Early Detection of Knee Arthritis Using AI: Enhancing Diagnostic Accuracy with Deep Learning

Introduction

Arthritis is among the most common conditions that usually result in chronic pain and loss of mobility in millions of people worldwide; the most susceptible part is knee arthritis. Treatment of arthritis, especially at early stages, requires the diagnosis to be precise (Reference). The X-ray image diagnosis of arthritis can be challenging since the differences among the stages of the disease are not easily marked. There is a need for an automated system to classify knee arthritis X-ray images into five stages: normal, doubtful, mild, moderate, and severe. Doctors might then identify conditions more rapidly and precisely using machine learning, which could imply better patient care.

Equally important is the limitation of manual diagnoses, which may be very slow and full of human errors. Automating this classification would make diagnoses more consistent and speedier, especially in areas with poor access to specialized doctors. The AI-based classification system will also recognize patients who require emergency medical care, thus ensuring that those with debilitating arthritis get quality and timely treatment. The approach hastens the diagnosis and ensures patients receive proper and timely treatment.

Artificial intelligence (AI) in medical imaging, particularly for musculoskeletal conditions like arthritis, has gained significant attention in recent years. Many works also target the implementation of machine learning algorithms, especially CNNs, for diagnosing and classifying arthritis from X-ray images. It was demonstrated by Tiulpin et al. 2018 that CNNs have great potential in quantifying the severity of osteoarthritis. The results obtained were close to those from trained radiologists. These results point to particular possibilities of AI improving diagnostic precision and facilitating workflow, possibly reducing human errors regarding the detection and classification of arthritis severity. However, specific challenges still exist to disseminating AI for arthritis diagnosis. First, there is an issue with the datasets these models have used for training. Most reviewed studies concluded that one needs to move toward more diverse and comprehensive datasets that would allow generalization across a wide range of populations and conditions by AI models. Although deep learning models are very effective, their "black box" nature has raised concerns about interpretability and trust in clinical decision-making. Ongoing research is directed toward improving the interpretability of these models so they will smoothly fit into clinical practice, providing accuracy with explainability in real-world healthcare settings.

The model is designed to automate the classification of knee arthritis X-ray images into five stages: Normal, Doubtful, Mild, Moderate, and Severe. This will help doctors to develop a better tool to identify the progression of arthritis with more consistency and precision. Early detection of the stages of arthritis is critical to ensure that patients receive the proper care at the right time, which helps to delay disease progression and improves long-term outcomes. The model aims to handle the limitations brought about by manual diagnoses, which are time-consuming, subjective, and prone to human error.

One of the key objectives of the model is to reduce the diagnostic variability between radiologists. Even experienced doctors have differing opinions on the severity of arthritis when reviewing the same X-ray images, particularly in the early stages of the disease, where differences are subtle. It also has the feature to perform diagnosis consistently so that inconsistencies can be reduced by bringing a uniform, automated classification system. Such diagnosis consistency is crucial in remote or underserved areas where well-trained specialists might not be readily available. With an AI-based tool, primary care doctors or general practitioners could make more accurate assessments to ensure that patients with severe arthritis get timely referrals to specialists.

Another critical objective is that it should allow doctors to follow the evolution of arthritis over time. This model will enable doctors to differentiate between the different stages of arthritis and, hence, provide better patient monitoring. Doctors could alter their treatment plans to slow further degradation if a patient progresses from mild to moderate in several months. Classifying patients stage-wise also enlightens them about more precise functioning or facts about their condition, helping them understand how their disease progresses and what interventions may be necessary.

Apart from achieving high accuracy in the classification tasks, the model seeks to address a key challenge with AI solutions in healthcare: trust and interpretability. Many current models in AI work as "black boxes" and provide results without explanations; hence, doctors are skeptical about relying on their output. This paper focuses on making the model's predictions interpretable so doctors understand each classification's reasoning. It brings in explainable AI to help the model gain the trust of healthcare professionals for better clinical adoption. The overall aim is to develop a practical and scalable solution that will enhance speed and accuracy in diagnosing arthritis, thereby helping doctors with better guidance for patients.

Dataset

The dataset utilized in this study consists of images designated for classifying various stages of knee arthritis. These images are organized into categories representing different severity levels: ‘Normal,’ 'Doubtful,' 'Mild,' 'Moderate,' and 'Severe.' Figure 1 illustrates an example of these different severity levels. The grayscale images undergo preprocessing steps, including resizing and normalization. After preprocessing, each image is resized to a consistent 300x160 pixels. This standardization ensures continuity across the dataset, which is essential for practical model training.

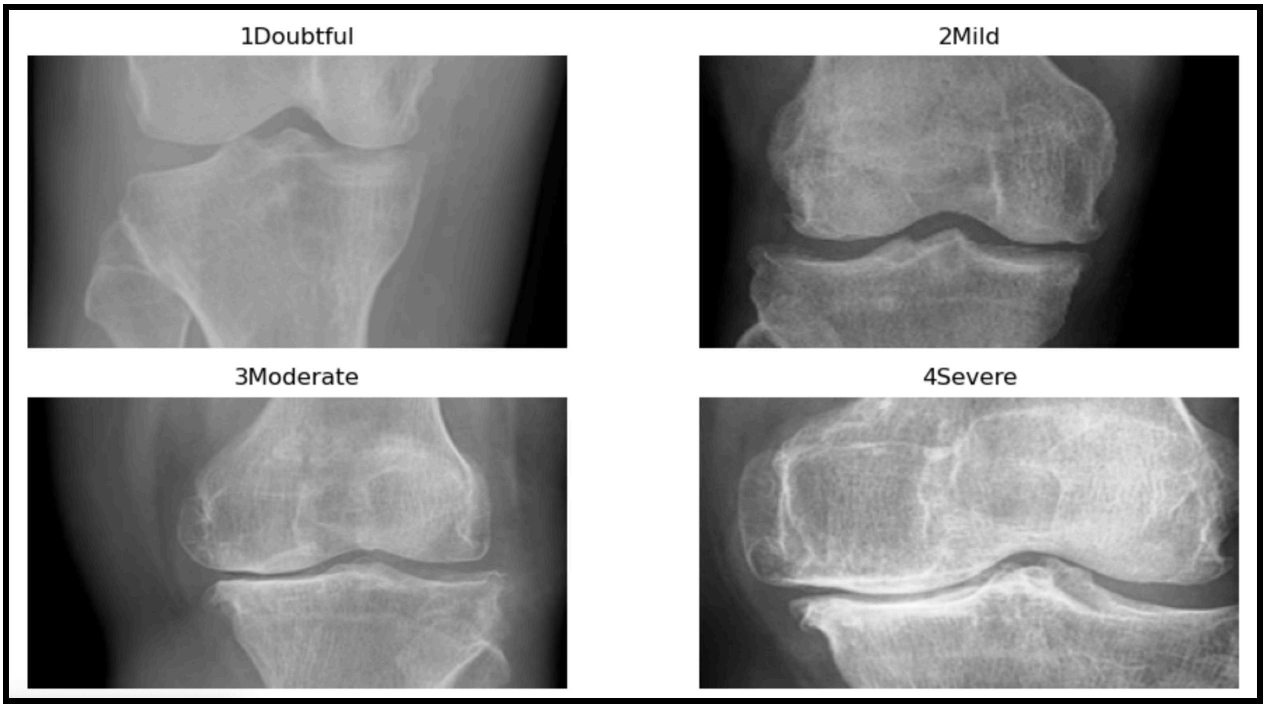


Figure 1. Images of doubtful, mild, moderate, severe cases of arthritis

The dataset contains many images, though the number varies across categories, as depicted in Figure 2. The 'Normal' and 'Doubtful' categories have the highest number of images, with 517 and 47,7, respectively, while the 'Mild,' 'Moderate,' and 'Severe' categories have fewer, with 232, 221, and 206 images each. It splits the images into training and testing subsets in preparation for model development. About 80% of the image dataset shall be used for training, while 20% will be reserved for testing. This split will allow for extensive model training and provide a way to test the model on new, unseen data.

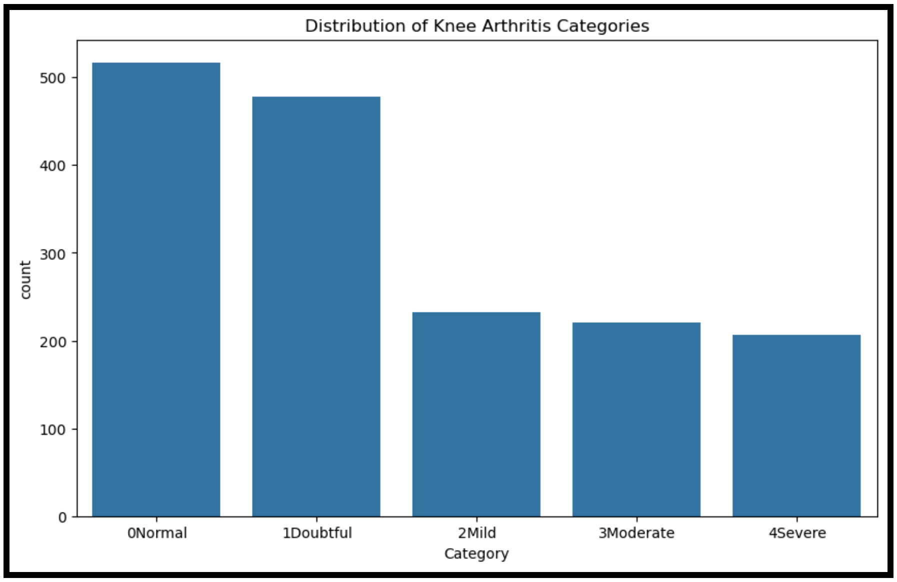


Figure 2. Count distribution of different knee arthritis categories

Methods and Models

CNN(Convolutional Neural Networks)

A CNN is a class of deep learning models visualized to operate on data presented in gridlike structures, for instance, images, videos, and time-series data. The most crucial reason why CNNs outperform other algorithms in computer vision tasks is that they automatically learn and detect spatial hierarchies of features, ranging from low-level details, such as edges, to high-level patterns, such as objects, without explicit manual feature engineering. The key word is that CNNs efficiently preprocess data with local dependencies using convolution operations. This is quite an improvement compared to traditional neural networks, where each input pattern is considered independently. Moreover, CNNs emulate the hierarchical processing of a visual signal in the human brain, where neurons in the visual cortex are responsible for feature recognition at different spatial scales. These bio-inspired architectures improve generalization capabilities in models for a wide range of visual contexts and tasks.

CNN architecture consists of three major components: convolutional, pooling, and fully connected layers. Convolutional Layers: In convolutional layers, filters or kernels slide over the input data, capturing features at a specific spatial location. Typically, pooling layers follow convolutional layers to down-sample data to decrease computational load while making the network invariant or robust to spatial transformations like translation and rotation. Fully connected layers serve to map the learned features to what is usually the ultimate output, typically a classification decision. Finally, an activation function, such as ReLU, follows each convolutional layer to introduce nonlinearities into the network to enrich its capabilities of capturing complicated patterns. Training the CNN involves updating weights through backpropagation and optimization algorithms, such as stochastic gradient descent, through iteratively minimizing the loss function over many iterations of data.

CNNs have revolutionized several areas besides computer vision, including medical imaging, speech recognition, and natural language processing. In medical applications, CNNs are employed to analyze radiological images to assist in anomaly detection, such as tumors or bone fractures, with a very high degree of accuracy. In this line, CNNs are also crucial in object detection in real-time for autonomous driving systems to find pedestrians, traffic signs, and other vehicles. More recently, innovation in CNN architectures such as AlexNet, VGG, ResNet, and Inception has shown impressive performance improvements, establishing new benchmarks on accuracy in image classification and object detection tasks. This has helped increase not only the functionalities but also the application areas for the deep learning models.

InceptionV3

InceptionV3 is a very advanced, deep convolutional neural network that tries to achieve an optimal balance between computational efficiency and image classification accuracy. It was developed as part of the Google Inception family, which introduced the notion of "Inception modules," which simultaneously perform convolutions at multiple scales.

Using parallel convolutions with different filter sizes (1x1, 3x3, and 5x5) and aggregating the outputs, InceptionV3 extracts fine-grained and large-scale image features, making it particularly effective for complex classification tasks.

This work employed a pre-trained InceptionV3 model, fine-tuned for classifying knee arthritis stages. Pre-trained layers of the model were trained on the ImageNet dataset and were initially frozen to leverage their ability to detect basic visual patterns. Then, a custom classification head was added consisting of a GlobalAveragePooling2D layer, a dense layer with 512 neurons and ReLU activation, a dropout layer to mitigate overfitting, and a softmax output layer to classify the five arthritis stages. After training the added layers, fine-tuning was done by unfreezing the last few layers of the InceptionV3 base model and training these with a reduced learning rate, adapting the model to the specifics of the dataset.

EfficientNet

EfficientNet is a state-of-the-art CNN architecture that achieves remarkable performance with minimal computational cost. Unlike traditional CNNs, where the depth or width of the network is often scaled up arbitrarily, EfficientNet uses a principled compound scaling method to systematically scale up all three dimensions of the network: depth, width, and resolution.

For this study, EfficientNetB0, the base variant of the EfficientNet family, was fine-tuned for arthritis classification. The pre-trained model was frozen to retain the generic visual features learned from the ImageNet dataset. A custom classification head was added, consisting of a GlobalAveragePooling2D layer, a dense layer with 512 neurons using ReLU activation, a dropout layer to prevent overfitting, and a softmax layer for multi-class classification. The model was then fine-tuned by unfreezing the last few layers and training them with a lower learning rate to adapt the pre-trained weights to the arthritis dataset.

EfficientNet's compound scaling and squeeze-and-excitation modules Enhanced the model's ability to focus on the most significant regions of the X-ray images while reducing irrelevant noise. This efficiency enabled EfficientNet to excel in medical imaging tasks where computer resources were sometimes in short supply. Combining EfficientNet's scalability with data augmentation and techniques like the ReduceLROnPlateau callback, the model achieved balanced performance across all arthritis stages, even for the underrepresented classes.

DenseNet

DenseNet is a modern CNN architecture that maximizes the information flow between layers by connecting each layer to every subsequent layer in a feed-forward fashion. Such dense connections help reduce the vanishing gradient problem, diminish redundancy, and improve feature reuse, making DenseNet highly effective and efficient for medical imaging tasks.

In this study, the authors used a 121-layer model of DenseNet and fine-tuned it for the classification of arthritis. The pre-trained weights from the model were frozen, which had been trained on ImageNet. A custom classification head was added, which consisted of a GlobalAveragePooling2D layer, a dense layer of 512 neurons using ReLU activation, a dropout layer to handle overfitting, and finally, a softmax layer for the classification of five stages of arthritis and fine-tuning involved unfreezing the last few dense blocks and retraining them with a reduced learning rate.

Densely connected layers in DenseNet allowed the model to grab features from shallow to deep, making it very proficient in describing minute differences among stages in arthritis. Data augmentation techniques and the Adam optimizer were also employed to enhance the model's performance. This efficient use of parameters decreases the chances of overfitting while maintaining the high classification accuracy in DenseNet; hence, it was selected as a robust model for this work.

Evaluation Metrics

A confusion matrix is a simple evaluation device that helps compare predicted and actual labels. This confusion matrix tabulates the actual classes on the rows and predicted classes on the columns to give a complete breakup of model performance across categories. It shows four key outcomes: true positives, indicating correct predictions for positive instances; true negatives, indicating correct predictions for negative instances; false positives, meaning incorrectly predicted positive instances; and false negatives, which are incorrectly predicted negative cases. That would allow for nuancing in the model's mistakes to identify better the specific weaknesses, such as over-predicting some classes or under-predicting others. For example, a high rate of false negatives could be critical in medical diagnosis because it would mean that your model is missing the disease cases.

While the confusion matrix gives some insight into the raw results, the classification report goes into more detail for quantitative analysis of the model's performance with metrics such as precision, recall, F1-score, and support. On the other hand, precision is defined as the ratio of accurate optimistic predictions to all positive predictions; this is very useful in applications where the cost of false positives is prohibitive, such as in spam detection. Recall or sensitivity is the measure that informs about how many relevant instances in the dataset the model can detect; therefore, it is a valuable metric when there are false negatives one would not want to see. The F1-score expresses precision and recall in a single metric by calculating their harmonic mean. Therefore, this gives a good balance between the two when evaluating the accuracy of a model, especially in those cases where precision and recall are at odds. It is the support metric that will essentially provide the number of actual instances per class, thus helping in the evaluation of model performance within the context of a class distribution, making sure example performance is not only for the majority classes but also for the minority ones.

Transfer Learning Model Modifications

In this study, several transfer learning models were fine-tuned to classify knee arthritis stages more accurately and efficiently. Transfer learning involves leveraging pre-trained models that have already learned general features from large datasets and adapting them to the specific task. We want to improve these models by reducing training time, improving their work with small medical datasets, and helping the model perform better. We selected three top CNN architectures for this study: EfficientNetB0, InceptionV3, and DenseNet121. Each model was modified in some ways to improve its performance on the knee arthritis dataset. The base layers of these models were initially frozen to retain the pre-trained knowledge of general image features, such as edges and textures. Subsequently, new layers were added to the models to adapt them for classifying arthritis stages. After training the new layers, fine-tuning was performed by unfreezing select layers from the base models to improve performance by allowing the models to learn more task-specific features from the dataset.

Layer Modifications

For each model, the number of layers unfrozen varied depending on the architecture's depth and complexity. In the case of **EfficientNetB0**, the final 20 layers of the base model were unfrozen to allow the network to learn task-specific features from the arthritis images. Similarly, for **InceptionV3**, the last 50 layers were unfrozen, while for **DenseNet121**, the final dense blocks were unfrozen to refine the model's performance on the target dataset. After unfreezing these layers, we added custom classification heads tailored to the arthritis classification task. The new layers included a **GlobalAveragePooling2D** layer to reduce the dimensionality of the feature maps and retain the most critical features. This was followed by a fully connected **Dense** layer with 512 neurons and **ReLU activation**, which introduced non-linearity to enhance the model's learning capacity. A **Dropout layer** with a dropout rate 0.5 was added to mitigate overfitting by randomly deactivating 50% of neurons during each training iteration. Finally, a **Dense output layer** with five neurons and a **softmax activation function** was added to classify the images into the five categories of arthritis severity: Normal, Doubtful, Mild, Moderate, and Severe.

Compilation and Optimization

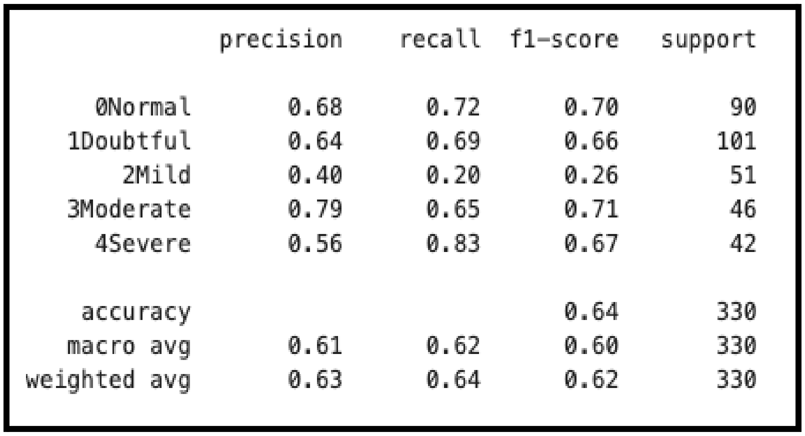
Each modified model was compiled using the **Adam optimizer**, widely used for its adaptive learning rate and efficient handling of sparse gradients. The loss function selected was **sparse categorical cross entropy**, appropriate for multi-class classification problems where the target labels are integers. The model's performance was evaluated using **accuracy** as the primary metric, ensuring the model's ability to classify images into the correct arthritis stages. A **learning rate scheduler (ReduceLROnPlateau)** was employed to optimize training further. This callback function monitored the validation loss during training and reduced the learning rate by 0.5 when the validation loss plateaued for five consecutive epochs. Dynamical learning rate adaptation helped to enhance the convergence of the model by allowing it to take more significant steps during its earlier training while taking smaller steps when it gets closer to finding the optimal solution.

Training Procedure

The training process consisted of two stages: initial training with frozen base layers and subsequent fine-tuning with unfrozen layers. Only the newly added layers were trained during the initial training phase, allowing the model to adapt the high-level features learned from the base model to the arthritis dataset. The fine-tuning phase began once the new layers reached satisfactory performance. In the fine-tuning phase, select layers from the pre-trained base models were unfrozen, and the entire model was trained with a reduced learning rate. This step allowed the models to adjust their pre-trained features better to capture the specific patterns in knee arthritis X-ray images. The models were trained for **100 epochs**, with early stopping criteria to prevent overfitting. Data augmentation techniques were applied during training to improve the model's robustness by introducing variations in the input images, such as rotations, zooms, and flips. Adding previous knowledge, tailor-made classification heads, and fine-tuning improved the performance measures of all three models, thus yielding balanced accuracy scores for the different stages of arthritis, including even the rarer classes like Mild and Severe.

Results and Discussion

Base CNN Model:



```
              precision    recall  f1-score   support

     0Normal       0.68      0.72      0.70        90
   1Doubtful       0.64      0.69      0.66       101
       2Mild       0.40      0.20      0.26        51
   3Moderate       0.79      0.65      0.71        46
     4Severe       0.56      0.83      0.67        42

    accuracy                           0.64       330
   macro avg       0.61      0.62      0.60       330
weighted avg       0.63      0.64      0.62       330
```

Figure X - Testing data classification report.

The classification report presented in **Figure X** provides a detailed evaluation of the Base CNN model’s performance across five categories of knee arthritis severity: **Normal**, **Doubtful**, **Mild**, **Moderate**, and **Severe**. The primary metrics analyzed in the report include **Precision**, **Recall**, **F1-score**, and **Support**, which collectively offer insights into the model's capability to classify X-ray images accurately.

The overall accuracy rate of this model was 64%, meaning it correctly classified 64% of all test images regarding the multiple categories. A macro average F1 Score of 0.60 was obtained, depicting the model's relatively good performance across all classes independent of the sample size within each class. Meanwhile, the weighted average F1 Score, considering the class imbalances, showed a slightly better result, having a score of 0.62—a token of this model's better performance when dealing with classes of more significant populations.

Among the five classes, the model shows the highest performance for the **Moderate** category, achieving an **F1-score of 0.71**. This indicates that the model effectively identifies X-ray images with moderate arthritis symptoms. The **Severe** category also demonstrates strong performance, with a **recall of 0.83**, meaning that the model correctly identified 83% of the actual severe cases in the test set. The model's high recall rate for severe cases is significant in healthcare settings because it ensures that those with advanced arthritis are correctly identified for urgent medical treatment.

On the other hand, the model has a lot of trouble correctly classifying the Mild category, reaching a low F1 score of 0.26. The recall rate for instances classified in the Mild category is only 0.20, indicating that the model does not classify 80% of these cases as such. This situation brings about great concern since the timely diagnosis of arthritis is necessary for slowing disease progression and improving long-term outcomes. The suboptimal performance in this class suggests that the model has difficulty distinguishing subtle differences between early-stage arthritis and other disease stages. This highlights the need for further model refinement.

The **Normal** and **Doubtful** categories show moderate performance, with **F1 scores of 0.70** and **0.66**, respectively. The model's **recall for Normal cases is 0.72**, indicating that most healthy knee X-rays are correctly identified. However, some confusion exists between **Normal** and **Doubtful** cases, likely due to overlapping visual features in the X-ray images.

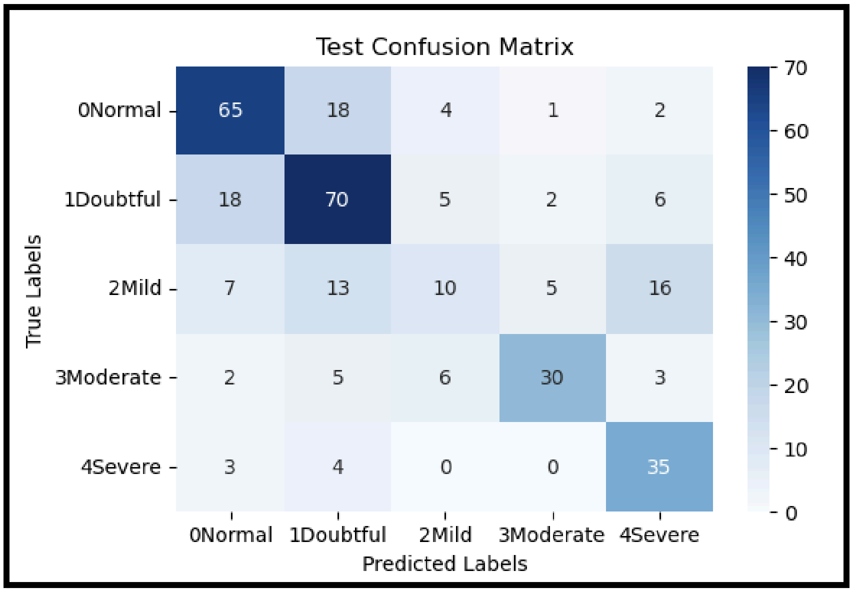


Figure Y - Testing Data Confusion Matrix

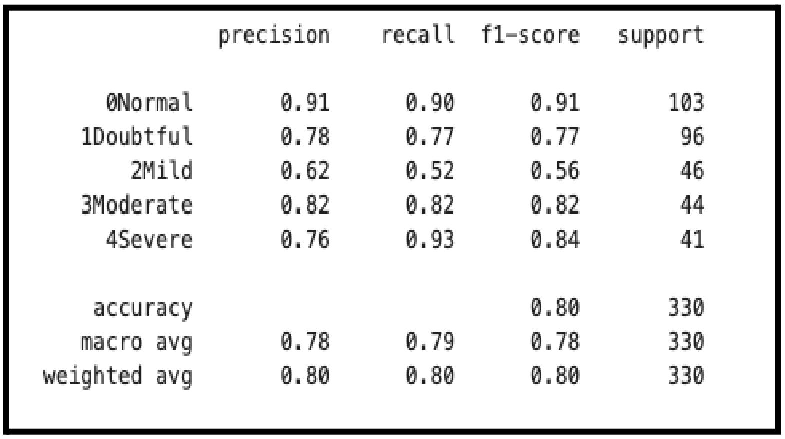
Another good way to judge performance is to use the confusion matrix shown in Figure Y. It offers a more thorough look at model predictions than the test set's actual labels. In that matrix, rows represent actual classes, and columns represent predicted classes. The diagonal values indicate correct predictions, while off-diagonal values represent misclassifications.

The Base CNN model correctly classified 65 out of 90 Normal cases, giving this class an accuracy of 72%. However, it misclassified 18 Normal cases to the Doubtful class, proving that sometimes the model finds it challenging to classify minor abnormalities from healthy knees. Similarly, 70 out of 101 doubtful cases were correctly classified; however, 18 were misclassified as Normal.

This makes the model's performance in the mild class very worrying. Only 10 out of the 51 cases classified as Mild were correctly classified, with 16 misclassified as severe and 13 as Doubtful. This indicates that the model tends to overestimate the severity of mild cases, which could lead to unnecessary alarm in real-world medical applications. The inability to correctly classify mild cases may be due to the subtle differences in early-stage arthritis X-ray images, which are harder to detect without more refined features.

On the other hand, the model shows good performance in the Moderate and Severe classes. It correctly classified 30 of 46 cases labeled as Moderate and 35 of 42 instances labeled as Severe, which shows that it can also reliably detect more progressed stages of arthritis. Notably, the Severe class has a high recall value of 0.83, meaning the model rarely misses severe cases. However, the confusion matrix also indicates that a small number of **Severe cases** were misclassified as **Normal**, which could be problematic in clinical practice if patients with advanced arthritis are overlooked.

DenseNet121 CNN implementation Model:



```
              precision    recall  f1-score   support

     0Normal       0.91      0.90      0.91       103
   1Doubtful       0.78      0.77      0.77        96
       2Mild       0.62      0.52      0.56        46
   3Moderate       0.82      0.82      0.82        44
     4Severe       0.76      0.93      0.84        41

    accuracy                           0.80       330
   macro avg       0.78      0.79      0.78       330
weighted avg       0.80      0.80      0.80       330
```

Figure E - Testing Data Classification Report (DenseNet121)

The classification report for the DenseNet121 model (shown in **Figure E**) demonstrates its overall performance across five categories of knee arthritis severity: **Normal**, **Doubtful**, **Mild**, **Moderate**, and **Severe**. The report presents key evaluation metrics, including **Precision**, **Recall**, **F1-score**, and **Support**, which provide insights into the model's ability to classify X-ray images into different arthritis stages accurately.

The model achieved an overall **accuracy of 80%**, indicating that it correctly classified 80% of the test images. The **macro average F1 Score** was **0.78**, and the **weighted average F1 Score** was also **0.80**, showing balanced performance across all categories.

The model shows the highest performance for the **Normal** category, achieving a **precision of 0.91**, **recall of 0.90**, and an **F1-score of 0.91**. This indicates that the model can reliably identify healthy knee X-rays with minimal false positives and negatives. For the **Doubtful** category, the model achieved an **F1-score of 0.77**, with a precision of **0.78** and a recall of **0.77**. These metrics suggest the model performs well in identifying doubtful cases but still misclassifies some images, as shown in the confusion matrix.

The performance for the **Moderate** category is strong, with an **F1 Score of 0.82**, indicating that the model accurately identifies patients in the intermediate stages of arthritis. In the severe category, the model achieved an F1 Score of 0.84 and a recall rate of 0.93, correctly identifying 93% of the severe cases. The high recall rate in this category guarantees that most patients with advanced arthritis will get the medical care they need.

The model's most significant challenge is the mild class, with an F1 Score of 0.56. A recall of 0.52 means that the model correctly identifies only 52% of actual mild cases, indicating difficulty distinguishing early-stage arthritis from other classes.

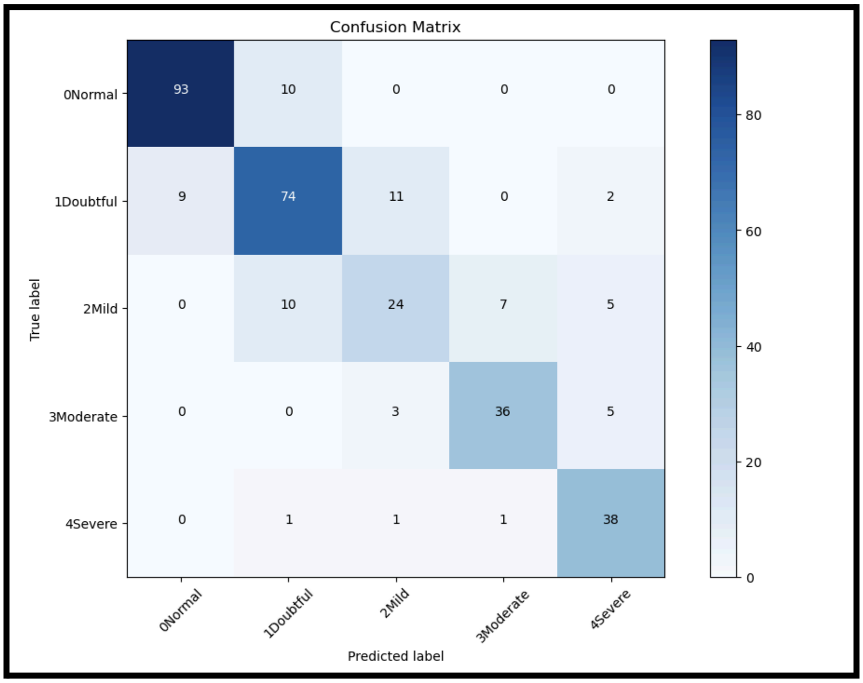


Figure F - Testing Data Confusion Matrix (DenseNet121)

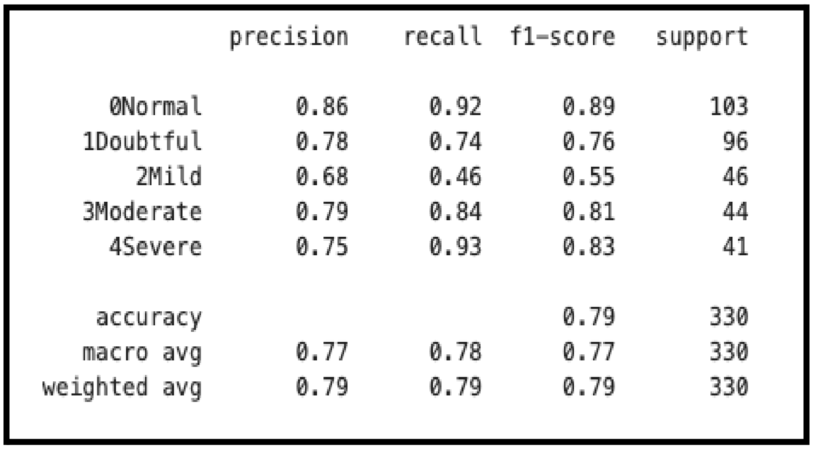
The confusion matrix (shown in **Figure F**) visually represents the model’s predictions compared to the actual labels in the test set.

The DenseNet121 model correctly classified **93 out of 103 Normal cases**, resulting in a **90% accuracy** for this category. However, **10 Normal cases** were misclassified as **Doubtful**, suggesting that some healthy images share features with borderline cases. For the **Doubtful** category, the model correctly classified **74 out of 96 cases**, but **11 cases were misclassified as Mild** and **9 as Normal**, indicating that the model occasionally struggles to distinguish between mild and doubtful stages of arthritis.

The confusion matrix shows that the mild class is the most difficult for this model to recognize. Of the 46 samples assigned to this class, only 24 are correctly identified, while 10 are wrongly classified as Doubtful, seven as Moderate, and five as severe. This misclassification pattern tends to mean that the model often overestimates the mild class, which might lead to false alerts within clinical settings.

The model performs well in the **Moderate** category, correctly classifying **36 out of 44 cases**. The confusion matrix shows minimal misclassification for moderate cases, with only **three misclassified as Mild** and **five as severe**. The model correctly classified **38 out of 41 cases for the severe category.**

InceptionV3 CNN implementation Model:



```
              precision    recall  f1-score   support

     0Normal       0.86      0.92      0.89       103
   1Doubtful       0.78      0.74      0.76        96
       2Mild       0.68      0.46      0.55        46
   3Moderate       0.79      0.84      0.81        44
     4Severe       0.75      0.93      0.83        41

    accuracy                           0.79       330
   macro avg       0.77      0.78      0.77       330
weighted avg       0.79      0.79      0.79       330
```

Figure G - Testing Data Classification Report (inceptionV3)

The model achieved an overall **accuracy of 79%**, indicating that it correctly classified 79% of the test images across all categories. The **macro average F1 Score** was **0.77**, and the **weighted average F1 Score** was **0.79**, demonstrating consistent performance across majority and minority classes.

The model performs best in the **Normal** category, achieving a **precision of 0.86**, a **recall of 0.92**, and an **F1 Score of 0.89**. This indicates that the model can accurately identify healthy knee X-rays and minimize false positives and negatives.

The performance metrics for the Doubtful category are also strong, with an F1 Score of 0.76, precision of 0.78, and recall of 0.74. Those metrics indicate that the model is quite good at accurately identifying doubtful instances, although there is room for improvement concerning the nearby classes.

The Moderate category has a high F1-score of 0.81, which means the model correctly identifies patients with moderate arthritis. With a recall of 0.84 in this category, the model captures the mildest cases necessary for initiating early medical intervention.

The model performs exceptionally well in the severe category, achieving a high F1 Score of 0.83 and a high recall of 0.93. The high recall indicates that the model is very good at identifying cases of severe arthritis, thus ensuring that patients with advanced conditions are identified for further medical care. However, the precision score of this class is 0.75, which means there are some false positives, meaning that the model sometimes mistakenly classifies non-severe cases as severe.

The **Mild** category remains the most challenging for the model, with an **F1 Score of 0.55** and a **recall of 0.46**. This indicates that the model struggles to identify early-stage arthritis cases correctly.

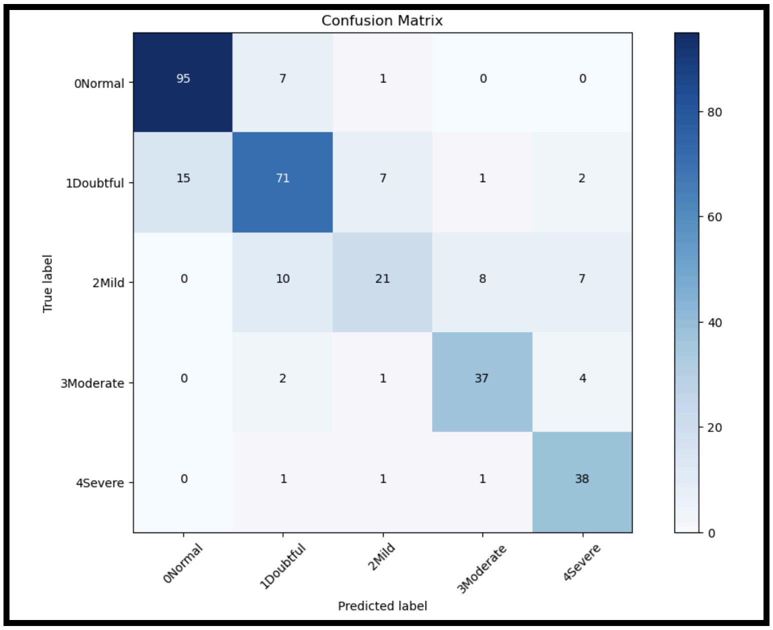


Figure H - Testing Data Confusion Matrix (InceptionV3)

The InceptionV3 model correctly classified **95 out of 103 Normal cases**, resulting in a **92% accuracy** for this category. However, **7 Normal cases** were misclassified as **Doubtful**, and **1 was misclassified as Mild**, indicating that the model occasionally confuses healthy knees with borderline cases.

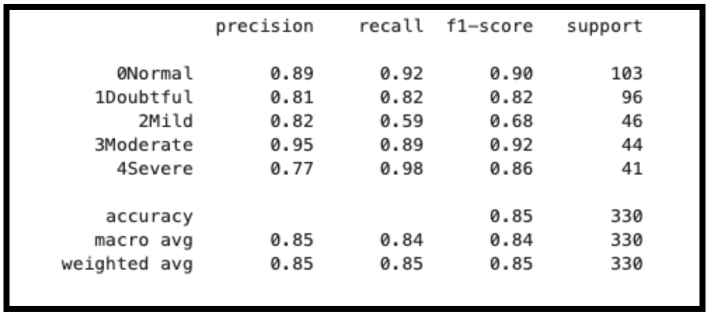
For the **Doubtful** category, the model correctly classified **71 out of 96 cases**, but **15 cases were misclassified as Normal**, and **7 were misclassified as Mild**. These results suggest that the model occasionally struggles to distinguish between doubtful cases and adjacent categories, as well as exceptionally normal and mild cases.

The confusion matrix shows that the Mild class is quite challenging for the model. Of the 46 cases labeled Mild, only 21 were correctly classified, while 10 were misclassified as Doubtful, 8 as Moderate, and 7 as severe. These misclassifications indicate that the model tends to overestimate the severity of mild cases, which could lead to unnecessary concern in clinical practice.

The model performs well for the **Moderate** category, correctly classifying **37 out of 44 cases**. The confusion matrix shows minimal misclassification for moderate cases, with only a few misclassified as Mild or Severe.

In the severe category, the model correctly classified 38 of the 41 cases, resulting in a high recall rate of 93%. This indicates that the model is very good at identifying patients with advanced arthritis. In particular, one severe case was misclassified as Doubtful, and another was misclassified as Mild.

EfficientNetB80 CNN implementation Model:



```
              precision    recall  f1-score   support

     0Normal       0.89      0.92      0.90       103
   1Doubtful       0.81      0.82      0.82        96
       2Mild       0.82      0.59      0.68        46
   3Moderate       0.95      0.89      0.92        44
     4Severe       0.77      0.98      0.86        41

    accuracy                           0.85       330
   macro avg       0.85      0.84      0.84       330
weighted avg       0.85      0.85      0.85       330
```

Figure I - Testing Data Classification Report (EfficientNetB80)

The classification report for the EfficientNetB0 model (shown in **Figure I**)

The model achieved an overall **accuracy of 85%**, indicating that it correctly classified 85% of the test images across all categories. The **macro average F1 Score** and the **weighted average F1 Score** were both **0.85**, demonstrating consistent performance across all five categories.

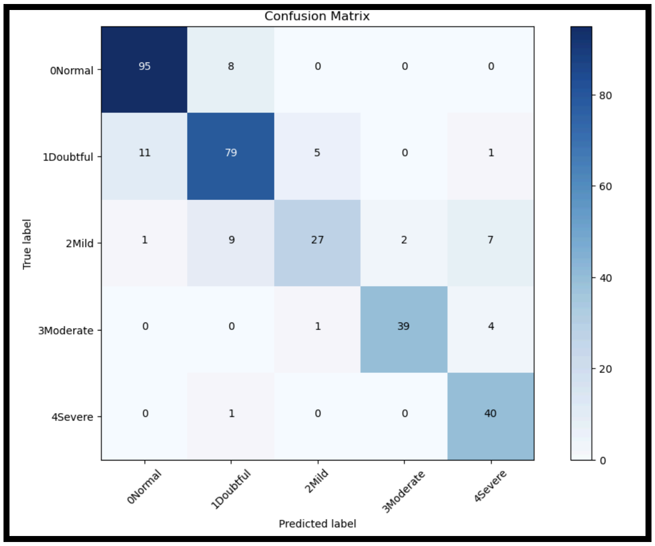
The Normal class performs best, with the model achieving a precision of 0.89, a recall of 0.92, and an F1 Score of 0.90. This implies that the model is excellent at consistently finding healthy knee X-rays and keeping false positives and negatives low, meaning it retains many correct classifications for typical cases.

For the Doubtful category, the model achieved a precision of 0.81, a recall of 0.82, and an F1-score of 0.82. These metrics suggest that the model performs well in identifying borderline cases of arthritis but occasionally misclassifies them as usual or mild, as reflected in the confusion matrix.

The model performs well in the Moderate category, with an F1 Score of 0.92 and a precision of 0.95. These values mean that the model classifies instances of moderate arthritis with high confidence. The class's 0.89 recall rate ensures that most mild cases are correctly identified, essential for initiating early medical treatment.

The severe class performs well, scoring 0.86 and recalling 0.98. The high recall rate indicates the model could recognize nearly all severe cases from the test data. Nevertheless, its precision score of 0.77 shows that specifications were misclassified; cases of less severe diseases were put into the class of the more severe diseases, resulting in false positives.

The **Mild** category remains the most challenging for the model, with an **F1 Score of 0.68** and a **recall of 0.59**. This indicates that the model struggles to identify early-stage arthritis cases accurately. The confusion matrix shows mild cases are often misclassified as doubtful or moderate.



	0Normal	1Doubtful	2Mild	3Moderate	4Severe
0Normal	95	8	0	0	0
1Doubtful	11	79	5	0	1
2Mild	1	9	27	2	7
3Moderate	0	0	1	39	4
4Severe	0	1	0	0	40

Figure J - Testing Data Confusion Matrix (EfficientNetB80)

The EfficientNetB0 model correctly classified 95 of the 103 Normal cases, which corresponds to an accuracy of 92% for this class. Nevertheless, 8 Normal cases were incorrectly assigned to the Doubtful class, showing that the model can sometimes be confused when dealing with healthy knees and borderline cases. Additionally, 1 Normal case was misclassified into the Mild class, proving the model is relatively reliable but imperfect in discriminating standard cases from early-stage arthritis.

For the **Doubtful** category, the model correctly classified **79 out of 96 cases**, but **11 were misclassified as Normal**, and **5 were misclassified as Mild**. These results suggest that the model occasionally struggles to distinguish doubtful cases from adjacent categories and exceptionally normal and mild cases.

The confusion matrix shows that the Mild category represents significant challenges for the model: Out of 46 Mild cases, only 27 have been correctly predicted, while 9 have been classified as Doubtful, 2 as Moderate, and seven as severe. Those misclassifications may indicate the model's inclination to overpredict mild cases, leading to inappropriate alerts in a clinical setting.

The model performs well for the **Moderate** category, correctly classifying **39 out of 44 cases**. The confusion matrix shows minimal misclassification for moderate cases, with only a few misclassified as Mild or Severe.

It correctly classified 40 of the 41 cases within the severe category, giving a tremendous recall rate of 98%. This result would mean the model has learned to distinguish patients with advanced arthritis. Indeed, one severe case was misclassified as Doubtful.

Conclusion

The main goal of this research was to develop and optimize convolutional neural network models for automatically classifying knee arthritis stages based on X-ray images. This study attempted to overcome the drawbacks of manual diagnosis., which can be subjective, time-consuming, and prone to human error. Among the models tested, EfficientNetB0 outperformed the others, achieving the best balance between precision, recall, and scores across all categories.

The EfficientNetB0 model identified Normal, Moderate, and severe arthritis cases; it achieved a high F1 Score for the moderately affected group and good recall for cases belonging to the severe class, an essential factor for early medical intervention. However, it struggled to classify cases with mild arthritis, indicating that more refinement is required to diagnose the early stages of arthritis.

The DenseNet121 and InceptionV3 architectures achieved similar performances, with overall accuracies of 80% and 79%, respectively. All models distinguished between the regular and severe classes well but performed significantly poorly in correctly classifying cases with mild effects. The baseline convolutional neural network performed the worst, further ascertaining the benefits of using pre-trained transfer learning models in medical imaging tasks.

Significance and Impact

The findings of this study have important implications for the medical community. Deep learning can help automate the classification of stages of knee arthritis, decreasing diagnostic variability and increasing diagnostic accuracy. This can speed up the detection of arthritis, mainly in resource-poor areas. Eventually, this will lead to earlier treatment interventions that have the potential to slow disease progression and thereby reduce the need for expensive imaging techniques like MRIs. The paper presents how artificial intelligence can help radiologists manage their workload and engage with patients. An automated framework can also be used as a decision-support system that consistently and reliably offers classifications, improving patients' outcomes. Additionally, these models integrated into mobile applications or cloud-based platforms bear enormous potential for enhancing accessibility to diagnostics in remote and under-resourced locations.

Limitations of Your Study

While these findings are encouraging, the present study has several limitations: the dataset used to test the approach included a relatively small number of examples; this may be insufficient to generalize the model to more extensive and more diverse populations. Additionally, the dataset could include biases tied to specific demographics that may lead to less accurate generalization when applying the model to patients belonging to other demographics. The model also struggled to classify **Mild** cases accurately, highlighting the need for further improvements in feature extraction to capture subtle differences in early-stage arthritis.

Work Cited

1. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., & Saarakkala, S. (2018). **Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach**. *Scientific Reports, 8*(1), 1727. https://doi.org/10.1038/s41598-018-20132-7
2. Antony, J., McGuinness, K., O'Connor, N. E., & Moran, K. (2016). **Automatic Detection of Knee Osteoarthritis from X-ray Images Using Deep Learning**. *International Symposium on Biomedical Imaging*. https://doi.org/10.1109/ISBI.2016.7493411
3. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). **ImageNet Classification with Deep Convolutional Neural Networks**. *Advances in Neural Information Processing Systems*, 25, 1097-1105. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks
4. Zhang, X., Wang, Z., Liu, D., & Li, Y. (2020). **EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks**. *Proceedings of the 36th International Conference on Machine Learning*. https://arxiv.org/abs/1905.11946
5. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., & Ghafoorian, M. (2017). **A Survey on Deep Learning in Medical Image Analysis**. *Medical Image Analysis, 42*, 60-88. https://doi.org/10.1016/j.media.2017.07.005
6. Wang, S., Yin, Y., Cao, G., Wei, B., Zheng, Y., & Yang, G. (2021). **Interpretability of Deep Learning in Healthcare: A Technical Overview**. *IEEE Access, 9*, 103931-103946. https://doi.org/10.1109/ACCESS.2021.3089824
7. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., & Mehta, H. (2018). **Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists**. *PLOS Medicine, 15*(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686
8. Shamir, L., Ling, S. M., Scott, W. C., & Ferrucci, L. (2009). **Knee X-ray Image Analysis to Detect Osteoarthritis**. *BMC Medical Imaging, 9*(1), 1-9. https://doi.org/10.1186/1471-2342-9-10

9. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., & Blau, H. M. (2017). **Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks**. *Nature, 542*(7639), 115-118. https://doi.org/10.1038/nature21056
10. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). **Artificial Intelligence in Radiology**. *Nature Reviews Cancer, 18*(8), 500-510. https://doi.org/10.1038/s41568-018-0016-5
11. Zhang, Z., Yang, L., & Zheng, Y. (2021). **Classification of Osteoarthritis Severity Using Deep Learning on X-ray Images**. *Journal of Medical Imaging, 8*(4), 041206. https://doi.org/10.1117/1.JMI.8.4.041206
12. Ronneberger, O., Fischer, P., & Brox, T. (2015). **U-Net: Convolutional Networks for Biomedical Image Segmentation**. *Medical Image Computing and Computer-Assisted Intervention*. https://arxiv.org/abs/1505.04597
13. LeCun, Y., Bengio, Y., & Hinton, G. (2015). **Deep Learning**. *Nature, 521*(7553), 436-444. https://doi.org/10.1038/nature14539
14. Ghorbani, A., Ouyang, D., Abid, A., He, B., & Chen, P. H. C. (2019). **Deep Learning Interpretation of Echocardiograms**. *Nature Medicine, 26*(5), 886-891. https://doi.org/10.1038/s41591-019-0447-x
15. Thomas, D. J., & Young, S. W. (2020). **The Role of AI in Predicting Osteoarthritis Progression**. *Orthopaedic Journal of Sports Medicine, 8*(1), 2325967120903084. https://doi.org/10.1177/2325967120903084
16. Ma, D., He, X., & Zhang, Y. (2019). **Explainable AI: Developing Trust in Medical AI Applications**. *Computers in Biology and Medicine, 114*, 103491. https://doi.org/10.1016/j.compbiomed.2019.103491
17. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., & Anguelov, D. (2015). **Going Deeper with Convolutions**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2015.7298594
18. Liu, F., Zhou, Z., Samsonov, A., Blankenbaker, D., Larison, W., & Kanarek, A. (2018). **Deep Learning for Knee MRI Analysis**. *Radiology, 289*(1), 160-169. https://doi.org/10.1148/radiol.2018171843
19. Chesbrough, H. W., & Appleyard, M. M. (2007). **Open Innovation and Arthritis Research**. *California Management Review, 50*(1), 57-76. https://doi.org/10.2307/41166416
20. WHO. (2023). **Global Burden of Arthritis**. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/arthritis