Severity Classification of Knee Osteoarthritis from X-Ray Images using Deep Learning

MSc Research Project
MSc Data Analytics

Pranav Prakash
Student ID: x23134682

School of Computing
National College of Ireland

Supervisor: Furqan Rustam

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Pranav Prakash		
Student ID:	x23134682		
Programme:	MSc. Data Analytics	**Year:**	Jan2024 - Jan2025
Module:	MSc. Research Project		
Supervisor:	Furqan Rustam		
Submission Due Date:	December 12, 2024		
Project Title:	Severity Classification of Knee Osteoarthritis from X-Ray Images using Deep Learning		
Word Count:	6633	**Page Count:**	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Pranav Prakash

Date: December 12, 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Severity Classification of Knee Osteoarthritis from X-Ray Images using Deep Learning

Pranav Prakash

x23134682

Abstract

Osteoarthritis of the knee (KOA) is the most common cause of disability caused by cartilage degeneration and consequent joint failure. Getting diagnosed early is key to avoiding long-term pain and degeneration, mental health, and mobility issues. While traditional manual analysis of X-ray images can be subjective, time consuming and inconsistent, automated solutions are desired. Advances in deep learning have recently demonstrated effective use of CNN architectures including DenseNet121, EfficientNetB0 and MobileNet to perform superior to traditional machine learning methods on medical imaging tasks. This study explores the effectiveness of these models for classifying KOA severity at three levels of granularity: A three-class, five-class, and binary classification, on a dataset graded using the Kellgren–Lawrence system. Accuracy, F1-score and confusion matrices were used in evaluation of the models. Results were found to indicate that binary classification with DL models consistently outperformed conventional ML methods, with DenseNet121 being the most accurate (78.27%). As classification granularity decreased, performance improved, confirming the contribution of simplified tasks to ameliorating class imbalance and improve generalization. ML models such as Random Forest had a moderate outcome, but they exhibited a failure to deal with high dimensional data. Finally, this research has shown the basis for DL as an automation method of KOA diagnostics, and potential areas are suggested for future research exploiting hybrid models and clinical metadata integration.

Keywords: CNN (Convolutional Neural Network), KOA (Knee Osteoarthritis), Classification, Accuracy, F1-Score, Machine Learning.

1 Introduction

Knee osteoarthritis, where degeneration leads to loss of articular cartilage and ultimately joint failure is a major source of mobility and quality of life impairment and is a major contributor to global disability and socioeconomic burden. Manual analysis of X-ray images through traditional diagnostic methods are subjective, time consuming and prone to inconsistencies leading to early and accurate diagnosis as a challenge. Early diagnosis is important as delays can lead to long term pain, restricted mobility and mental health problems known as anxiety and depression.

CNNs have recently revolutionised medical image analysis with automated, efficient, and accurate diagnostic solutions. Deep CNN architectures like VGG16, ResNet, DenseNet, Inception are getting popular in medical image tasks including detecting tumours in the brain and grading severity of KOA (Mukesh Kumar and Sree Sankar, 2024; Divyanshi and Bansal, 2024). These deep learning models rely on hierarchical feature extraction that provide a superior performance on the use case of complex image classification compared to traditional models.

This study investigates the research question:
How does the performance of deep learning models and machine learning models change when varying the classification granularity of detecting severity of knee osteoarthritis?

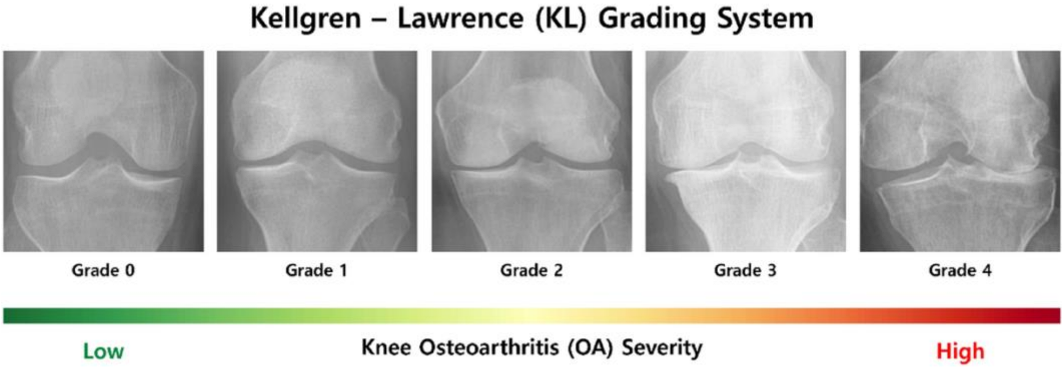

Kellgren – Lawrence (KL) Grading System
Grade 0
Grade 1
Grade 2
Grade 3
Grade 4
Low
Knee Osteoarthritis (OA) Severity
High

Figure 1: KL Grading System.

As a solution to this, the research compares the classification's ability of these advanced CNN architecture along with traditional machine learning models for the KOA severity classification. The analysis starts with a five-class classification based on Kellgren Lawrence grading and then several reductions to three class and two class models to study the impact of classification granularities on model performance. The above models are assessed in terms of evaluation metrics including accuracy and F1 score and so on.

The purpose of this study was to use its results to shed light on the comparative advantages and shortcomings of state-of-the-art CNN and ML architectures for KOA severity assessment. This work makes contributions to developing such efficient and accurate diagnostic tools that can improve clinical decision making and patient outcomes or to providing a demonstration of when and how deep learning and machine learning can advance medical imaging technology.

The thesis is structured as follows: In Section 2, a review of related work on deep learning and machine learning application in medical imaging is given, particularly in the KOA severity classification. The methodology of the research is detailed in Section 3, where the research starts with the dataset overview, then goes for pre-processing, followed by model selection and lastly, evaluation metrics. Section 5 explains the implementation process of the models and the tools used, and Section 4 contains the design specifications for those models. The results of the evaluation and their discussion are presented in Section 6. Finally, a summary of key findings, limitations and potential directions for future research are presented in Section 7.

2 Related Work

Medical imaging has stepped forward markedly using deep learning as well as machine learning approaches, allowing for highly accurate automated diagnostic systems. In this section, important studies of employing these techniques in datasets, models, results, and limitations are reviewed with a focus to KOA severity classification and allied fields.

2.1 Review on Medical Image Classification using Deep Learning Architectures

The use of Convolutional Neural Networks for classifying the severity of knee osteoarthritis using X-ray images was studied by `Divyanshi and Bansal (2024), Kant (2024). Both studies utilized datasets graded based on the Kellgren and Lawrence scheme, categorizing arthritis into five severity levels: — Normal; Doubtful, Mild, Moderate, Severe. They used a CNN model with three layers of convolution and SoftMax activation, their overall accuracy being 89% and precision scores of 77% for 'doubtful' and 100% for 'moderate' severity. Instead, Kant pre-processed their data with steps including resizing, normalization and implementing data augmentation to assist in training and get a test accuracy of 58%. Both studies showed that CNN could automate KOA diagnostics, but also noted issues in separating individual severity levels. Based upon the results, the authors also suggest future directions including the addition of clinical data, applying more advanced architectures such as transfer learning, and overcoming overfitting to reach higher diagnostic accuracy and applicability.

Deep learning methodologies for KOA severity classification are offered in Kitukale et al. (2024), Chandu et al. (2024), and Pandey and Kumar (2023), based on DenseNet-201, AlexNet enhanced CNN, and enhanced EfficientNet-B0 architecture, respectively. DenseNet-201 with 83.9% accuracy, shown by Kitukale et al., surpassed the traditional models like CNN and Xception, while Chandu et al. with 96.08% of accuracy using AlexNet surpassed the other models such as DenseNet-121, VGG-19 and MobileNetV2. An improved EfficientNet B0 was suggested by Pandey and Kumar, with an accuracy of 69.74%, when using a dual-branch strategy to enhance feature extraction and classification performance. Taken as a whole, this work highlighted those advanced architectures can automate further towards diagnosing KOA, with external validation, clinical data integration, and real-world deployment, as further recommendations for precise diagnostics and practical utility.

A comparative analysis of CNN architectures such as the VGG16, ResNet, Inception ResNet, and DenseNet for brain tumour detection from MRI images was performed by Kumar and Sankar (2024), whereas Gill et al. (2023) used the pre trained ResNet50V2 model to classify pneumonia using chest X-ray images. Pre-processing techniques (intensity normalization, spatial normalization, skull stripping) were used to prepare MRI sets for binary classification with VGG16 outperforming other architectures with accuracy of 92%, then DenseNet (85%), Inception (84%), and ResNet (80%). The same is done by Gill et al. for fine tuning ResNet50V2 with Adam optimizer and binary cross entropy loss and ends up with 86% post fine tuning. The two studies also highlighted the promise of CNNs in automating medical diagnostic tasks, suggesting that VGG16 may be useful when computational efficiency and model interpretability are needed, and ResNet50V2 useful in resource limited settings. Future work focused on working with various datasets, tuning hyperparameters, and developing useful applications to further improve real world diagnostic ability.

2.2 Review on Medical Image Classification using Machine Learning Models

Various machine learning approaches for the prediction and classification of Rheumatoid Arthritis (RA) have been studied by Sakaria et al. (2023), Singh et al. (2019), and Sundaramurthy et al. (2020), which claim to improve diagnostic accuracy as well as decision making in healthcare. On a dataset of 500 images, five ML models were compared: Convolutional Neural Networks, K-Nearest Neighbours, XGBoost, Gaussian Naive Bayes and Support Vector Machine. They found that CNN performed the best with an accuracy of 98% and the worst at 60% by KNN, which proved to be a weak function for image-based data. On the other hand, they also suggested for Explainable AI (XAI) to boost the prediction transparency, and for an integration of Generative Adversarial Network, and a larger dataset in the long term.

Singh et al. applied the K-means, hierarchical clustering, and DBSCAN algorithms to a dataset of 60 patient profiles through attributes of Rheumatoid Factor and Swollen Joint Count. It was found that the DBSCAN algorithm was the most effective algorithm and adapted well to irregular data patterns and robust to noise. K-Means and hierarchical clustering failed against outliers and high computational cost. The cluster identification performed was important to early RA detection, and the improvement to clustering algorithms is needed to increase diagnostic precision.

Ensemble ML methods were investigated by Sundaramurthy et al. which combines classifiers SVM, AdaBoost, and Random Subspace with baseline models k-Nearest Neighbours and Random Forest. As such, they tested several classifiers on a dataset consisting of a scroll of 1,000 patient profiles and found that SVM based ensemble classifiers achieved the highest accuracy (94%), followed by AdaBoost and RSS. They found the ensemble methods effective in utilizing the powers of individual models skipping the flaws of others. Additional sources of data were proposed to integrate into future RA diagnostic systems, and algorithms were suggested to be refined and increased model robustness remained areas for further work to advance these diagnostic systems. Finally, this work demonstrates that ML techniques have potential for aiding RA diagnosis and further corroborates the usefulness of continued ML research to support this goal.

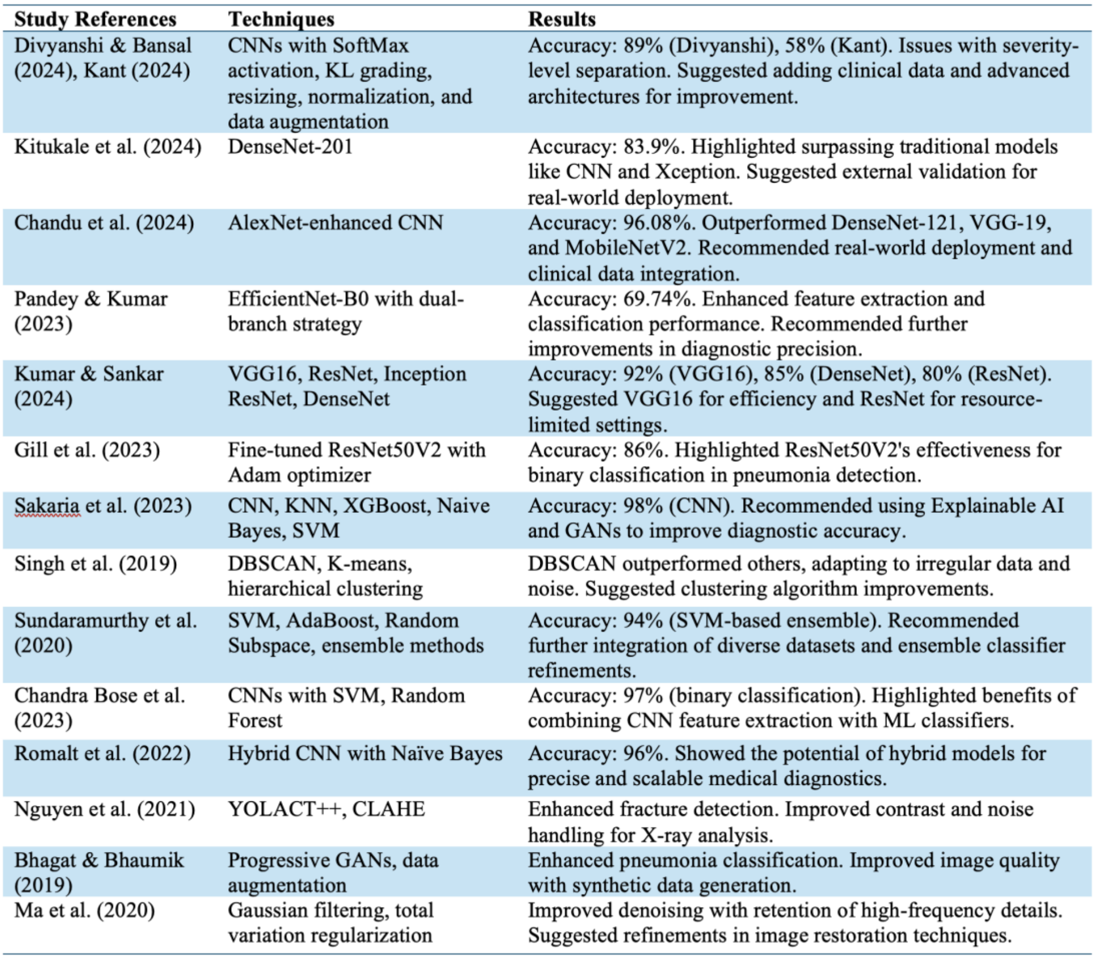
2.3 Review on Combining Deep Learning Architectures with Machine Learning Models

Chandra Bose et al. (2023), Gill et al. (2023), Romalt et al. (2022) have looked at how Medical Diagnostics can benefit from integration of Machine Learning and Deep Learning methods and the possibility of its improved accuracy and efficiency in conjunction. In Chandra Bose et al., CNNs were used to derive features from X-ray images of Knee Osteoarthritis for both binary and multiclass classification. Further these features were processed with ML classifiers like Support Vector Machine produced best result for binary classification task and Random Forest performed better on multi-class task. Using a CNN model DenseNet201, Gill, et al. applied a similar approach for classifying kidney diseases from X-ray images. By using ML algorithm on extracted high level features extracted by DenseNet201, the impressive 97% accuracy was achieved, which shows the power of joining DL's feature extraction with ML's predictive capability.

Additionally, Romalt et al. (2022) presented a hybrid model by combining CNN and Naïve Bayes model for the prediction of cardiovascular disease, using unstructured data to structure it for feature extraction by using CNN and later classification by using Naïve Bayes. This approach combined the DL ability to process complex data with ML efficiency in classification to obtain an average accuracy of 96%. Collectively, these studies show that hybrid ML—DL models can combine the best of both worlds to achieve precise, automated, and scalable medical diagnostics, and call for the future research on their robustness and applicability.

2.4 Review on Image Pre-processing for X-ray Images

YOLACT++ and CLAHE were used to segment and contrast enhancing for arm fracture detection, tackling detection on noisy and low contrast Xray images by Nguyen et al. (2021). With data augmentation techniques such as flipping and brightness adjustments, this approach enhanced their YOLOv4 based fracture detector. To enrich datasets and boost DCNN accuracy, Bhagat and Bhaumik (2019) demonstrated utilizing Progressive GANs (ProGAN) to generate synthetic high-resolution images for detection of pneumonia in chest X-rays along with conventional augmentations. Gaussian filtering and Split Bregman based total variation regularization is introduced by Ma et al. (2020) for retention of high frequency details in denoised industrial x-ray images using partition-based restoration. These studies collectively show the improvement of X-ray image analysis by means of pre-processing techniques, such as segmentation, contrast enhancement, augmentation, and denoising.



Study References	Techniques	Results
Divyanshi & Bansal (2024), Kant (2024)	CNNs with SoftMax activation, KL grading, resizing, normalization, and data augmentation	Accuracy: 89% (Divyanshi), 58% (Kant). Issues with severity-level separation. Suggested adding clinical data and advanced architectures for improvement.
Kitukale et al. (2024)	DenseNet-201	Accuracy: 83.9%. Highlighted surpassing traditional models like CNN and Xception. Suggested external validation for real-world deployment.
Chandu et al. (2024)	AlexNet-enhanced CNN	Accuracy: 96.08%. Outperformed DenseNet-121, VGG-19, and MobileNetV2. Recommended real-world deployment and clinical data integration.
Pandey & Kumar (2023)	EfficientNet-B0 with dual-branch strategy	Accuracy: 69.74%. Enhanced feature extraction and classification performance. Recommended further improvements in diagnostic precision.
Kumar & Sankar (2024)	VGG16, ResNet, Inception ResNet, DenseNet	Accuracy: 92% (VGG16), 85% (DenseNet), 80% (ResNet). Suggested VGG16 for efficiency and ResNet for resource-limited settings.
Gill et al. (2023)	Fine-tuned ResNet50V2 with Adam optimizer	Accuracy: 86%. Highlighted ResNet50V2's effectiveness for binary classification in pneumonia detection.
Sakaria et al. (2023)	CNN, KNN, XGBoost, Naive Bayes, SVM	Accuracy: 98% (CNN). Recommended using Explainable AI and GANs to improve diagnostic accuracy.
Singh et al. (2019)	DBSCAN, K-means, hierarchical clustering	DBSCAN outperformed others, adapting to irregular data and noise. Suggested clustering algorithm improvements.
Sundaramurthy et al. (2020)	SVM, AdaBoost, Random Subspace, ensemble methods	Accuracy: 94% (SVM-based ensemble). Recommended further integration of diverse datasets and ensemble classifier refinements.
Chandra Bose et al. (2023)	CNNs with SVM, Random Forest	Accuracy: 97% (binary classification). Highlighted benefits of combining CNN feature extraction with ML classifiers.
Romalt et al. (2022)	Hybrid CNN with Naïve Bayes	Accuracy: 96%. Showed the potential of hybrid models for precise and scalable medical diagnostics.
Nguyen et al. (2021)	YOLACT++, CLAHE	Enhanced fracture detection. Improved contrast and noise handling for X-ray analysis.
Bhagat & Bhaumik (2019)	Progressive GANs, data augmentation	Enhanced pneumonia classification. Improved image quality with synthetic data generation.
Ma et al. (2020)	Gaussian filtering, total variation regularization	Improved denoising with retention of high-frequency details. Suggested refinements in image restoration techniques.

Table 2: Summary Table

3 Research Methodology

In this part of the study, a step by step methodology utilized within this study for classifying the severity of knee osteoarthritis using both deep learning and traditional machine learning approach has been described.

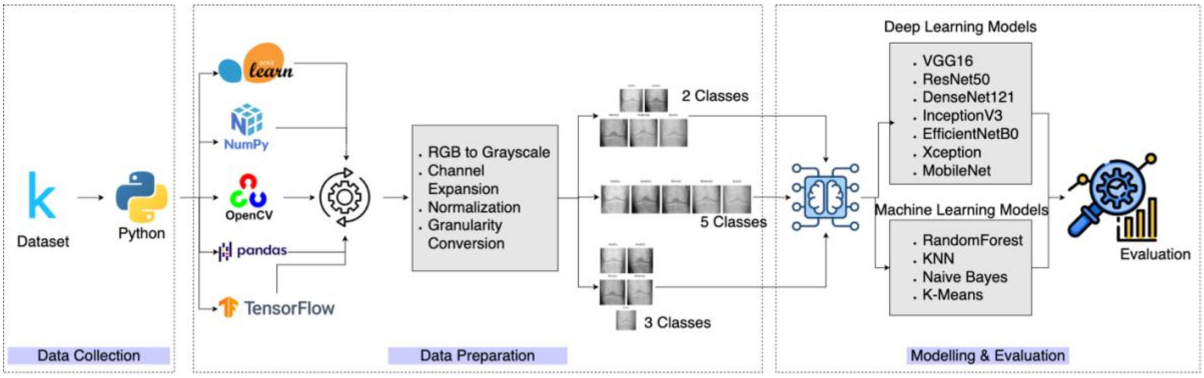


Figure 2: Methodology.

3.1 Dataset Overview

This study used the dataset which was obtained from Kaggle repository which is a popular repository for machine learning and deep learning projects with publicly available datasets. The dataset consists of X-ray images of the knee joint labelled using the Kellgren and Lawrence grading system, which is a well-established standard for rating the severity of knee osteoarthritis This system allows to classify severity at five levels: Class 0 (Normal) to Class 4 (Severe). It provides a rich resource for medical image analysis and machine learning applications research.

The dataset contains an almost 10,000 images, distributed across the following five classes:

Grade 0	Healthy knee image.
Grade 1 (Doubtful)	Joint narrowing but with doubt about osteophytic lipping.
Grade 2 (Minimal)	Osteophytes present, joint space narrowing possible.
Grade 3 (Moderate)	Several osteophytes, definite joint space narrowing, with minimal sclerosis.
Grade 4 (Severe)	Large joint narrowing, severe sclerosis, and large osteophytes.

Table 2: Classes

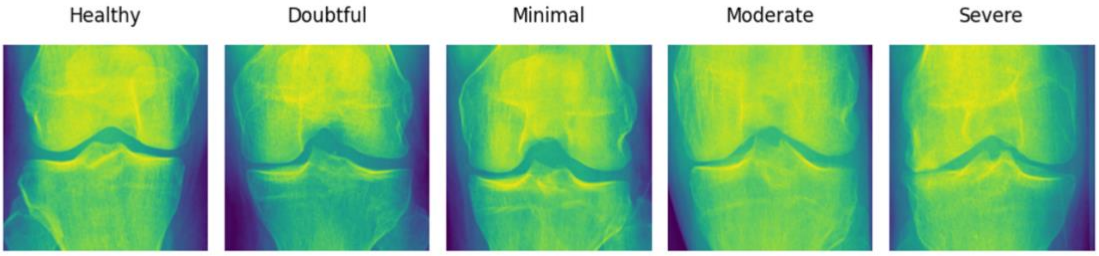


Figure 3: Dataset in RGB.

The planning of the study was much dependent on the size and distribution of the dataset. However, an analysis of the class distribution revealed imbalances in the data where the total number of images is sufficient to train complex deep learning models. For example, a smaller number of images exist in some of the classes like Class 4 which will necessarily affect training model performance. To understand the external dataset composition, visualizations like bar charts showing the number of images per class have been created. To rectify this, main attention during the data preparation phase was directed to address these imbalances through techniques such as data augmentation, oversampling and under sampling.

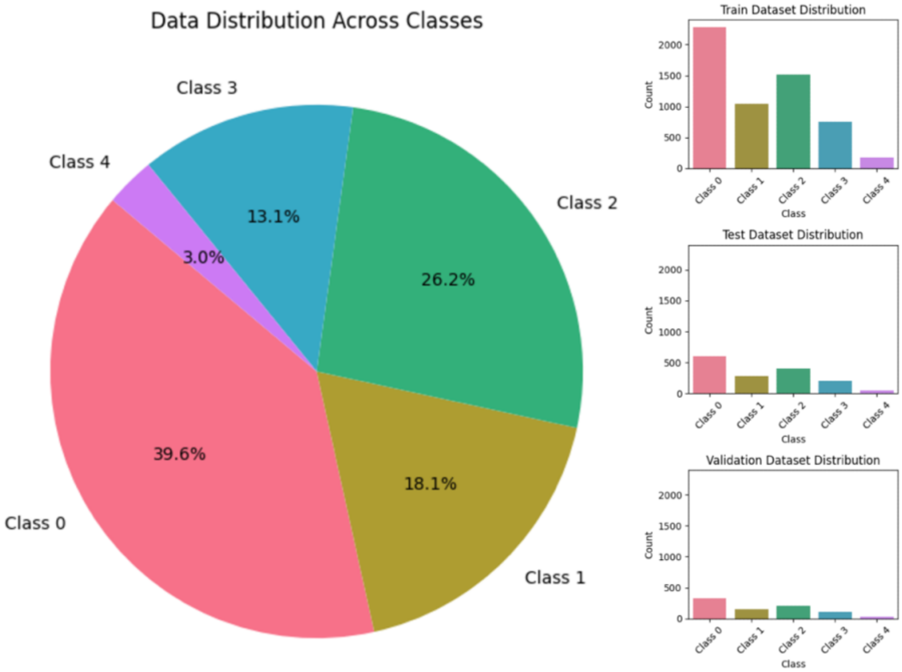


Figure 4: Dataset Distribution.

Data quality and relevance of the dataset was validated by inspecting sample images for the different classes, showcasing unique features, like cartilage loss and bone deformities, that characterize the evolution of knee osteoarthritis severity between Class 0 and Class 4. Representative images, including, for example, a grid of X-rays covering classes, would neutrally illustrate these differences. The structured labelling scheme and broad coverage across severity levels of the dataset make it a sound reference for training, validation, and testing of automated KOA severity classification with machine learning and deep learning approaches.

3.2 Data Preparation

Several steps were done during the data preparation, which aims at improving or enhancing the data quality and utility for use while training deep learning and machine learning models. These made the models able to feed and learn with the data.

3.2.1 RGB to Grayscale Conversion

The data was images in RGB format, that is, three colour channels (red, green, blue), for representing visual information. RGB images provide rich information on colour but due to medical image classification task such as diagnostic based on texture or intensity, the colour is not necessary. The basis for the first step was the transfer of images to grayscale to decrease the computational complexity and concentrate on the relevant features. Converting this data simplifies it down to intensity values, reducing memory needs and speeding up processing without compromising of critical diagnostic features.

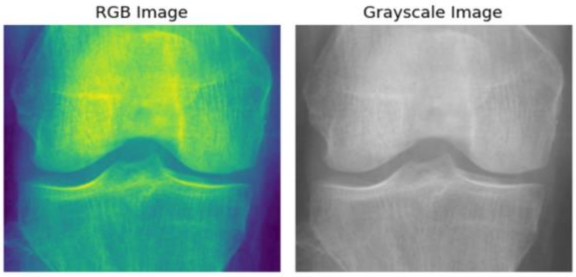


Figure 5: RGB to Grayscale.

3.2.2 Channel Expansion

Grayscale images tend to reduce computations but many pre trained convolutional neural network (CNN) like VGG16, ResNet50, and DenseNet are modelled in the assumption of three channel (RGB) images. Because of these models, the single grayscale channel was expanded to three identical channels to ensure compatibility. This does not introduce additional information but makes it easy to plug in pre trained models, that take inputs of a certain shape, such as height × width × 3. This step is part to allow the feature maps generated from the pre trained network to match the given architecture design of the networks.

Figure 6: 3-Channel Grayscale.

3.2.3 Normalization

Images have raw pixel values in range of 0 to 255. To increase model training efficiency, and to make the dataset stable, we performed min-max normalization on pixel values to put them in the range [0,1]. Normalizing data becomes a necessity to ensure that the numerical scale of the input data is fixed since many optimization algorithms converge faster with fixed numerical scale and exploding or vanishing gradients during backpropagation are reduced. Moreover, inputs normalized with this method are further aligned with the input distributions for which the pre trained models were trained on, yielding better performance.

3.2.4 Granularity Conversion

To evaluate model performance across varying levels of classification complexity, the dataset was reorganized into three levels of granularity:

1. **Five-class dataset:** This is the original dataset with 5 classes of knee osteoarthritis X-ray images.
2. **Three-class dataset:** To make the classification easier, and reduce class imbalance, we grouped severity levels into broader categories.

    - Class 0: Normal and Doubtful
    - Class 1: Mild and Moderate
    - Class 2: Severe

   The reorganization here maximized the clinical significance of groupings but retained interpretability.

3. **Two-class dataset:** For binary classification, we merged lower severity levels of a prescription (Normal and Doubtful) into one class (healthy), and higher severity levels of a prescription (Mild, Moderate, Severe) into another class (severe). An ideal approach for healthy vs. pathological case discrimination in a simplified framework.

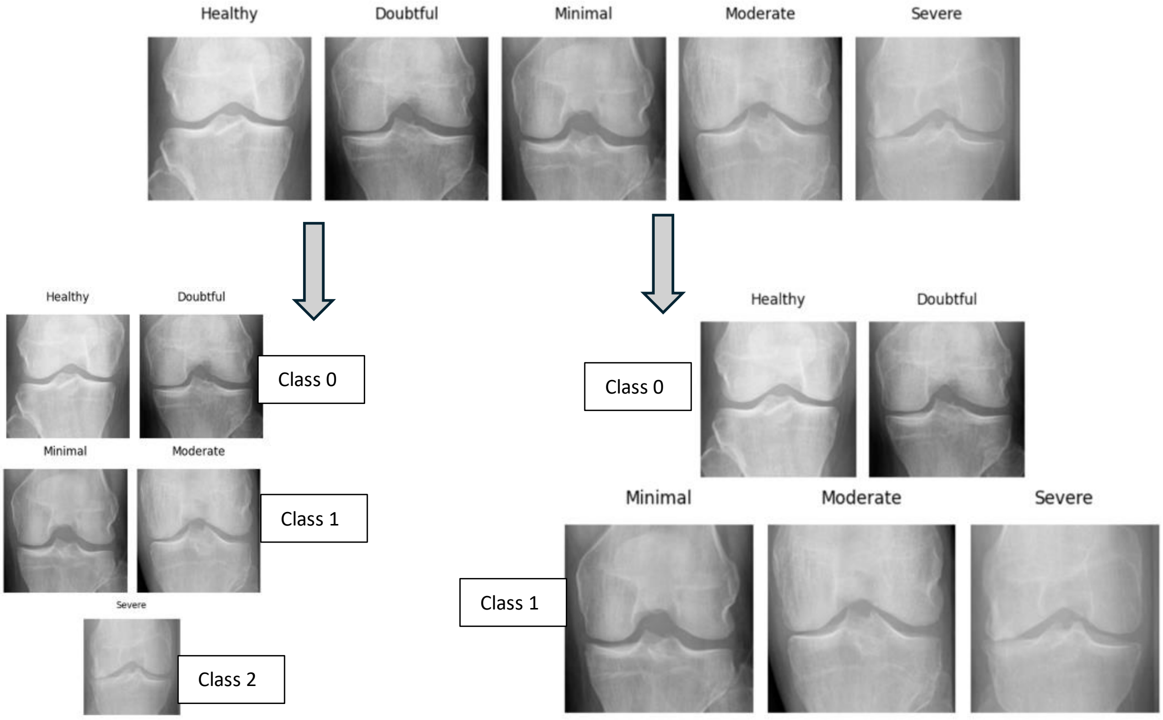


Figure 7: Granularity Conversion.

3.3 Modelling & Evaluation

The study evaluates two types of models for classifying knee osteoarthritis (KOA) severity: Traditional machine learning models and deep learning models. Following is a detailed architecture, strengths, and weaknesses of each model and the evaluation metrics.

3.3.1 Deep Learning Models

a. VGG16

VGG16 is the 16 layered sequential convolutional neural network. It has a very simple architecture with convolutional, pooling and fully connected layers and it's working with small 3x3 kernels. It is strong in fine grained feature extraction so is good for image classification. But it has an enormous number of parameters, which increase computational cost and memory usage. (Kumar & Sankar, 2024)

b. ResNet50

A 50-layer deep residual network, ResNet50, which uses skip connections to solve the vanishing gradient problem is also implemented. Training deeper networks without a performance degradation is made possible through this architecture. One of its strengths is having the ability to model complex features whilst remaining computationally efficient. The downside, however, is that it has many layers to tune. (Kumar & Sankar, 2024)

c. DenseNet121

DenseNet121 connects all layers with all other layers and therefore promotes the use of features and efficient gradient flow. This simplifies the computation to reduce complexity and a better point in data generalization on smaller dataset. Feature efficiency is its main strength, for which it may need more memory because of the dense connections. (Kumar & Sankar, 2024)

d. InceptionV3

The InceptionV3 uses a modular architecture to capture multi scale features by using filters of various shapes on the same layer. Computationally it is efficient and also effective to learn complex patterns. However, its architecture is complex and as a result harder to implement and tune than simpler models. (Kumar & Sankar, 2024)

e. EfficientNetB0

Compound scaling is used in EfficientNetB0 to trade distribution of parameters between network depth, width, and resolution, to achieve optimal performance. It achieves very high accuracy while requiring little computational resources thus it is perfect for constrained environments. However, it doesn’t perform as well when on large, highly complex datasets without proper tuning. (Pandey & Kumar, 2023)

f. Xception

To increase computational efficiency, and reduce parameters, Xception replaces standard convolutions with depth wise separable convolutions. It is best at dealing with large datasets. This is however less interpretable than simpler models.

g. MobileNet

MobileNet is a mobile and embedded system architecture that uses depth wise separable convolutions to minimize the computational overhead. Its strength is its low resource setting efficiency, but it may not work as good as heavier models on complex datasets.

3.3.2 Machine Learning Models

a. Random Forest

Random Forest is an ensemble model consisting of several decision trees and combined predictions of the built decision trees. It is very robust to overfitting and has good performances on the data corrupted by noise and unbalanced. Its merit, however, comes at a price: single decision trees are much more interpretable than a decision forest and may sometimes be computationally expensive.

b. Naïve Bayes

The probabilistic model Naive Bayes assumes features that are independent. It’s simple, fast, and very good with small datasets. However, it can restrict its performance since it assumes the features independent.

c. K-Nearest Neighbours

It is a non-parametric algorithm that classifies data points using the majority vote of their closest neighbours (k neighbours). For small datasets its simple and effective but when the dataset gets large or the data dimension is high, it becomes computationally expensive and less accurate. (Sakaria, Jain & Rana, 2023)

d. K-Means

The widely used clustering algorithm k-Means, partitions data into 'k' clusters, where they are clustered such that the total variance within each cluster is minimized. k-Means is iteratively and assigns each data point to the cluster of nearest cluster centroid, updated the centroids via the mean of assigned points. It is efficient and is found to do well on large datasets with well separated clusters. However, it has limitations: It assumes clusters to be spherical of same size, sensitive to initial centroid placement and requires the number of clusters ('k') to be specified beforehand which is not always straightforward. (Singh, Gupta & Choudhury, 2019).

3.3.3 Evaluation Metrics

The performance of classification model is evaluated by using Accuracy, F1-Score, and confusion matrix. Accuracy portrays a general performance of a model in percentage of right classified instances out of total predictions. It tends to be misleading in imbalanced datasets in the sense that some classes overwhelm others. This is tackled using the F1 Score, a harmonic mean of precision and recall, used when the trade-off between false positives and false negatives needs to be balanced, and can be especially useful in imbalanced classification cases. The confusion matrix allows you to understand the model’s performance better, by dividing predictions into how many predicted positives are true positives, true negatives, false positives, and false negatives, so that you can understand where such performance is poor. These metrics together give a complete evaluation about the accuracy of the model, how precise is it and its reliability.

4 Design Specification

The architectures used for knee osteoarthritis severity classification such as DenseNet121, EfficientNetB0 and Random Forest, are considered in this section. Preliminary evaluations were conducted to choose these models for their superior performance in hierarchical feature extraction, computational efficiency and in dealing with class imbalance with overfitting, of which the DenseNet121 and EfficientNetB0 both did quite well, while Random Forest also performed well. The performance over five class, three class and binary classification tasks is analysed to understand the strengths of these and to recommend when each is suitable for use.

4.1 DenseNet121

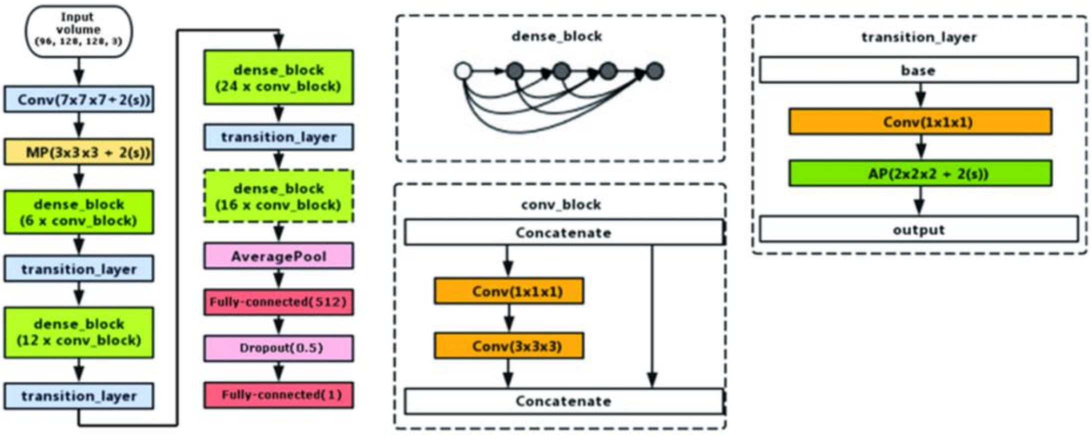


Figure 8: DenseNet121 Architecture (Solovyev et al., 2021).

DenseNet121 is a convolutional neural network of the dense connectivity family which connects all layers together. On contrary to summation of the outputs like residual networks DenseNet concatenates the output of preceding layers in form of feature maps to reuse the feature and provide efficient gradient flow.

The structure consists of several dense blocks, separated by transition layers. These are transition layers, and use down sampling operations, such as pooling layers, to reduce spatial dimensions. Each dense block incorporates:

- Batch Normalization (BN) for both improving training stability and accelerating training.
- Using ReLu activation for creating non linearity.
- 1x1 convolutions (bottleneck layers) that brought down feature dimensionality.
- 3x3 convolutions for spatial feature extraction.

The number of parameters is minimal, and the features are extracted with efficiency, thus rendering this design effective with small or imbalanced datasets.

4.2 EfficeientNetB0

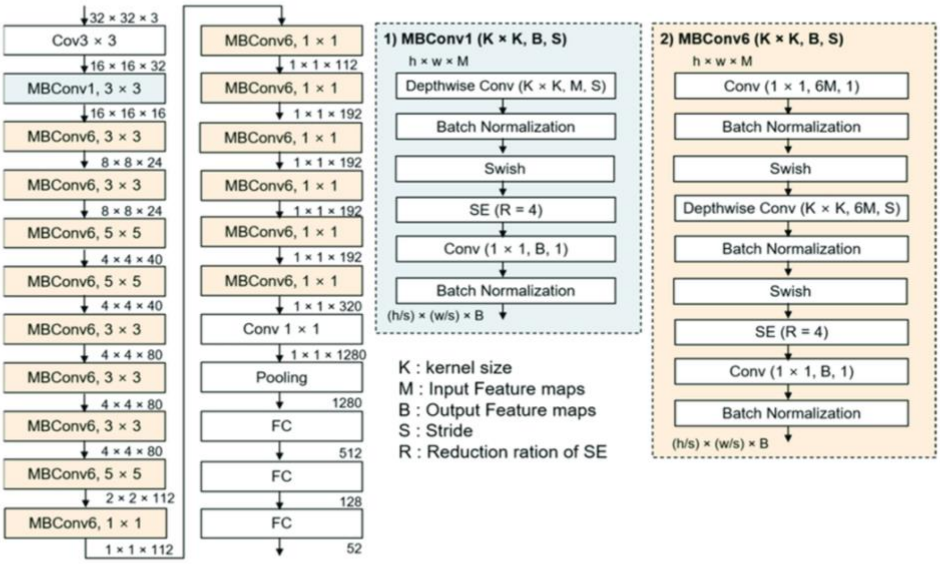


Figure 9: EfficentNetB0 Architecture (Gang et al., 2021).

EfficientNetB0 is a convolutional neural network optimized for good performance at the cost of minimal resources. Unlike traditional networks, EfficientNetB0 uses compound scaling to increase depth, width, and resolution collectively to maintain a good trade-off between effectiveness and computational cost.

Each architecture is structured as a sequence of MBConv blocks, as the backbone. Pooling and fully connected layers are interleaved among these blocks to obtain lower dimensional features on the spatial dimension. Each MBConv block incorporates:

- Depth wise convolutions to avoid computational complexity where they convolute independently on each channel.
- Feature dimensionality reduction (via pointwise convolutions (1x1 convolutions) that combine channel outputs in an effective way).
- Batch Normalization (BN) is for improving the training stability and training speed.
- Used the swish activation function to introduce nonlinearity and improve gradients flow.
- Utilizes Squeeze-and-Excitation (SE) blocks to recalibrate channel-wise feature importance and has the model specializing on the most essential regions.

Depth wise separable convolutions in EfficientNetB0 lower the number of parameters and SE blocks increase feature representation. Compared to the typical case, this design is computationally efficient and especially suited to complex data, as in medical X-rays. The algorithm's ability to generalize from one granularity level of classification to another demonstrates its robustness and it is suitable for both low level resource environment and high precision tasks.

4.3 Random Forest

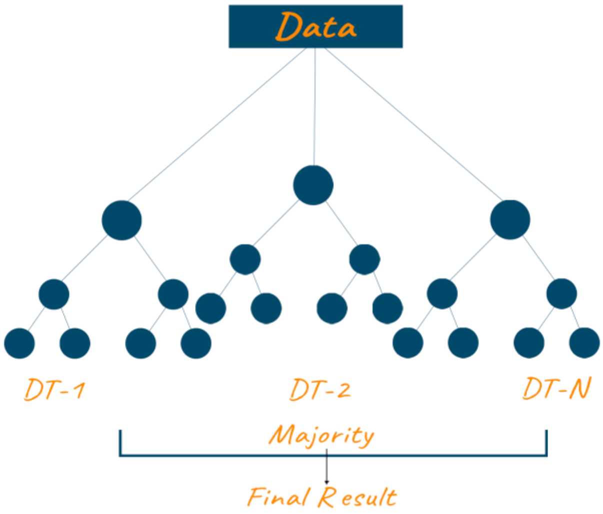


Figure 10: Random Forest Architecture

Random Forest is an ensemble learning algorithm that trains by creating multiple decision trees and produces a classification prediction by combining those trees outputs. A Random Forest is a collection of trees compared to a single decision tree that can be overfit on the training data.

The structure of Random Forest consists of the following key elements:

- Bootstrap Aggregation (Bagging): However, in each tree the training data is a random sample without replacement of the complete data. This prevents overfitting and ensure diversity between those trees.
- Random Feature Selection: Furthermore, in addition, only a random subset of features is considered for splitting at each node to further decorrelate the trees and improve generalization.
- Decision Trees: Each tree splits the data independently recursively in data based on the feature thresholds until it hits a leaf node and then independently predicting the class of given input.
- Majority Voting (Classification): For prediction, we predict by aggregating the predictions of all individual trees through majority voting for classification tasks.

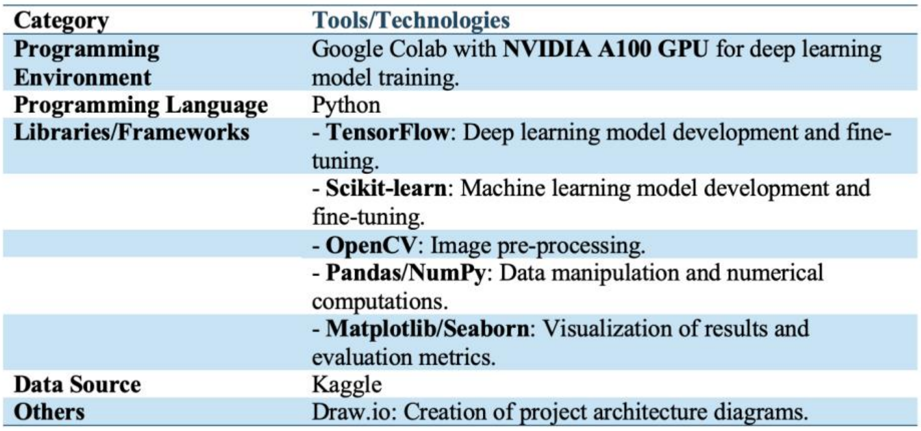
Random Forest is very interpretable and resists noise and class imbalance. But thanks to this ability to model non-linear relationships and interactions between features, it is still able to be effective on high dimensional datasets. However, its bagging approach also leads to a lower variance with a low bias and hence a lower tendency to generalize poorly on unseen data.

It proved to work well in binary classification tasks, with accuracy of 71.67%. The reason it is strong is because it can address imbalanced data and noisy features and can solve easily with simpler classification problems where there are limited computational requirements. Easily implementable and robust, Random Forest is a good choice for problems that require interpretability in efficiency.

5 Implementation

This section includes description of processes followed, tools and technologies applied, and outputs produced while executing this project. The implementation was divided into data preparation, model training, and evaluation stages, covering three classification setups: Severity classification into 5, 3 and 2 classes.

5.1 Tools and Technologies



Category	Tools/Technologies
Programming Environment	Google Colab with **NVIDIA A100 GPU** for deep learning model training.
Programming Language	Python
Libraries/Frameworks	- **TensorFlow**: Deep learning model development and fine-tuning.
	- **Scikit-learn**: Machine learning model development and fine-tuning.
	- **OpenCV**: Image pre-processing.
	- **Pandas/NumPy**: Data manipulation and numerical computations.
	- **Matplotlib/Seaborn**: Visualization of results and evaluation metrics.
Data Source	Kaggle
Others	Draw.io: Creation of project architecture diagrams.

Table 3: Tools and Technologies

5.2 Process

5.2.1 Data Preparation

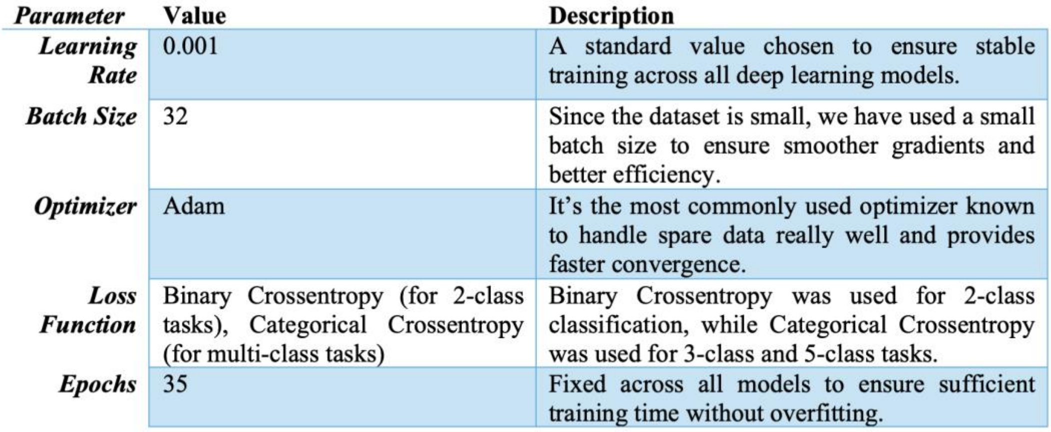
Knee X-ray images on Kaggle were classified among five severity levels namely: Healthy, Doubtful, Minimal, Moderate, Severe. To explore varying classification granularities, the dataset was reorganized into:

- **Five-Class Setup:** The original five severity levels for instruction are stated.
- **Three-Class Setup:** Gathered into Normal (Healthy), Mild (Doubtful, Minimal), and Severe (Moderate, Severe).
- **Two-Class Setup:** It will be the binary classification of Normal (Healthy, Doubtful) and Arthritis (Minimal, Moderate, Severe).

Images were converted to grayscale for emphasis on structural details in pre-processing and normalized in the pixel range [0,1]. To match the input dimensions that the pre-trained CNN architectures needed, each image has been resized to 224×224 pixels. Since these architectures take the input in the form of three channel images, grayscale images were expanded to three channels. We augmented the data with random flips, rotations, and zooms to improve generalization. Class imbalance was treated by calculating and using class weights during model training.

5.2.2 Model Training

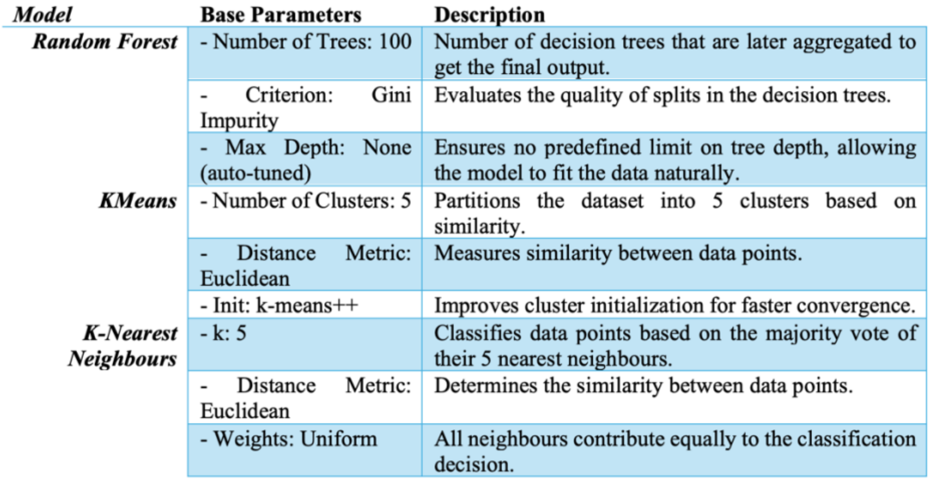
- **Deep Learning Models:** All classification scenarios were fine-tuned with seven pre-trained CNN architectures such as EfficientNetB0, DenseNet121, InceptionV3, Xception, MobileNet, ResNet50, and VGG16. Additional layers including Global Average Pooling and the last several fully connected layers were incorporated to each of the models. The final layer used:
    - Softmax layer for multi-class problems (Five-class and Three-class layouts).
    - Sigmoid activation for binary action: Two classification models.
    - For optimization, we used the Adam optimizer with learning rate scheduling performed using the ReduceLROnPlateau method.



Parameter	Value	Description
Learning Rate	0.001	A standard value chosen to ensure stable training across all deep learning models.
Batch Size	32	Since the dataset is small, we have used a small batch size to ensure smoother gradients and better efficiency.
Optimizer	Adam	It's the most commonly used optimizer known to handle spare data really well and provides faster convergence.
Loss Function	Binary Crossentropy (for 2-class tasks), Categorical Crossentropy (for multi-class tasks)	Binary Crossentropy was used for 2-class classification, while Categorical Crossentropy was used for 3-class and 5-class tasks.
Epochs	35	Fixed across all models to ensure sufficient training time without overfitting.

Table 4: CNN Parameters

- **Machine Learning Models:** Features describing the images were flattened and fed into classical artificial intelligence algorithms including Random Forest, KNN, Naïve Bayes and K-Means clustering. The hyperparameters of the algorithm were tuned using grid search with cross-validation.



Model	Base Parameters	Description
Random Forest	- Number of Trees: 100	Number of decision trees that are later aggregated to get the final output.
	- Criterion: Gini Impurity	Evaluates the quality of splits in the decision trees.
	- Max Depth: None (auto-tuned)	Ensures no predefined limit on tree depth, allowing the model to fit the data naturally.
KMeans	- Number of Clusters: 5	Partitions the dataset into 5 clusters based on similarity.
	- Distance Metric: Euclidean	Measures similarity between data points.
	- Init: k-means++	Improves cluster initialization for faster convergence.
K-Nearest Neighbours	- k: 5	Classifies data points based on the majority vote of their 5 nearest neighbours.
	- Distance Metric: Euclidean	Determines the similarity between data points.
	- Weights: Uniform	All neighbours contribute equally to the classification decision.

Table 5: ML Parameters

5.2.3 Model Evaluation

The performances of the models were assessed on the test data using the basic test measures including accuracy, F1-score, and confusion matrix. For the CNN models, the training and validation loss and accuracy plots were created for the analysis of trend and overfitting issues, and for confirmation of stable convergence during the training process.

6 Evaluation

This section evaluates the performance of the models across three classification granularities: Each task was 5-class, 3-class, and 2-class severity classification tasks. For each case study, we carry out a comparison of the models based on key metrics such as accuracy, loss trends and confusion matrices. An overall summary of the findings is presented in the final subsection.

6.1 Case Study 1: 5–Class Classification

For the 5-class classification task, EfficientNetB0 was found to be the best performing model with accuracy 67.33% and overcoming class imbalance, as well as smooth convergence. Although traditional models like K-Means and Random Forest performed poorly because of difficulties associated with working with high dimensional and imbalanced data, other deep learning models, such as InceptionV3, DenseNet121, and MobileNet presented competitive performance. However, VGG16 clearly performs poorly here and indeed fails to generalize due to overfitting. Deep learning models outperformed the traditional methods and, in general, demonstrated better performance in the task of complex image classification.

Model	Accuracy (%)	Insights
EfficientNetB0	**67.33**	Best performance; smooth convergence and effective handling of class imbalance.
InceptionV3	**64.79**	Multi-scale feature extraction contributed to strong performance and generalization.
DenseNet121	**64.01**	Feature reuse and gradient flow improved accuracy but slightly behind top performers.
MobileNet	**63.77**	Lightweight and efficient; competitive accuracy for a resource-efficient model.
Xception	**62.08**	Robust feature extraction; handled class imbalance effectively but lagged behind EfficientNetB0.
ResNet50	**60.81**	Decent performance; residual connections helped, but struggled with minority classes.
Random Forest	**42.82**	Best among traditional models; struggled with non-linear and imbalanced data.
K-Means	**38.59**	Heavily biased toward majority class; not suitable for supervised classification.
K-Nearest Neighbours	**34.60**	Poor performance due to sensitivity to high-dimensional data and class imbalance.
Naive Bayes	**22.06**	Unrealistic independence assumptions rendered it ineffective for image data.
VGG16	**3.08**	Failed to generalize; only predicted majority class due to overfitting.

Table 6: 5-Class Model Summary

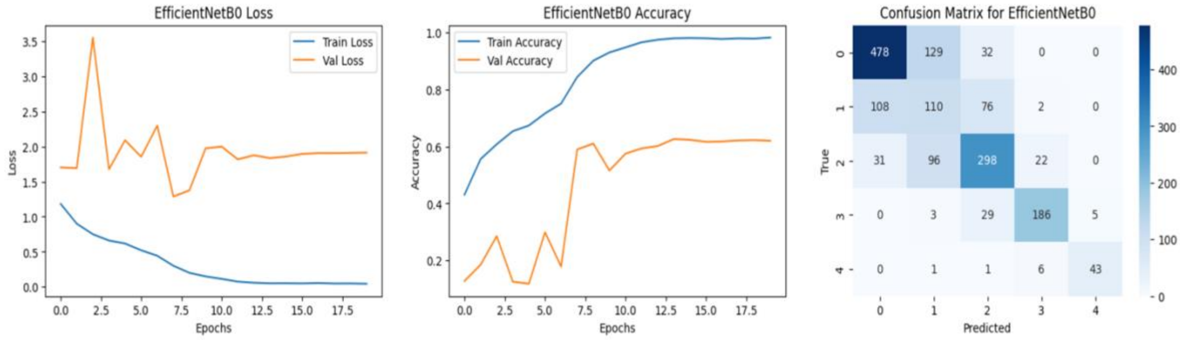


Figure 11: EfficientNetB0 5-Class Performance.

The trend of training and validation loss of EfficientNetB0 shows that it has a good learning ability, the training accuracy is continuously improving, but the validation accuracy has reached a maximum of 67% and stagnates, and there are problems with generalization. In the confusion matrix, we have presented a good accuracy of the majority class but problems estimating the minority class, as evidenced by the effect of class imbalance.

6.2 Case Study 2: 3–Class Classification

In the 3-class classification task, DenseNet121 showed the best performance with accuracy equal to 77.68 %, followed by InceptionV3 and EfficientNetB0 which both also demonstrated quite good performance. Traditional models were outperformed by deep learning models, wherein the best accuracy was achieved by Random Forest (65.95%). Classification performance improves across all the models when the classification task is simplified, but traditional methods still struggle with complex relationship and large dimension data.

Model	Accuracy (%)	Insights
DenseNet121	**77.68**	Best performance; feature reuse and gradient flow contributed to high accuracy and generalization.
ResNet50	**70.50**	Strong performance; residual connections helped achieve effective feature extraction.
InceptionV3	**75.02**	Robust multi-scale feature extraction enabled excellent performance in a reduced class setup.
EfficientNetB0	**74.14**	High accuracy; optimized architecture ensured efficiency and balance.
Xception	**73.75**	Consistent performance with advanced feature extraction capabilities.
MobileNet	**72.66**	Lightweight and efficient; good trade-off between accuracy and computational requirements.
VGG16	**65.88**	Significant improvement compared to 5-class classification, but still struggled with generalization.
Random Forest	**65.95**	Best among traditional models; limited ability to handle complex feature relationships.
K-Means	**65.88**	Comparable to Random Forest but limited due to its unsupervised nature and class bias.
K-Nearest Neighbours	**60.66**	Struggled with high-dimensional data, resulting in lower accuracy.
Naive Bayes	**33.33**	Performed poorly due to its unrealistic independence assumptions for image data.

Table 7: 3-Class Model Summary

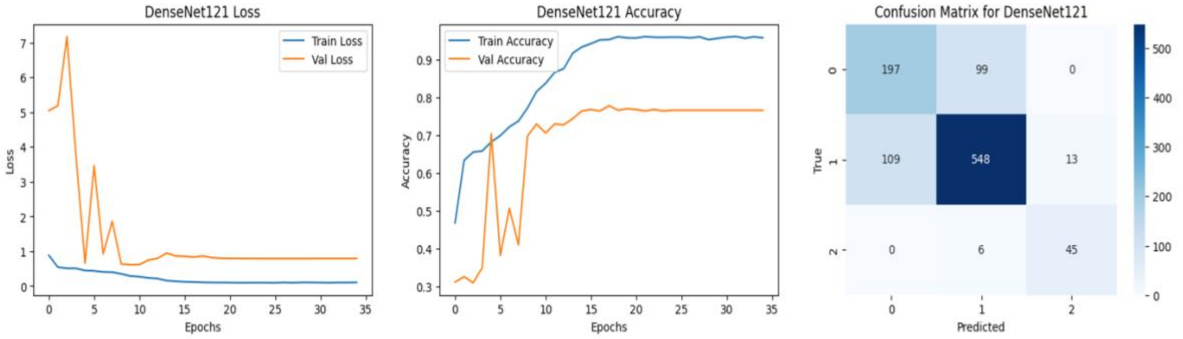
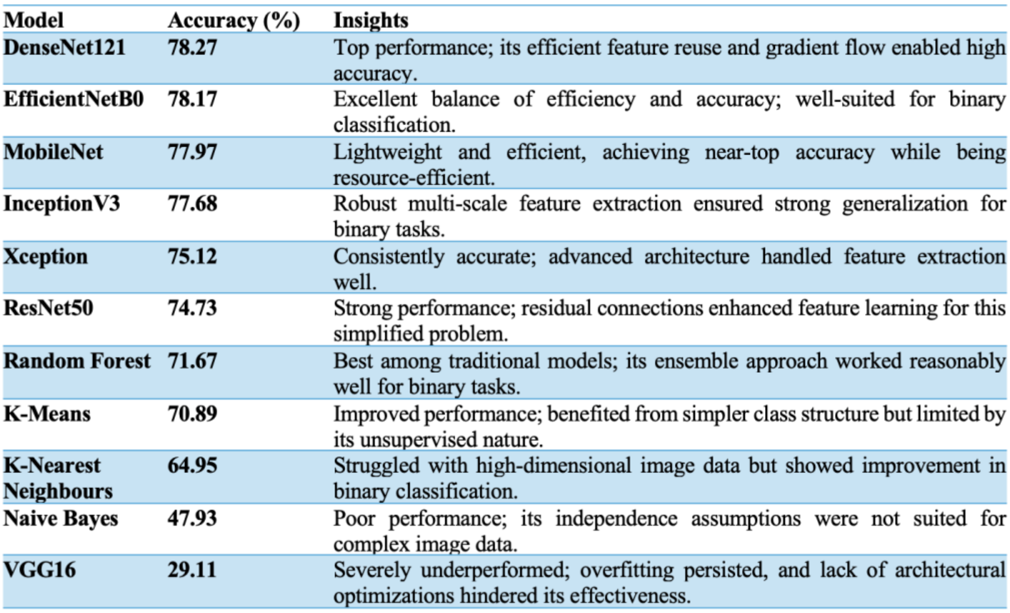


Figure 12: DenseNet121 3-Class Performance

The training performance of this model is good since its loss steadily decreases and accuracy increases. Validation loss fluctuates and accuracy seems to plateau, but it is clear there are possible overfitting or generalization issues. From the confusion matrix we can see that the performance is very uneven, a few classes (class 0 in this case) are easier to predict, but there are also big mistakes, especially between adjacent severity levels. This needs further optimization or balanced data handling.

6.3 Case Study 3: 2–Class Classification

DenseNet121 recorded the highest accuracy of 78.27% in the 2-class classification task, closely trailed by EfficientNetB0 at about 78.17% and tops the MobileNet at 77.97% which proves their efficiency and adaptability in binary tasks. Previous tasks showed deeper learning models outperforming traditional models, such as Random Forest (71.67%), which our own models also did. This task was significantly simplified, resulting in better performance across all models, while traditional methods such as K-Means and K-Nearest Neighbours hit the limits of their inability to deal with high dimensional data. The importance of modern architectural optimizations was confirmed by the fact VGG16 continued to underperform.



Model	Accuracy (%)	Insights
DenseNet121	**78.27**	Top performance; its efficient feature reuse and gradient flow enabled high accuracy.
EfficientNetB0	**78.17**	Excellent balance of efficiency and accuracy; well-suited for binary classification.
MobileNet	**77.97**	Lightweight and efficient, achieving near-top accuracy while being resource-efficient.
InceptionV3	**77.68**	Robust multi-scale feature extraction ensured strong generalization for binary tasks.
Xception	**75.12**	Consistently accurate; advanced architecture handled feature extraction well.
ResNet50	**74.73**	Strong performance; residual connections enhanced feature learning for this simplified problem.
Random Forest	**71.67**	Best among traditional models; its ensemble approach worked reasonably well for binary tasks.
K-Means	**70.89**	Improved performance; benefited from simpler class structure but limited by its unsupervised nature.
K-Nearest Neighbours	**64.95**	Struggled with high-dimensional image data but showed improvement in binary classification.
Naive Bayes	**47.93**	Poor performance; its independence assumptions were not suited for complex image data.
VGG16	**29.11**	Severely underperformed; overfitting persisted, and lack of architectural optimizations hindered its effectiveness.

Table 8: 2-Class Model Summary

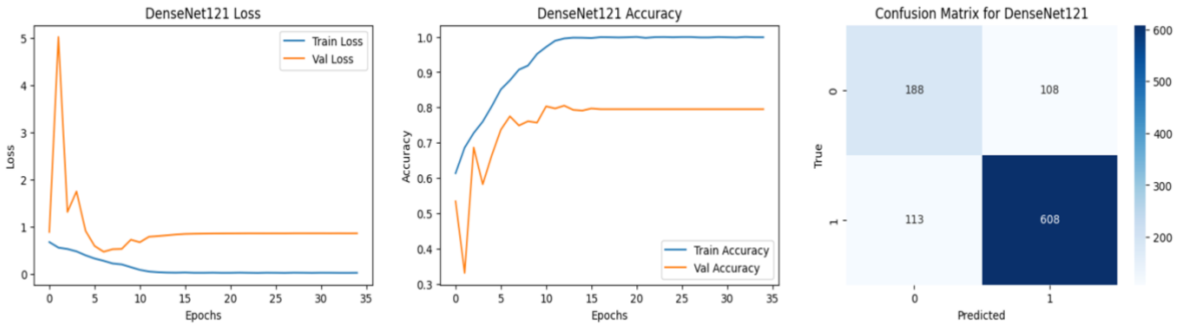


Figure 13: DenseNet121 2-Class Performance

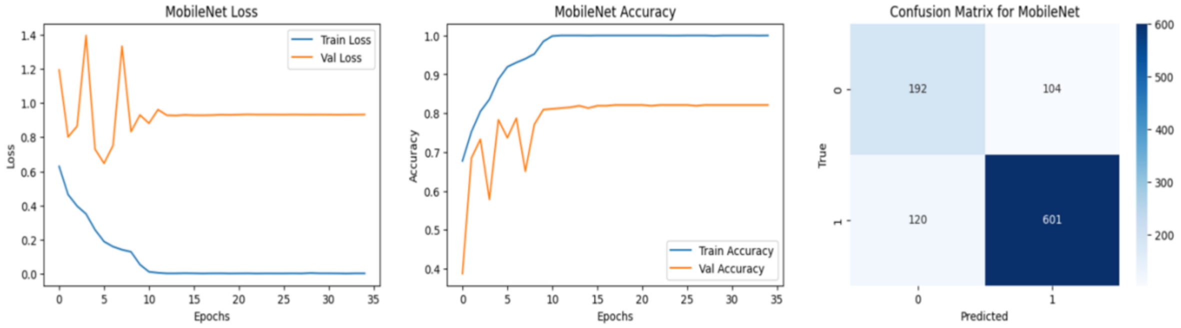


Figure 14: MobileNet 2-Class Performance

The training of the DenseNet121 and MobileNet models is effective as observed by steadily decreasing loss and increasing training accuracy while both shows overfitting, with validation loss oscillating or validation accuracy stays lower. The confusion matrices show stronger performance for class 0 and misclassifications especially in some other classes. With these trends, there is a demand to get better generalization using methods such as data augmentation or hyperparameter tuning.

6.4 Discussion

Results indicate that deep learning models perform better than traditional methods in all classification tasks, but with greater advantage as complexity of classes reduces. The top performers were DenseNet121 with Accuracy of 2-class classification 78.27%, EfficientNetB0 had Accuracy 78.17%, and MobileNet 77.97%. Advanced feature extraction and optimization techniques excelled these models in high dimensional image data by flattening each image down to a vector.

With fewer classes, the traditional machine learning models, like Random Forest and K-Means, also showed slight improvements, but they perform much worse than deep learning models. Of all traditional techniques, Random Forest attained the best accuracy of 71.67% in binary classification. However, models like Naive Bayes and KNN consistently struggled and Naive Bayes especially was unsuitable for image data as its independence assumptions.

Although a deep learning model, VGG16 achieved only 29.11% accuracy across all the tasks. The lack in modern optimizations such as residual connections make it outdated architecture that is less effective. Overall, we find that modern architectures like DenseNet121 and EfficientNetB0 are considerably superior to image classification, particularly for complex high dimensional data.

7 Conclusion and Future Work

Finally, based on all the classification granularities DenseNet121, EfficientNetB0, and MobileNet outperformed traditional machine learning models. With their advanced architectures extracting hierarchical and complex features from high dimensional image data, a clear advantage was exhibited. In the two-class classification task, DenseNet121 performed best (78.27%), utilizing efficient feature reuse, optimized gradient flow and redundancy reduction. In this study they show why the DenseNet family of CNNs has been a key landmark in medical image classification. The model performance increased considerably from classification tasks of five classes to two classes, as the classification problem was simplified by decreasing class overlap and misclassification, resulting in more distinct and generalizable features being identifiable by the models and reducing noise and class imbalance impact.

Binary classification tasks like Random Forest, K-Means performed well in traditional machine learning setup but could not better the performance of deep learning models, especially when working with high dimensional data. Their dependence on handcrafted features and shallow representations makes it incapable to characterize intricate patterns present in medical images. In contrast, by automatically learning task specific features via multiple levels of abstraction, deep learning models excel and are therefore the method of choice for complex images in medical diagnostics. The potential for deep learning architectures to revolutionize medical imaging technology is highlighted in this study.

To build on the findings of this study, the following areas of improvement and further exploration are proposed:

- **Incorporation of Clinical Metadata:** Integrate patient charts for demographics and medical history, into our system along with diagnostic data, from laboratory results for example, and use predictive analytics to improve diagnostic accuracy.
- **External Validation:** Their robustness is tested on external datasets for real world applications.

- **Hybrid Models:** Blend deep learning feature extraction with traditional classifiers such as Random Forests or SVM to take best of both worlds approaches.
- **Dataset Augmentation:** Generating images with Generative models such as GAN's and solving class imbalance problem.
- **Real-Time Deployment:** We also optimize a lightweight architecture (e.g. MobileNet) to run on edge devices to perform real time KOA classification.
- **Multi-Modal Analysis:** Xray images are combined with other imaging modalities (MRI or CT scan) for a complete view into KOA progression and severity.

These areas deserve addressing because they will advance automated KOA severity classification for earlier detection and decision making.

References

Divyanshi and S. Bansal, "Detecting Severity Level of Knee Osteoarthritis using CNN," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-5, doi: 10.1109/ICCCNT61001.2024.10724941.

C. Mukesh Kumar and J. Sree Sankar, "Comparative Analysis of Convolutional Neural Networks for Brain Tumor Detection: A Study of VGG16, ResNet, Inception, and DenseNet Models," 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2024, pp. 41-46, doi: 10.1109/ICAAIC60222.2024.10575770.

V. Kant, "CNN-Based Classification of Knee Arthritis Severity from X-ray Images," 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2024, pp. 1248-1252, doi: 10.1109/ICOSEC61587.2024.10722635

G. M. Vali Shaik, B. Zhou and Q. Liu, "A Comparison Study in Detecting Knee Osteoarthritis Severity with Deep Leaning," 2024 20th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guangzhou, China, 2024, pp. 1-6, doi: 10.1109/ICNC-FSKD64080.2024.10702300.

G. Kitukale, N. A. Shelke, R. Agrawal, N. P. Singh and S. Quamara, "Predicting Knee Osteoarthritis using Deep Neural Network," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp. 1-6, doi: 10.1109/I2CT61223.2024.10543978.

K. S. Gill, V. Anand and R. Gupta, "Classification of Pneumonia Disease Through Deep Learning Procedures and Fine-tuning on ResNet50V2 CNN Model Utilizing Chest Xray Images," 2023 Global Conference on Information Technologies and Communications (GCITC), Bangalore, India, 2023, pp. 1-5, doi: 10.1109/GCITC60406.2023.10426406.

S. D. Chandu, P. Revathi and N. A. S. Vinoth, "Discovering Knee Osteoarthritis Using CNN Enhanced with AlexNet," 2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2024, pp. 134-141, doi: 10.1109/ICICV62344.2024.00028.

A. Pandey and V. Kumar, "Enhancing Knee Osteoarthritis Severity Classification using Improved Efficientnet," 2023 10th IEEE Uttar Pradesh Section International Conference on

Electrical, Electronics and Computer Engineering (UPCON), Gautam Buddha Nagar, India, 2023, pp. 1351-1356, doi: 10.1109/UPCON59197.2023.10434740.

S. Sakaria, S. Jain and M. K. Rana, "Rheumatoid Arthritis Predictor Using ML Techniques and Explainable AI," 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballar, India, 2023, pp. 1-7, doi: 10.1109/ICDCECE57866.2023.10150759.

U. V. Singh, E. Gupta and T. Choudhury, "Detection of Rheumatoid Arthritis Using Machine Learning," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2019, pp. 25-29, doi: 10.1109/ICCIKE47802.2019.9004234.

S. Sundaramurthy, S. C and P. Kshirsagar, "Prediction and Classification of Rheumatoid Arthritis using Ensemble Machine Learning Approaches," 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2020, pp. 17-21, doi: 10.1109/DASA51403.2020.9317253.

A. S. Chandra Bose, S. C and I. J. S, "Machine Learning based Detection and Severity Classification of Knee Osteoarthritis," 2023 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICERCS57948.2023.10434162.

K. S. Gill, V. Anand and R. Gupta, "Kidney Disease Classification Using Machine Learning Approach on DenseNet201 Model using Xray Images," 2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI), Tiruchengode, India, 2023, pp. 1-4, doi: 10.1109/ICAEECI58247.2023.10370846.

A. A. Romalt and R. M. S. Kumar, "Prediction of Cardio Vascular Disease by Deep Learning and Machine Learning-A Combined Data Science Approach," 2022 International Conference on Computer, Power and Communications (ICCPC), Chennai, India, 2022, pp. 83-85, doi: 10.1109/ICCPC55978.2022.10072141.

H. P. Nguyen, T. P. Hoang and H. H. Nguyen, "A deep learning based fracture detection in arm bone X-ray images," 2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), Hanoi, Vietnam, 2021, pp. 1-6, doi: 10.1109/MAPR53640.2021.9585292.

V. Bhagat and S. Bhaumik, "Data Augmentation using Generative Adversarial Networks for Pneumonia classification in chest Xrays," 2019 Fifth International Conference on Image Information Processing (ICIIP), Shimla, India, 2019, pp. 574-579, doi: 10.1109/ICIIP47207.2019.8985892.

G. Ma, J. Lin, Z. Li and Z. Zhao, "Image Restoration Method based on Partition and Regularization for Industrial X-ray images," 2020 International Symposium on Autonomous Systems (ISAS), Guangzhou, China, 2020, pp. 254-257, doi: 10.1109/ISAS49493.2020.9378856.

Solovyev, Roman & Kalinin, Alexandr & Gabruseva, Tatiana. (2021). 3D Convolutional Neural Networks for Stalled Brain Capillary Detection. 10.48550/arXiv.2104.01687.

Gang, Sumyung & Fabrice, Ndayishimiye & Chung, Daewon & Lee, Joonjae. (2021). Character Recognition of Components Mounted on Printed Circuit Board Using Deep Learning. Sensors. 21. 2921. 10.3390/s21092921.