224 pixels, ensuring compatibility with the input dimensions required by MobileNet and ResNet models. Image normalization was performed by scaling pixel intensity values to the [0,1] range, which helps improve model convergence and training stability. To address the challenge of limited data and improve model generalization, data augmentation techniques such as random rotations, horizontal flipping, zooming, and contrast adjustments were applied. These augmentations introduced variability in the training data, reducing the risk of overfitting and enhancing model robustness. The dataset was then split into training (70%), validation (15%), and test (15%) sets to ensure a fair and robust evaluation of model performance.

Model Selection

We employed transfer learning, a technique that leverages pre-trained deep learning models, to classify knee osteoarthritis severity from X-ray images. Two well-established convolutional neural networks (CNNs), MobileNet and ResNet, were selected for this task due to their effectiveness in image classification tasks.

MobileNet

MobileNet is a lightweight deep learning architecture designed for efficient computation, making it particularly suitable for deployment on mobile and embedded devices. Unlike traditional CNNs, which use standard convolutions for feature extraction, MobileNet employs depthwise separable convolutions, significantly reducing the number of trainable parameters and computational cost. A depthwise separable convolution consists of two steps: Depthwise Convolution which is a single filter applied to each input channel separately, reducing the number of operations, and Pointwise Convolution which is a 1×1 convolution used to combine the outputs of the depthwise convolution, effectively learning spatial dependencies. This structure enables MobileNet to achieve high accuracy while being computationally efficient, making it ideal for medical image analysis where processing speed is a concern. MobileNet is known for its ability to retain key features despite its lightweight nature, making it a strong candidate for classifying knee OA severity.

ResNet (Residual Network)

ResNet (Residual Network) is a deep learning model designed to overcome the vanishing gradient problem, which occurs in very deep networks where gradients become extremely small, slowing down learning. ResNet introduces residual connections (skip connections), which allow information to bypass several layers, preserving important feature representations and making it easier to train deep networks. A residual block in ResNet consists of a shortcut connection that directly adds the input of a layer to its output before passing it to the next layer. This prevents the network from degrading in performance as depth increases, allowing the model to learn deeper, more complex features. ResNet is particularly effective for medical imaging tasks because it can capture fine-grained details in X-ray images, distinguishing subtle differences in joint space narrowing, osteophyte formation, and bone sclerosis. By using ResNet in knee OA classification, the model can learn high-level hierarchical features that differentiate between different KL grades, improving the overall classification performance.

Model Training and Fine-Tuning

The selected models, MobileNet and ResNet, were fine-tuned on the knee X-ray dataset to adapt their pre-trained weights to the specific task of knee osteoarthritis grading. The final fully connected layer of each model was replaced with a custom classification head, allowing for five-class classification corresponding to the KL grades. The training was performed using the Adam optimizer with an initial learning rate of 0.0001. The categorical cross-entropy loss function was used, as this is a multi-class classification problem. A batch size of 32 was selected to balance computational efficiency and training stability. The models were trained for 50 epochs, with early stopping applied to prevent overfitting if the validation loss did not improve for five consecutive epochs. To further improve classification accuracy, fine-tuning was applied by gradually unfreezing and training the deeper layers of MobileNet and ResNet. This process allowed the models to learn domain-specific features from knee X-ray images while retaining the general image representation knowledge from ImageNet.

Implementation Details

The models were implemented using the TensorFlow and Keras deep learning frameworks, running on a GPU-enabled system to accelerate training. Hyperparameter tuning was performed to optimize learning rate, batch size, and dropout rates, ensuring optimal model generalization. To enhance the reliability and robustness of our results, we conducted 10 independent training cycles, with each run consisting of 50 epochs. This approach allowed us to account for variations in model convergence and performance across different runs. The trained models were later evaluated on unseen X-ray images to assess their stability, consistency, and real-world applicability in clinical settings.

Results

The performance analysis of MobileNet and ResNet demonstrates that ResNet consistently outperforms MobileNet across all evaluation metrics, highlighting its superior ability to classify knee osteoarthritis severity based on X-ray images. In terms of accuracy, MobileNet achieved 74.1% (±0.0036), while ResNet demonstrated a significantly higher accuracy of 83.5% (±0.0058). This 9.4% improvement suggests that ResNet’s deeper architecture and residual connections enable it to capture more intricate patterns in knee joint structures, leading to better classification performance. The slightly higher standard deviation in ResNet’s accuracy indicates a small degree of variation across the 10 independent runs, but its performance remains consistently superior to MobileNet.

When analyzing precision, MobileNet attained 71.5% (±0.032), whereas ResNet achieved a notably higher 83.0% (±0.078). Precision reflects the model's ability to minimize false positives, and the substantial difference suggests that ResNet is more effective at correctly identifying osteoarthritis severity levels while avoiding misclassification errors. The larger standard deviation in ResNet’s precision score may indicate greater variability in its classification of certain KL grades, but overall, it remains more reliable and precise than MobileNet. Regarding sensitivity (recall), which measures the model’s ability to correctly identify true positive cases, MobileNet scored 71.0% (±0.0039), while ResNet achieved 83.1% (±0.0037). This indicates that ResNet is significantly better at correctly identifying knee osteoarthritis cases across different severity levels, reducing the likelihood of false negatives, which is critical in medical diagnosis. The nearly identical standard deviations for both models in sensitivity suggest stable and consistent recall performance across multiple runs.

Overall, these results indicate that ResNet is the superior model for knee osteoarthritis severity grading, offering higher accuracy, precision, and sensitivity. The improved performance can be attributed to its deep residual connections, which enable better feature extraction and prevent degradation in deeper networks. While MobileNet is a more lightweight model optimized for efficiency, its lower classification performance suggests that it may not be as effective for complex medical imaging tasks such as knee osteoarthritis diagnosis.

Table 1. Classification Performance of the Pre-trained Deep Learning Models

Model	MobileNet	ResNet
Accuracy	0.741±0.0036	0.835±0.0058
Precision	0.715±0.032	0.830±0.078
Sensitivity	0.710±0.0039	0.831±0.0037

Statistical Analysis

A statistical evaluation of differences in classification performance metrics (accuracy, precision, and recall) between the models can be conducted using appropriate statistical methods. One such approach is Analysis of Variance (ANOVA), which is particularly useful for comparing multiple groups simultaneously. ANOVA enables the assessment of whether significant differences exist in the mean performance of the models across various metrics. The p-values generated from ANOVA indicate whether these differences are statistically meaningful. If the p-value falls below a predefined significance threshold (e.g., 0.05), it confirms the presence of significant disparities between the models' classification performance.

To determine whether there were significant differences in model performance, an Analysis of Variance was conducted on accuracy, precision, and recall. The results indicated statistically significant differences across all three metrics. Specifically, accuracy showed a p-value of 4.36e-03, confirming meaningful variation between models. Similarly, precision had a p-value of 4.15e-03, reinforcing the presence of performance differences. The most pronounced distinction was observed in the recall, with a p-value of 8.22e-04, highlighting considerable variation in the model's ability to correctly identify true positive cases. These findings emphasize the critical role of model selection, as performance disparities can significantly impact classification effectiveness.

Conclusion

In this study, we developed an automated knee osteoarthritis severity grading system using transfer learning with MobileNet and ResNet models. The proposed approach leverages deep learning to classify knee X-ray images into Kellgren-Lawrence (KL) grades, offering an efficient and objective alternative to manual radiographic assessment. By employing pre-trained convolutional neural networks (CNNs), we aimed to improve diagnostic accuracy while reducing the subjectivity and variability associated with traditional grading methods. Our results demonstrate that ResNet significantly outperforms MobileNet across all evaluation metrics, including accuracy (83.5% vs. 74.1%), precision (83.0% vs. 71.5%), and recall (83.1% vs. 71.0%). The superior performance of ResNet can be attributed to its deep residual connections, which enhance feature extraction and prevent performance degradation in deeper networks. MobileNet, while more computationally efficient, exhibited lower classification performance, indicating that lightweight architectures may not be optimal for complex medical imaging tasks such as knee osteoarthritis grading.

While the proposed deep learning-based system demonstrates promising results in automated knee osteoarthritis severity grading, several limitations should be acknowledged. First, the study relies solely on X-ray images, which, despite being widely used in clinical practice, may not capture the full extent of osteoarthritis-related changes compared to multi-modal imaging techniques such as MRI or CT scans. Additionally, the dataset used for training, although well-structured, may not fully represent the diversity of real-world clinical scenarios, including variations in patient demographics, imaging quality, and disease progression. Another limitation is the potential misclassification of borderline cases, particularly between adjacent Kellgren-Lawrence (KL) grades, which could impact clinical decision-making. Future research should explore the integration of multi-modal imaging data to improve diagnostic accuracy, as well as the development of explainable AI (XAI) techniques to enhance the interpretability of model predictions for clinical adoption. Moreover, fine-tuning the models with larger, more diverse datasets and implementing self-supervised learning approaches could further enhance generalizability. Expanding this work to real-time clinical applications and assessing its effectiveness in prospective clinical trials would be valuable steps toward deploying AI-assisted knee osteoarthritis diagnosis in routine medical practice.

Conflict of Interest

The authors imply no conflict of interest.

References

[1] Sharma L. Osteoarthritis of the knee. New England Journal of Medicine. 2021 Jan 7;384(1):51-9.

[2] Hussain SM, Neilly DW, Baliga S, Patil S, Meek RM. Knee osteoarthritis: a review of management options. Scottish medical journal. 2016 Feb;61(1):7-16.

[3] Chen P. Knee osteoarthritis severity grading dataset. Mendeley Data. 2018 Jan;1(10.17632):30784984.

[4] Cozzi D, Albanesi M, Cavigli E, Moroni C, Bindi A, Luvarà S, Lucarini S, Busoni S, Mazzoni LN, Miele V. Chest X-ray in new Coronavirus Disease 2019 (COVID-19) infection: findings and correlation with clinical outcome. La radiologia medica. 2020 Aug;125:730-7.

[5] Rad AE, Rahim MS, Rehman A, Saba T. Digital dental X-ray database for caries screening. 3D Research. 2016 Jun 11;7(2):18.

[6] Sharifi S, Donyadadi A. Detection and Diagnosis of Congenital Heart Disease from Chest X-Rays with Deep Learning Models. International Journal of Applied Data Science in Engineering and Health. 2025 Jan 2;1(1):1-9.

[7] Saleem M, Farid MS, Saleem S, Khan MH. X-ray image analysis for automated knee osteoarthritis detection. Signal, Image and Video Processing. 2020 Sep;14(6):1079-87.

[8] Tariq T, Suhail Z, Nawaz Z. Knee osteoarthritis detection and classification using x-rays. IEEE Access. 2023 May 16;11:48292-303.
[9] Suganyadevi S, Seethalakshmi V, Balasamy K. A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval. 2022 Mar;11(1):19-38.
[10] Yeoh PS, Lai KW, Goh SL, Hasikin K, Hum YC, Tee YK, Dhanalakshmi S. Emergence of deep learning in knee osteoarthritis diagnosis. Computational intelligence and neuroscience. 2021;2021(1):4931437.
[11] Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Scientific reports. 2018 Jan 29;8(1):1727.
[12] Kijowski R, Fritz J, Deniz CM. Deep learning applications in osteoarthritis imaging. Skeletal radiology. 2023 Nov;52(11):2225-38.