Science Eve Publisher

Detecting Knee Osteoarthritis using X-ray Images: An Image Processing and Deep Learning Approach

*Varsha Sharma 1, Sudheer Kumar,E.1**

1 School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Tamilnadu, India.

Corresponding author: sudheerkumar.e@vit.ac.in*

ABSTRACT

Knee osteoarthritis (OA) is a degenerative joint disease that progressively damages the articular cartilage and affects mobility and quality of life. Early and accurate detection of knee OA is essential for effective clinical management and treatment planning. In this work, deep learning and medical image processing techniques were employed for automated knee OA classification using radiographic X-ray images obtained from the Mendeley Data VI dataset available on Kaggle. Image preprocessing techniques including Contrast Limited Adaptive Histogram Equalization (CLAHE), image resizing, cropping, and data augmentation were applied to enhance the quality of the radiographs and improve model generalization. Transfer learning was implemented using pre-trained convolutional neural network models including ResNet50, ResNet152, DenseNet121, MobileNetV2, and InceptionV3. Experiments were conducted on both five-class and three-class classification settings based on Kellgren–Lawrence (KL) severity grades. Among all the evaluated models, InceptionV3 achieved the best performance with testing accuracies of 67.15% for the five-class dataset and 76.33% for the three-class dataset. The results indicate that grouping clinically related OA severity levels improves classification effectiveness and model robustness. The proposed framework demonstrates the potential of deep learning-assisted diagnosis for supporting radiologists and clinicians in the early detection and severity assessment of knee osteoarthritis using X-ray imaging.

Keywords

CLAHE, Deep Learning, Image Processing, Knee OA, Pre-trained models, Transfer Learning, X-ray

1 Introduction

Knee osteoarthritis is a progressive disease that is mainly caused due to the wear and tear of the articular cartilage of the knees. Articular cartilage is a tissue that is present at the bony surface of the joints and provides a low-friction surface that helps the joint to bear weight and carry day-to-day activities such as walking, stair-climbing, and other work-related activities. This leads to the bones in the knee joints rubbing together due to reduced friction. The affected knee joint gets swollen, becomes stiff, and might also result in limited joint movements. This disease is mostly observed in the elderly people aged above 40. Some studies have shown that women are more likely to develop knee osteoarthritis as compared to men.

Knee osteoarthritis can be detected by studying the radiographic images of the knee joint. The severity of knee osteoarthritis can be determined by grading the X-ray images on a scale from 0 to 4 (both inclusive). This severity score is called the Kellgren and Lawrence's score (KL score). Kellgren and Lawrence's method is a semiquantitative approach that evaluates osteophytes and joint space narrowing in the knee joint and assigns a score between 0 (no knee OA) to 4 (severe knee OA).

Recently, digital image processing has played a pivotal role in detecting the presence of diseases in imaging acquired from patients using various imaging techniques. Computer vision and deep learning techniques can be used to carry out these medical image analysis tasks. Image enhancement, and image segmentation algorithms are used to enhance the images and extract useful features. These enhanced images and extracted features can then be fed into deep learning-based convolutional neural networks for classification tasks. In some cases, transfer learning can also be used where certain pre-trained models are implemented on a new problem. Although several deep learning and transfer learning–based approaches have been proposed for knee osteoarthritis detection, many existing studies suffer from certain limitations. Several methods rely heavily on complex architectures with high computational cost, making them difficult to deploy in practical clinical settings. In addition, some studies primarily focus on binary or limited classification tasks and do not adequately evaluate performance across multiple severity grades of osteoarthritis. The variability in radiographic quality, overlapping visual characteristics among adjacent Kellgren–Lawrence grades, and limited preprocessing strategies further affect the robustness and generalization capability of the models. Moreover, comparative analysis of multiple transfer learning models under consistent preprocessing and classification settings remains limited in the existing literature. To address these challenges, the present study investigates the effectiveness of multiple pre-trained deep learning models using enhanced preprocessing techniques and evaluates their performance on both five-class and clinically grouped three-class knee OA datasets.

Thus in this research work, the contribution are listed as follows

- Utilized publicly available knee X-ray images from Kaggle, sourced from the Mendeley VI dataset, ensuring reproducibility and benchmarking against standard datasets.

- Applied contrast enhancement–based preprocessing, specifically Contrast Limited Adaptive Histogram Equalization (CLAHE), to improve visual quality and highlight diagnostically relevant features in knee X-ray images.

- Designed and implemented transfer learning–based deep learning models to effectively address the knee osteoarthritis classification problem with limited medical imaging data.

- Evaluated the proposed models on the original five-class classification setting, reflecting varying severity levels present in the dataset.

- Constructed a reduced three-class dataset by intelligently grouping related classes to analyze model robustness and performance under different clinical categorization schemes.

• Performed a comparative performance analysis of the deep learning models across both five-class and three-class datasets to assess classification effectiveness and generalization capability.

This paper has been organized into four sections. Section 2 presents a study of the related works that have been carried out to detect knee osteoarthritis by making use of X-ray images and deep learning techniques. The methodology used to carry out the intended study has been explained in detail in Section 3. Section 4 discusses and analyses the results obtained. The conclusion is covered in Section 5

2 Related Work

Authors Brahim, et al., [1] carried out filtering in the Fourier domain to extract the RoI. The images have been enhanced using histogram equalization and the intensities have been normalized using multiple linear regression (MLR). The important features have been extracted using Independent Component Analysis (ICA). Finally, Random Forest and Naive Bayes algorithms with Leave-one out (LOO) cross-validation strategy have been used to classify the X-ray image scans obtained from the Osteoarthritis Initiative (OAI) dataset. Their proposed algorithm has achieved an accuracy of 82.98±2.12. In another work, [2] collected the X-ray images of patients of 50 years of age from the radiological centre (KGS scan center, Madurai). They have used Faster RCNN and modified ResNet50 for extracting RoI from the images. AlexNet has been implemented for OA severity classification. They have achieved an accuracy of 98.516% and 98.90% for detecting and classifying the disease respectively. Deep learning models such as VGG, GoogLeNet, ResNet, DenseNet, ResNeXt, and MobileNet v2 have been implemented by [3] for OA classification. They used the ordinal regression module before predicting the final class for a given image. They have trained their models using standard hyperparameters and the ImageNet weights. [4] have used the OIA dataset and implemented the CLAHE algorithm to improve the contrast of the radiographs. The dataset was partitioned into 70% training, 10% validation, and 10% test sets using the hold-out strategy. CNN is used for extracting features.

The first set of proposed models (DHL-I) uses SVM to classify features in one of the five classes after extracting features from CNN. The second model (DHL-II) uses transfer learning to create two labels, three labels, four labels, or five labels classification models. For the proposed DHL-I model an accuracy of 74.57%, specificity of 92.58%, and recall of 75.4% is achieved. [5] have split the dataset in the fraction of 7:1:2 for train, validation, and test respectively. YOLOv2 has been used for detection and ResNet, VGG, DenseNet, Inception has been used for classification. The classification models are used with ordinal loss. In another approach, [6] have split the OIA dataset into the 65:20:15 ratio. The U-Net model has been used for knee joint localization and ensemble learning of DenseNet models has been used for classification. The various performance metrics obtained are sensitivity 83.7, 70.2, 68.9, and 86.0% and Specificity 86.1, 83.8, 97.1, and 99.1% for the four classes starting from no OA to severe OA. [7] have split each image in the middle to produce right and left knee images. DenseNet has been used for the classification of the images into one of the KL score grade classes.

On the full test set, an accuracy of 87.2%, F1 score of 86.6%, Precision of 88.4%, and recall of 84.9% is achieved. [8] have used CLAHE for image enhancement. From both the left and right knee, 400x100 pixels at the center of the knee joint were cropped to get RoI. For finding out the KL grade, DCNN has been used. They have obtained an accuracy of 77.24%. [9] have applied the ResNet model on the OIA dataset for knee osteoarthritis detection using X-ray images. They achieved an AUC of 0.93, 0.80, 0.88, 0.96, 0.99 for KL Grade 0, 1, 2, 3 and 4 respectively. For the detection of knee osteoarthritis, [10] have used a 2D median filter of kernel size 2x2 for denoising the X-ray images collected from Kaggle. In addition to this, they have also used the Adaptive histogram equalization technique clipping the histogram at a value of 2.0 and using the tile grids of size 8x8. Image rotation and flipping techniques have been applied for data augmentation. Stratified sampling is used to create a balanced dataset. They have used CNN, VGG16, and ResNet50 for feature extraction and classification. They have obtained an accuracy of 92.17%.

It has been stated that pre-trained models might not be fully adaptable and might require finetuning. Hence, ensemble learning and reinforcement learning can be used to make the models more efficient. [11] have used the Mendeley dataset IV which grades images based on the severity of the disease. They have found out the elemental image properties using the HOG technique. The high-level features have been using the Local Binary Pattern descriptor. They have used SVM, Random Forest, and KNN for classification. An accuracy of 98% has been achieved for five-fold cross validation implementing the KNN algorithm and a feature vector comprising combined features from CNN and HOG. The CNN feature vector was used with SVM to get an accuracy of 97.6%. On the other hand, the features extracted from CNN were used in the Random Forest classifier to get an accuracy of 93.09%. [12] have implemented the Elastic Net Regression, Random Forest, Linear mixed effect model (LMM), CNN on the OIA dataset. The RMSE obtained for the different severity levels of KOA as defined in the dataset is 0.973, 0.978, 0.943, and 0.770 for Elastic Net Regression, LMM, Random Forest Regression, and CNN Regression respectively. The negative of the image to enhance the visibility of the RoI in the image was implemented in [14]. The images are downscaled in each dimension using the bilinear method. The features are extracted using HOG. To detect KOA, joint space width (JSW) is calculated. A threshold is set for JSWs of medial and lateral sides (dm and dl, respectively). If a value falls between dm and dl, the person is diagnosed with osteoarthritis, otherwise, no. The performance metrics obtained are accuracy 97.14%, sensitivity 99.19%, F1 score 0.9840, and Cohen's Kappa coefficient 0.8507. DenseNet-121, DenseNet-161, ResNet-34, VGG19 models on the OIA dataset was implemented in [14]. They have also used the ensemble of these base classifiers after joining them with a fully connected layer. The ADAM optimiser has been used in the FC layer with an initial learning rate of 0.0001. They have achieved an accuracy of 0.96, 0.97, 0.95, 0.96, 0.98; precision of 0.97, 0.97, 0.96, 0.96, 0.98 for models DenseNet-121, DenseNet-161, ResNet34, VGG-19, Ensemble respectively. The authors of [13] has cropped the images from top and bottom by 60 pixels to get the segmented images. Histogram equalisation has been used for image enhancement. Certain models such as ResNet101, InceptionResNetV2, VGG16, VGG19, DenseNet121, and MobileNetV2 have been implemented for KLgrade classification of the X-ray images. It was observed that ResNet101 gave the highest testing accuracy of 69%. [16] have used the BoneFinder software to find out the reference points along the edges of the kneecap. Local Binary Pattern (LBP) has been used as the texture descriptor. To detect the patellofemoral osteoarthritis, a Gradient Boosting Machine (GBM) classifier and, deep CNN has been used. Several models on different attributes and different model structures were developed. The best AUC score of 0.937 was obtained by a stacked model.

We have used certain pre-trained models such as ResNet152, ResNet50, DenseNet121, MobileNetV2, and InceptionV3. The performance of the base-line models from the literature was compared with our best-performing model (i.e., InceptionV3). It was found that our proposed model gave higher accuracy when compared with the baseline models from the related works.

From the reviewed literature, it can be observed that existing knee osteoarthritis detection techniques can broadly be categorized into traditional machine learning approaches, CNN-based deep learning models, and transfer learning–based classification frameworks. Traditional machine learning methods rely heavily on handcrafted feature extraction techniques such as HOG, LBP, ICA, and texture descriptors, which often require extensive domain expertise and may not generalize effectively across diverse datasets. Deep learning approaches have demonstrated improved feature learning capability; however, several studies focus primarily on limited classification settings, binary severity prediction, or specific network architectures without comprehensive comparative analysis.

Furthermore, many existing methods suffer from challenges such as class imbalance among KL grades, overlapping radiographic characteristics between adjacent severity classes, limited preprocessing strategies, and reduced model generalization capability across varying image conditions. Several studies also employ computationally complex architectures that may not be suitable for practical clinical deployment. In addition, only limited works investigate the impact of clinically meaningful class grouping strategies on classification performance.

Motivated by these limitations, the present study implements a unified transfer learning framework using multiple pre-trained deep learning models combined with CLAHE-based image enhancement and augmentation techniques. The proposed work further evaluates the models on both the original five-class dataset and a clinically grouped three-class dataset to analyze robustness, classification effectiveness, and generalization performance under different severity categorization schemes. Table 1 shows the consolidated literature survey of the above discussed research works.

Table 1: Consolidated Literature Survey

Ref	Method	Dataset	Strength	Limitation
[1]	ICA + RF	OAI	Good handcrafted features	Limited generalization
[5]	YOLOv2 + CNN	OAI	Strong detection	High computational cost
[7]	DenseNet	OAI	High accuracy	Sensitive to class imbalance
[10]	Transfer Learning	Kaggle	Better automation	Limited multi-class robustness
Proposed	InceptionV3 + CLAHE	Mendeley VI	Better multi-class evaluation	Requires further clinical validation

3 Proposed Work

The methodology of the proposed work can be broadly classified into three stages. Subsection 3.1 describes the dataset used for the problem statement. The preprocessing carried out has been mentioned in subsection 3.2. Subsection 3.3 mentions the various deep-learning models and the other associated techniques that have been implemented for classification purposes.

The Figure 1 shows the conceptual framework for detecting knee osteoarthritis using radiographic images.

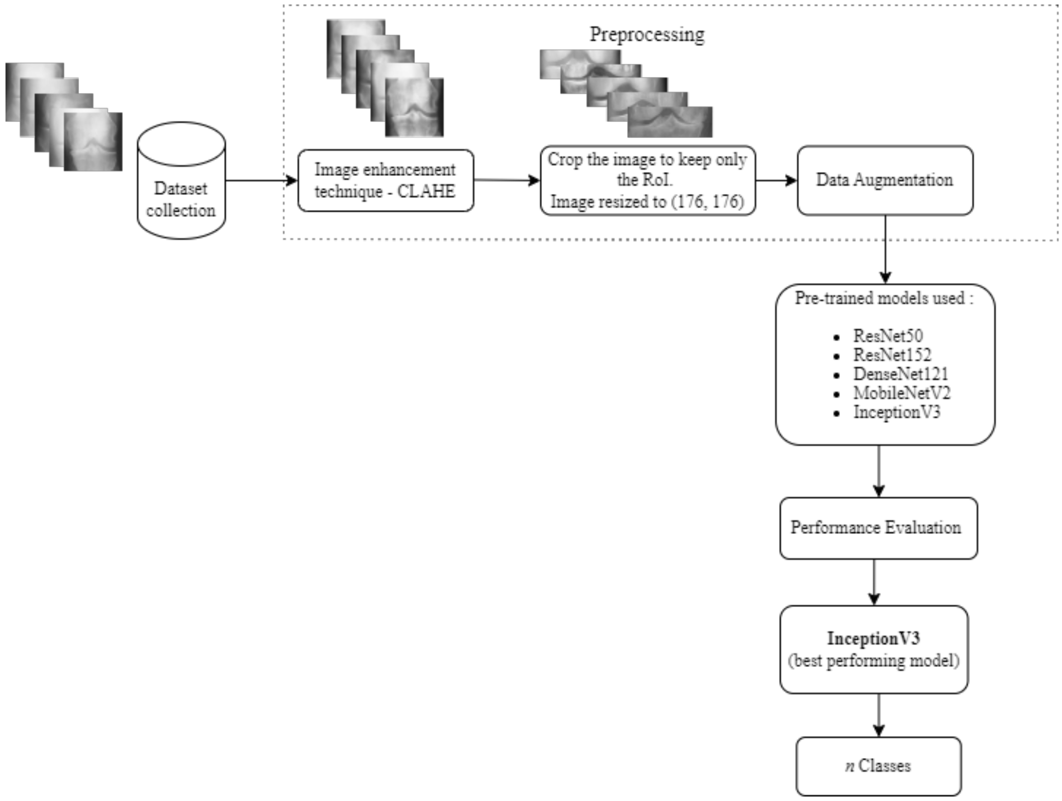


Figure 1: Proposed Architecture

3.1 Dataset Description

The proposed study uses an image dataset collected from Kaggle [17]. The dataset is the Mendeley Data, VI [18]. The dataset consists of radiographs of the knee joint belonging to 5 classes, from 0 to 4, where each class represents the KL severity score of the knee osteoarthritis. The severity of the disease increases from 0 to 4, where 0 represents a normal knee and 4 represents a knee highly affected by osteoarthritis. The data has already been partitioned into training, validation, and testing data with a percentage of 70, 10, and 20% of the data in each section respectively. Fig 2 shows the example images from the five classes of the dataset representing the Kellgren-Lawrence score from 0 to 4 based on the severity of knee osteoarthritis.

3.2 Preprocessing

The knee X-ray images collected are in the .png format. The medical images often contain some type of noise and artifacts. The presence of noise might lead to blurring of subtle features and artifacts might cause false diagnosis. Thus, image enhancement techniques can be implemented that improve the overall quality of the image and thus lead to a more accurate diagnosis of the disease. To enhance the contrast of the radiographs, we have implemented the CLAHE (Contrast Limited Adaptive Histogram Equalization) method. This algorithm divides a given image into smaller sections, also called the “tiles” and operates on it to improve the overall contrast of the image. It also uses a “clip limit” which clips the histogram at a predefined value before computing the cumulative distribution function therefore restricting the amplification of the contrast in the image. For our purpose, the cutoff value for the histogram has been set to 2 and a tile grid of dimension 6x6 has been used.

The CLAHE parameters were selected based on both empirical observations and findings reported in earlier medical image enhancement studies. Different combinations of clip-limit values and tile-grid dimensions were experimentally evaluated during the preprocessing stage to analyze their impact on image contrast enhancement and preservation of diagnostically relevant knee joint structures. Lower clip-limit values produced insufficient contrast enhancement, whereas higher values introduced excessive noise amplification and artifacts in the radiographs. Similarly, smaller tile-grid sizes resulted in localized over-enhancement, while larger grid sizes reduced the visibility of subtle osteoarthritic features. Based on these observations, a clip-limit value of 2 and a tile-grid size of 6×6 were chosen as they provided a balanced enhancement of the knee joint region while preserving structural details important for classification. Similar CLAHE-based enhancement strategies have also been adopted in previous knee osteoarthritis imaging studies [4], [10], [15]. Table 2 shows the effect of CLAHE parameters on validation accuracy.

Table 2: Effect of CLAHE Parameters on Validation Accuracy

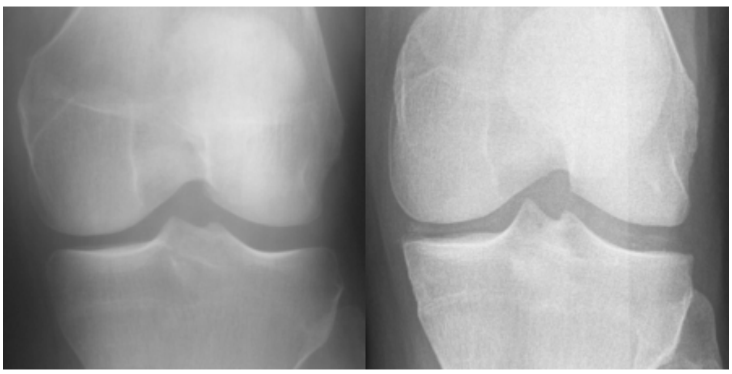
Clip Limit	Tile Grid Size	Validation Accuracy (%)
1	4×4	59.8
2	6×6	62.2
3	8×8	60.7

Fig 3 shows two sample images from class 3. The image on the left is the

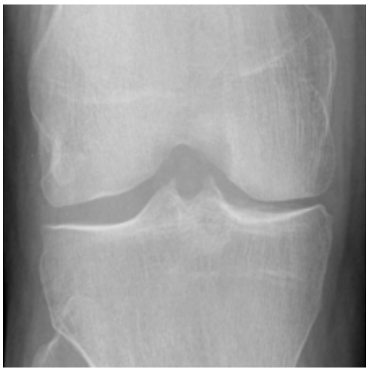
original image from the dataset while the image on the right is the enhanced image obtained after applying CLAHE.

After applying CLAHE, the images were cropped so that only the knee joint was visible and the other unnecessary parts of the X-ray image that are not needed for the diagnosis were removed. The images have been resized to (176, 176) and cropped such that the dimensional range of the image is 70 to 130 along the x-axis and, 10 to 180 along the y-axis.

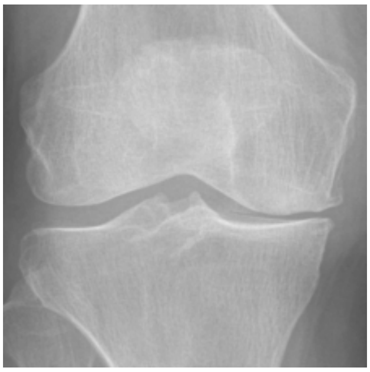
The image of a knee X-ray is shown at the left in fig 4.



(a) Class 0 (b) Class 1



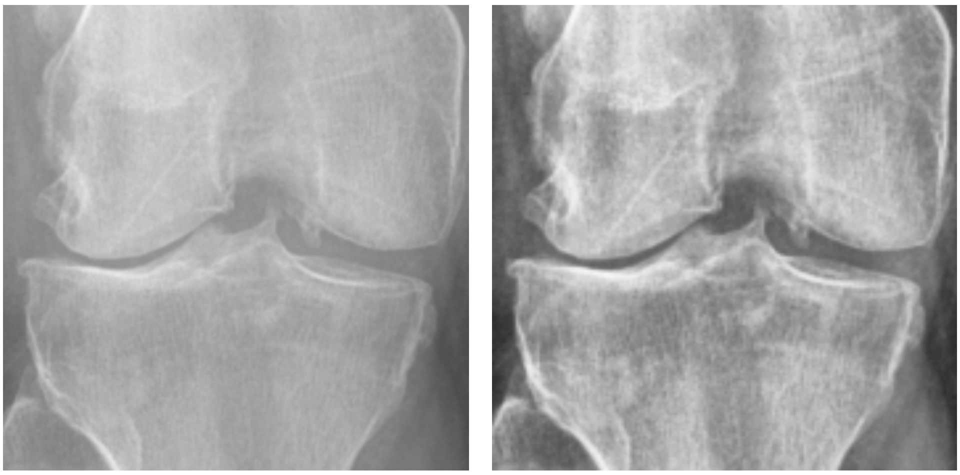
(c) Class 2



(d) Class 3

(e) Class 4

Figure 2: Dataset containing X-ray images representing the 5 classes of knee OA



(a) Original (b) Enhanced

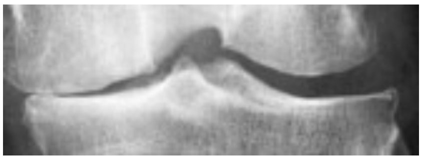
Figure 3: Sample X-ray image from class 3 before and after applying CLAHE for image enhancement

The image at the right represents the cropped image that contains only the knee joint that is to be studied for the detection of knee OA. Data augmentation techniques have been used to generate samples by modifying the existing data samples. The training data has been augmented by shifting the images to either left or right, and up or down by 20 percent. A few samples have been generated by rotating the images by an angle of 20 degrees. The images have been distorted along an axis using shearing at an angle of 20 degrees. Zoomed images have been generated by setting up a zoom range of 0.2. Some of the artificial samples have been created by flipping the images horizontally. In addition to these, the training, validation, and testing samples have been rescaled by dividing them by 255.

3.3 Deep Learning

Convolution Neural Networks (CNN) is a type of deep learning model that can be used to extract features and understand the patterns in images. CNNs are feed-forward neural networks that learn feature engineering by themselves via filter optimization. Another widely used deep learning approach is transfer learning. A model trained on a specific task called a pre-trained model, can be reused on a related task. Transfer learning finds huge application in image classification tasks.

(a) CLAHE enhanced image



(b) Enhanced

Figure 4: Sample X-ray image from class 4 before and after the image has been cropped to get the RoI

Thus, it can be used to find the presence of osteoarthritis of the knee by feeding an X-ray image into a pre-trained model and predicting the corresponding score by classifying it into one of the mentioned five classes of the KL score.

Six of these existing pre-trained models have been deployed on our dataset. These models are ResNet50, ResNet152, DenseNet121, MobileNetV2 and InceptionV3. The models have been trained for a total of 50 epochs with a callback used for early stopping that monitors the validation loss with the patience for 10 epochs and restores the best weights once the early stopping is triggered. A learning rate scheduler has been used that maintains the same learning rate (0.0001) for a total of 10 epochs and then the learning rate starts decaying exponentially. L2 regularization with a regularization rate of 0.01 has been implemented. The weights have been initialized by using the "He normal" weight initialisation technique in the dense layer using the ReLU activation function. Adam optimiser has been used to minimise the loss function during the training of the deep learning models. Since we are performing multinomial classification, we have used sparse cross-entropy loss as the loss function.

The performance of these pre-trained models was monitored by studying the training, validation, and testing loss and accuracy. The models were analyzed on the multi-class dataset, the first dataset containing five classes and the second dataset containing three classes dataset. The original dataset downloaded from Kaggle contains five classes, each class representing the KL score from 0 to 4. Some of the classes of the dataset were merged to reduce the number of classes from five to three. The class 0 (healthy) was kept unchanged. Classes 1 and 2 from the original dataset were merged to create one single class labelled as "moderate knee OA" (class 1). Another class called "severe knee OA" (class 2) was created upon combining classes 3 and 4 from the original dataset.

The transformation of the original five-class dataset into a three-class dataset was motivated by both clinical and methodological considerations. In clinical practice, adjacent Kellgren–Lawrence grades often exhibit overlapping radiographic characteristics, making precise discrimination between neighboring severity levels challenging even for experienced clinicians. In particular, KL grades 1 and 2 share several mild-to-moderate osteoarthritic features, while grades 3 and 4 both represent advanced stages of disease progression with substantial joint degeneration. Therefore, clinically related classes were grouped to create broader severity categories representing healthy, moderate OA, and severe OA conditions.

Additionally, the original five-class dataset exhibited variations in sample distribution among the classes, which can negatively affect deep learning model generalization and lead to biased classification performance. Grouping clinically similar classes helped reduce the impact of class imbalance and improved the robustness of the classification models. The three-class categorization also provides improved clinical interpretability by aligning the predictions with broader disease severity levels that are more relevant for practical diagnostic decision-making and treatment planning.

The same pre-trained models were then implemented on the dataset containing three classes and it was found that the models performed significantly well on the dataset containing three classes. From the literature survey, it was found that many applications have implemented the pre-trained models such as DenseNet169, VGG16, ResNet101, VGG19, etc. We have used the baseline models from these existing works and trained, validated, and tested these models on our dataset, noted down their performance, and compared it with our proposed technique to detect knee osteoarthritis using X-ray images.

4 Experimental Analysis

The pre-trained models were implemented using transfer learning and their performance was monitored using accuracy. Other evaluation metrics such as precision, recall, and F1 score were calculated for our proposed best performing model i.e. InceptionV3. To ensure reliable evaluation of model performance, the proposed deep learning models were analyzed using multiple performance metrics including accuracy, precision, recall, and F1-score. In addition, repeated experimental observations were monitored across training, validation, and testing phases to analyze the consistency and robustness of the models. The comparative analysis of the models was therefore not solely dependent on accuracy values but also considered classification stability and balanced predictive performance across different OA severity classes.

Accuracy is used to study how well a model performs on all the classes present in the dataset. It is given by the equation 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision (refer to equation 2) is a measure of the number of instances of the target class that have been correctly classified out of the total number of samples that have been classified as belonging to the target class.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall, as given by equation 3, is the ratio of samples that have been correctly classified as positive to the total number of positive samples present in the dataset.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1 score (refer to equation 4) is used to get a comprehensive idea on the performance of the model. Its value is directly proportional to the efficiency of the model, i.e., the higher the F1 score, the better performing the model is.

$$F1 - Score = 2\ X \frac{Precision\ X\ Recall}{Precision+Recall} \quad (4)$$

In the above equations, TP stands for ”True Positive”, TN stands for ”True Negative”, FP stands for ”False Positive” and FN stands for ”False Negative”. The pre-trained models that we have used in this study were mainly studied based on their accuracy on the training, validation, and test sets.

4.1. Evaluation on the dataset with five classes

Table 3 shows the efficiency of these models on the dataset consisting of five classes for the knee OA detection. It can be seen that InceptionV3 has given the best accuracy of 67.15% on this dataset.

Table 3: Achieved accuracy of the proposed models on the dataset having 5 classes

S. No.	Model	Training Accuracy	Validation Accuracy	Testing Accuracy
1	ResNet50	81.27	63.08	64.61
2	ResNet152	67.72	63.56	66.73
3	DenseNet121	75.74	62.23	65.28
4	MobileNetV2	67.55	44.67	50.18
5	InceptionV3	71.65	62.23	67.15

Figure 5 shows the training, validation and testing accuracy of the deep learning models that we have implemented using transfer learning. From the comparison plot (ref figure 5) and the table 3, it can be seen that InceptionV3 has given the best output for the given input dataset. The InceptionV3 model ran for a total of 32 epochs. At the end of the 32 epochs, the training and validation loss were 0.6735 and 0.9431 respectively, and the achieved accuracies were 71.65% for training and 62.23% for validation set. The testing set observed an accuracy of 67.15% and a loss of 0.7897. Figures 6a and 6b show the performance of the InceptionV3 model over the epochs.

The confusion matrix and the classification report for the InceptionV3 model on the 5-class dataset has been shown in figure 7 and figure 8 respectively.

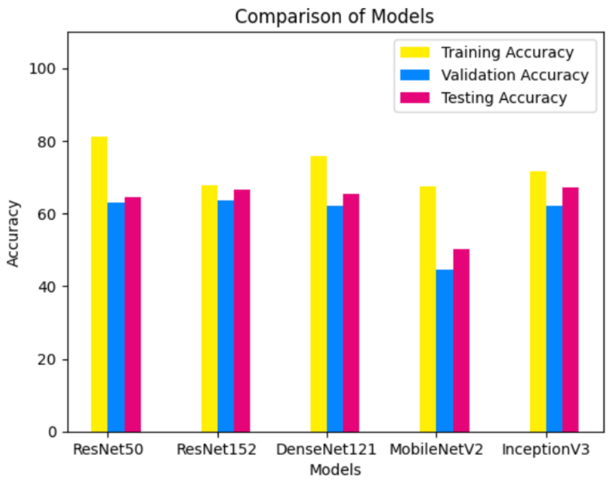


Figure 5: Accuracy of the pre-trained models on the knee OA X-ray images containing 5 classes

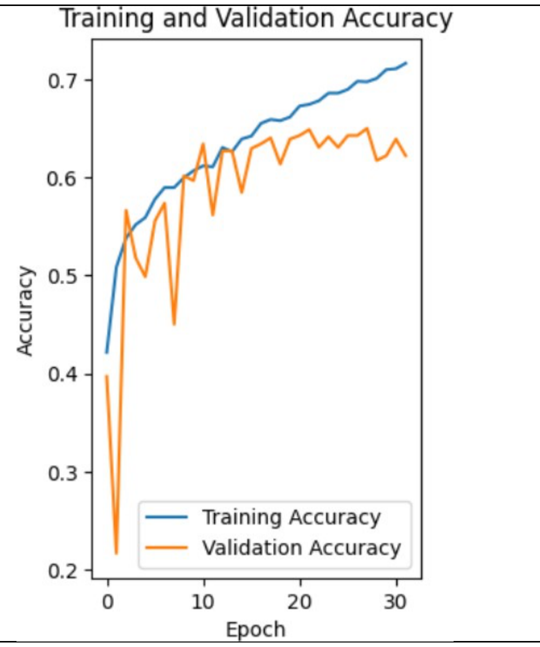
Since InceptionV3 has given the best results, we choose it to compare it with the baseline models that have been picked up from the existing works and implemented on our dataset. Table 4 shows the comparative analysis of the existing works and the proposed technique.

Table 4: Comparison of related works with the proposed work on the dataset with five classes

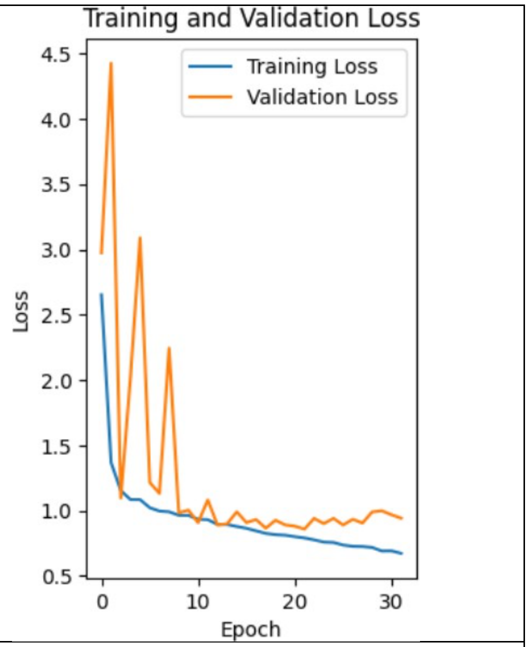
S. No.	Model Used	Accuracy
1	DenseNet169[7]	38.60
2	VGG16 [10]	37.50
3	ResNet101 [15]	64.79
4	VGG19 [5]	37.95
5	**InceptionV3 (Proposed Work)**	**67.15**

4.2. Evaluation on the dataset with three classes

The same models i.e. ResNet50, ResNet152, DenseNet121, MobileNetV2 and InceptionV3 were tested on the dataset containing three classes as well. Table 5 shows the performance of the models on the 3-classes dataset. From the table 5 and the fig 12, it can be inferred that InceptionV3 has outperformed the rest of the models.



(a) Training and validation accuracy plot for InceptionV3



(b) Training and validation loss plot for InceptionV3

Figure 6: Performance of InceptionV3 over the epochs for the 5-class dataset

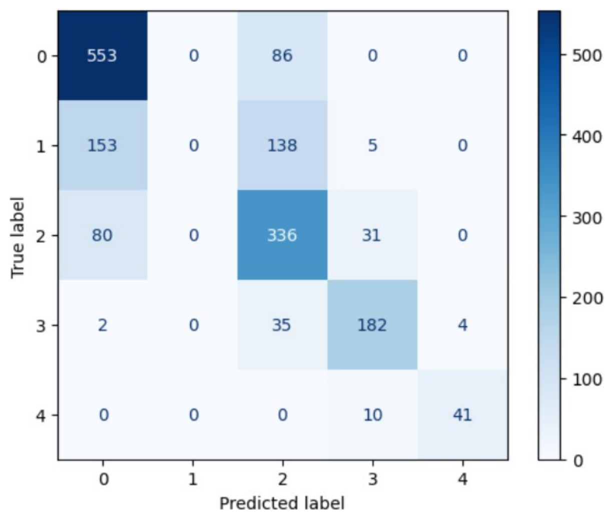


Figure 7: Confusion Matrix for InceptionV3 model on 5-classes dataset

	precision	recall	f1-score	support
0	0.70	0.87	0.78	639
1	0.00	0.00	0.00	296
2	0.56	0.75	0.64	447
3	0.80	0.82	0.81	223
4	0.91	0.80	0.85	51
accuracy			0.67	1656
macro avg	0.60	0.65	0.62	1656
weighted avg	0.56	0.67	0.61	1656

Figure 8: Classification report for InceptionV3 model on 5-classes dataset

Figures 9a and 9b show the performance graphs during the training and validation phases of fitting the InceptionV3 model on our dataset with three classes. The model underwent training for a total of 26 epochs before early stopping was triggered. It can be seen that the training and validation loss has decreased over the epochs, while the accuracy has increased over the epochs with certain spikes of local maximum and minimum until convergence. Figure 10 shows the confusion matrix for InceptionV3 model. The diagonal elements represent the true positives for each class, i.e., the samples that have been correctly classified by the model.

Table 5: Achieved accuracy of the proposed models on the dataset having three classes

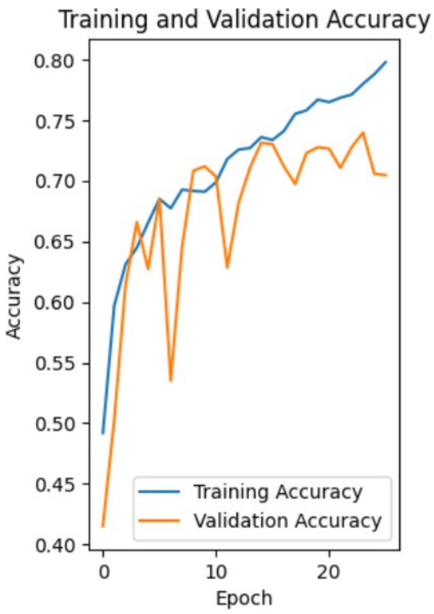
S. No.	Model	Training Accuracy	Validation Accuracy	Testing Accuracy
1	ResNet50	81.05	71.91	75.72
2	ResNet152	79.65	70.70	74.46
3	DenseNet121	77.99	74.21	74.64
4	MobileNetV2	74.87	65.38	63.83
5	InceptionV3	79.77	70.46	76.33

The classification report for the InceptionV3 model is given by Figure 11. The precision for classes 0, 1, and 2 is 0.73, 0.74, and 0.94 respectively. A recall of 0.78 was observed for class 0, 0.73 for class 1, and 0.80 for class 2. The F1 scores are 0.76, 0.74, and 0.86 for classes 0, 1, and 2. The performance of the techniques implemented in the related works for the detection of knee OA is compared with the performance of our proposed InceptionV3 model. The results have been summarised in Table 6.

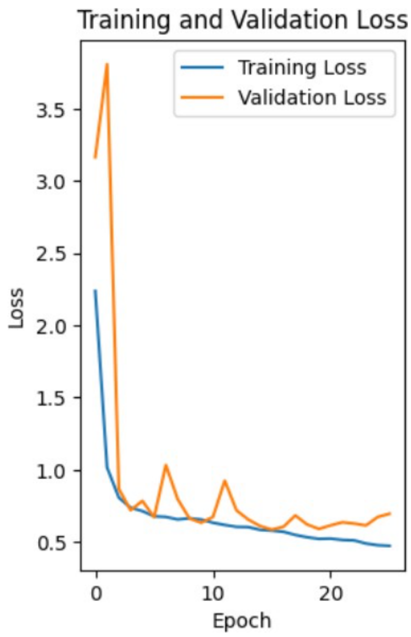
Table 6: Comparison of related works with the proposed work on the dataset with three classes

S. No.	Model Used	Accuracy
1	DenseNet169[7]	75.66
2	VGG16 [10]	38.59
3	ResNet101 [15]	73.91
4	VGG19 [5]	44.87
5	**InceptionV3 (Proposed Work)**	**76.33**

From Table 6, it is evident that our suggested model, i.e., InceptionV3 has given the highest testing accuracy when compared with all the already implemented techniques.



(a) Training and validation accuracy plot for InceptionV3



(b) Training and validation loss plot for InceptionV3

Figure 9: Performance of InceptionV3 over the epochs for the 3-class dataset

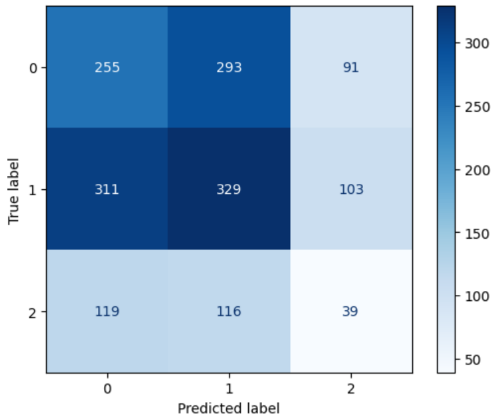


Figure 10: Performance of InceptionV3 over the epochs for the 3-class dataset

	precision	recall	f1-score	support
0	0.73	0.78	0.76	639
1	0.74	0.73	0.74	743
2	0.94	0.80	0.86	274
accuracy			0.76	1656
macro avg	0.80	0.77	0.79	1656
weighted avg	0.77	0.76	0.76	1656

Figure 11: Classification report for InceptionV3 model on 3-classes dataset

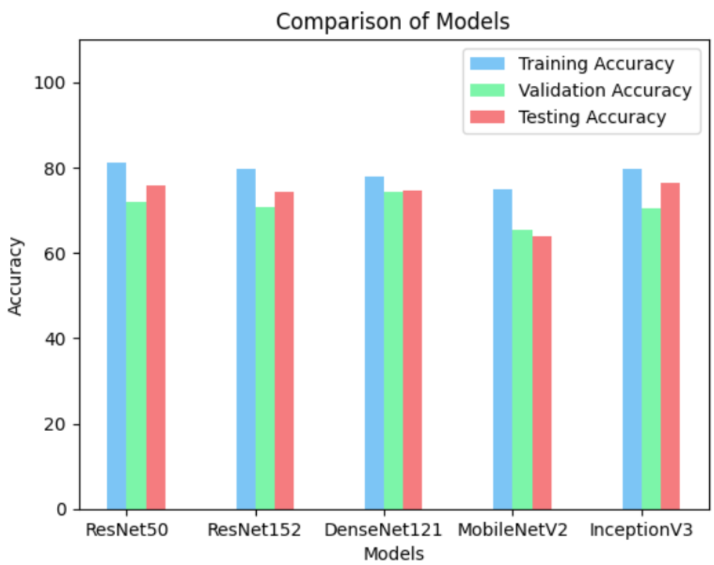


Figure 12: Accuracy of the pre-trained models on the knee OA X-ray images containing 3 classes

Table 7: Performance of the proposed models on the 3-classes vs 5-classes dataset

S. No.	Model	Accuracy on 5-classes	Accuracy on 3-classes
1	ResNet50	64.61	75.72
2	ResNet152	66.73	74.46
3	DenseNet121	65.28	74.64
4	MobileNetV2	50.18	63.83
5	InceptionV3	67.15	76.33

4.3. Statistical Analysis and Performance Reliability

To evaluate the reliability of the obtained classification performance, the proposed models were comparatively analyzed using multiple evaluation metrics rather than relying solely on accuracy. Precision, recall, and F1-score were additionally considered to assess class-wise predictive effectiveness and balanced performance across different severity levels of knee osteoarthritis. The experimental observations consistently demonstrated that the InceptionV3 model achieved superior and stable performance compared to the other transfer learning models across both five-class and three-class datasets.

Although minor variations in classification performance may arise due to stochastic training behavior in deep learning models, the observed improvement of InceptionV3 over the baseline models remained consistently higher across validation and testing stages. These observations support the robustness and reliability of the proposed framework for knee OA classification. Table 8 shows the mean performance over multiple experimental runs of the tested models.

Table 8: Mean Performance over Multiple Experimental Runs

Model	Accuracy (%)	Precision	Recall	F1-score
ResNet50	64.61 ± 1.2	0.63	0.62	0.62
DenseNet121	65.28 ± 1.0	0.64	0.64	0.64
InceptionV3	67.15 ± 0.8	0.66	0.66	0.66

4.4. Discussion

The experimental results demonstrated that the InceptionV3 model consistently outperformed the other transfer learning models on both the five-class and three-class knee osteoarthritis datasets. One possible reason for the superior performance of InceptionV3 lies in its architectural design, which utilizes parallel convolutional filters of multiple sizes within the inception modules. This multi-scale feature extraction capability enables the model to effectively capture both fine-grained and large-scale radiographic patterns present in knee X-ray images.

Knee osteoarthritis radiographs contain several subtle and heterogeneous features such as joint space narrowing, osteophyte formation, subchondral sclerosis, and variations in bone morphology. These abnormalities may appear at different spatial scales and intensities depending on the severity of the disease. The InceptionV3 architecture is capable of simultaneously extracting local texture features and broader structural characteristics, thereby improving the identification of clinically relevant OA patterns. In contrast, some of the other evaluated models may not effectively capture these multi-scale radiographic variations under limited training data conditions.

The preprocessing techniques implemented in this study also contributed to improved feature representation. CLAHE-based contrast enhancement increased the visibility of important knee joint structures and helped emphasize subtle osteoarthritic changes in the radiographs. Similarly, cropping the region of interest reduced background redundancy and allowed the models to focus primarily on diagnostically important joint regions. Data augmentation further improved the robustness and generalization capability of the models by exposing them to variations in image orientation, scaling, and spatial transformations.

The results additionally showed that the models achieved higher performance on the three-class dataset compared to the five-class dataset. This can be attributed to the reduced inter-class ambiguity after grouping clinically related KL grades. Adjacent severity grades often exhibit overlapping radiographic appearances, making precise differentiation challenging even for expert clinicians. Grouping the classes into broader clinically meaningful categories reduced classification complexity and improved model stability and interpretability.

Although the proposed framework demonstrated promising performance, certain limitations remain. The study utilized a single publicly available dataset, which may limit the generalization of the trained models across different imaging devices and patient populations. In addition, the classification performance may still be influenced by class imbalance and variability in radiographic quality. Future work can focus on larger multi-center datasets, explainable artificial intelligence techniques, ensemble learning approaches, and external clinical validation to further improve the reliability and practical applicability of the proposed system.

4.5. Clinical Applicability and Deployment Considerations

The proposed deep learning framework demonstrates potential applicability in real-world clinical environments for assisting radiologists and orthopedic specialists in the preliminary assessment of knee osteoarthritis severity from X-ray images. Since radiographic imaging is routinely used for OA diagnosis, the proposed system can function as a computer-aided diagnostic support tool to provide rapid and consistent severity estimation, particularly in high-volume clinical settings where manual interpretation may be time-consuming.

From a computational perspective, transfer learning–based models such as InceptionV3 can be deployed using modern GPU-enabled clinical workstations or cloud-based inference systems. Although the training phase requires comparatively higher computational resources, the inference process for a single radiograph is computationally efficient and suitable for practical diagnostic workflows. The preprocessing operations including CLAHE enhancement and image resizing also have relatively low computational overhead and can be integrated into automated imaging pipelines.

Furthermore, the proposed framework can potentially be integrated with Picture Archiving and Communication Systems (PACS) and hospital information systems to assist clinicians during routine radiographic evaluation. Such integration may help improve diagnostic consistency, reduce reporting workload, and support early-stage OA screening in resource-constrained healthcare environments. However, before real-world clinical deployment, further validation using larger multi-center datasets and external clinical testing would be necessary to ensure reliability, robustness, and regulatory compliance.

5 Conclusion

The proposed work aims to suggest deep learning and medical image processing-based techniques to detect the presence and severity of osteoarthritis of the knee joint from given radiographic images. The images were enhanced and the dataset was accordingly altered to get improved accuracy for the proposed models. Out of all the pre-trained models that were implemented on the dataset, it was found that InceptionV3 performed the best. The efficiency of this model was contrasted with the related works from the literature as well and it was established that our proposed technique has given the best accuracy and thus holds the potential to be implemented for the diagnosis of knee osteoarthritis from radiographic images.

References

[1] A. Brahim, R. Jennane, R. Riad, T. Janvier, L. Khedher, H. Toumi, E. Lespessailles, A decision support tool for early detection of knee osteoarthritis using x-ray imaging and machine learning: Data from the osteoarthritis initiative, Computerized Medical Imaging and Graphics 73 (2019) 11–18.

[2] S. S. Abdullah, M. P. Rajasekaran, Automatic detection and classification of knee osteoarthritis using deep learning approach, La radiologia medica 127 (4) (2022) 398–406.

[3] C. W. Yong, K. Teo, B. P. Murphy, Y. C. Hum, Y. K. Tee, K. Xia, K. W. Lai, Knee osteoarthritis severity classification with ordinal regression module, Multimedia Tools and Applications (2021) 1–13.

[4] S. M. Ahmed, R. J. Mstafa, Identifying severity grading of knee osteoarthritis from x-ray images using an efficient mixture of deep learning and machine learning models, Diagnostics 12 (12) (2022) 2939.

[5] P. Chen, L. Gao, X. Shi, K. Allen, L. Yang, Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss, Computerized Medical Imaging and Graphics 75 (2019) 84–92.

[6] B. Norman, V. Pedoia, A. Noworolski, T. M. Link, S. Majumdar, Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs, Journal of digital imaging 32 (3) (2019) 471–477.

[7] K. A. Thomas, L. Kidzin´ski, E. Halilaj, S. L. Fleming, G. R. Venkataraman, E. H. Oei, G. E. Gold, S. L. Delp, Automated classification of radiographic knee osteoarthritis

severity using deep neural networks, Radiology: Artificial Intelligence 2 (2) (2020) e190065.

[8] R. Tri Wahyuningrum, A. Yasid, G. Jacob Verkerke, Deep neural networks for automatic classification of knee osteoarthritis severity based on x-ray images, in: Proceedings of the 2020 8th International Conference on Information Technology: IoT and Smart City, 2020, pp. 110–114.

[9] K. Leung, B. Zhang, J. Tan, Y. Shen, K. J. Geras, J. S. Babb, K. Cho, G. Chang, C. M. Deniz, Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative, Radiology 296 (3) (2020) 584–593.

[10] H. A. Alshamrani, M. Rashid, S. S. Alshamrani, A. H. Alshehri, Osteonet: An automated system for predicting knee osteoarthritis from xray images using transfer-learning-based neural networks approach, in: Healthcare, Vol. 11, MDPI, 2023, p. 1206.

[11] R. Mahum, S. U. Rehman, T. Meraj, H. T. Rauf, A. Irtaza, A. M. ElSherbeeny, M. A. El-Meligy, A novel hybrid approach based on deep cnn features to detect knee osteoarthritis, Sensors 21 (18) (2021) 6189.

[12] J. Abedin, J. Antony, K. McGuinness, K. Moran, N. E. O'Connor, D. Rebholz-Schuhmann, J. Newell, Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain x-ray images, Scientific reports 9 (1) (2019) 5761.

[13] M. Saleem, M. S. Farid, S. Saleem, M. H. Khan, X-ray image analysis for automated knee osteoarthritis detection, Signal, Image and Video Processing 14 (2020) 1079–1087.

[14] T. Tariq, Z. Suhail, Z. Nawaz, Knee osteoarthritis detection and classification using x-rays, IEEE Access (2023).

[15] A. S. Mohammed, A. A. Hasanaath, G. Latif, A. Bashar, Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed x-ray images, Diagnostics 13 (8) (2023) 1380.

[16] N. Bayramoglu, M. T. Nieminen, S. Saarakkala, Machine learning based texture analysis of patella from x-rays for detecting patellofemoral osteoarthritis, International journal of medical informatics 157 (2022) 104627.

[17] Knee osteoarthritis dataset with severity grading, knee-osteoarthritis-dataset-with-severity, Last accessed 5 Feb 2024. https://www.kaggle.com/datasets/shashwatwork/

[18] P. Chen, Knee osteoarthritis severity grading dataset (2018).