Detecting Early Stage Knee Osteoarthritis Using Deep Transfer Learning

Lokesh Meesala
001078109
lmeesala@students.kennesaw.edu

Abstract— **Knee osteoarthritis is one of the most prevalent forms of the disease, and its diagnosis can be challenging, especially in its early stages. Imaging techniques such as X-Ray are commonly used to diagnose osteoarthritis, but the interpretation of these images can be subjective and prone to error, especially when detecting subtle changes. In this research, I aim to develop a deep learning network that can classify Knee X-ray images into 5 categories, i.e. 0 - Normal/Healthy, 1 - Doubtful, 2 - Minimal, 3 - Moderate, and 4 - Severe. I propose to use Convolutional Neural Networks (CNN) for multi-class image classification. The baseline model will be a CNN-based deep learning network, which will be trained on a dataset of knee X-ray images. The effectiveness of transfer learning is investigated by applying state-of-the-art CNN architectures such as ResNet, and VGG Nets. To handle the class imbalance, a selective augmentation technique is used. An iterative model training process is used for fine-tuning.**

Keywords—Osteoarthritis, CNN, VggNet, ResNets, Multi-class Image Classification, Transfer Learning, Selective Augmentation.

I. Introduction

Osteoarthritis is the most common form of arthritis, affecting millions of people worldwide. It occurs when the protective cartilage that cushions the ends of the bones wears down over time. It can be difficult to diagnose in its early stages because the early signs and symptoms mimic those of many other diseases. During the diagnosis, the doctor will recommend imaging tests like MRI and X-Ray. These are analyzed by radiologists to determine the severity. They use metrics like the [1] Kellgren-Lawrence grading system; which classifies osteoarthritis into five grades: Grade 0 - Normal, Grade 1 - Doubtful, Grade 2 - Mild, Grade 3 - Moderate, and Grade 4 - Severe. It is often difficult for radiologists to detect this condition in its early stages based on imaging techniques alone. Though this disease cannot be cured, early detection can help patients and medical providers to make the right decisions by controlling the symptoms with medication or therapy. It can slow down the progression of the disease to severe stages where surgery is needed.

In recent years, machine learning techniques have shown great potential in improving the accuracy and efficiency of medical diagnosis, including the detection of osteoarthritis. Many researches and cohort studies are conducted to gather the datasets needed for training these models. In this research I choose the labeled version [2] of the publicly available Osteoarthritis Initiative (OAI) database.

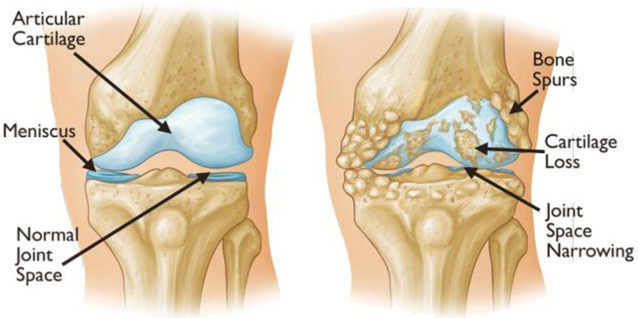


Fig 1: This illustration shows a healthy knee and a knee with osteoarthritis [3]

Convolutional Neural Networks (CNN) are a type of deep learning architecture that is specifically designed for image recognition tasks. They are known to be effective in identifying patterns and features in 2D images by using convolutional and pooling operations while also reducing the number of parameters required to train the model. In this research, I aim to develop a CNN-based deep learning network for detecting osteoarthritis in knee X-Ray images. My contribution is as follows,

- Transfer learning from ResNet and fine tune it by adding custom layers.
- Selective Augmentation is implemented to deal with the class imbalance problem.
- To improve the fine tuning results, iterative model training is implemented.
- Achieved an F1 score of 62.3%.

II. Background and Related work

To improve the early detection of knee osteoarthritis, various studies and research have been conducted. One such study is the Osteoarthritis Initiative (OAI), which is a large, ongoing study that aims to identify biomarkers for the onset and progression of knee osteoarthritis. The OAI study began in 2002 and enrolled over 4,000 participants from the US who were at high risk of developing OA or had early signs of the disease. The study collects various data including clinical, demographic, imaging, and biomarker measurements. The data is available to researchers worldwide to support further research in the field. [2] Chen, Pingjun (2018) labeled these X-Rays and classified them into five categories.

Some studies have used machine learning models such as Logistic Regression, XGBoost, Random Forest, Support Vector Machine (SVM), and K-nearest Neighbor (KNN) to perform binary classification. [4] Kokkotis, Christos et al used physical activity indexes, questionnaire data, and self-reported symptoms for binary classification and reported an accuracy of 77.88% using Logistic Regression. While this looks promising the data used for these models is challenging to collect when compared with easily available and non-invasive x-rays. For these reasons I decided to use the x-ray dataset.

Others have used Fully Convolutional Networks (FCN), Convolutional Neural Networks (CNN), VGG-16, ResNet, and DenseNet for multi-class classification. [5] Antony, J. et al. used X-Ray Data and trained Fully Convolutional Networks (FCN) and Convolutional Neural Networks (CNN) to perform Multi-class classification. They used joint training on regression and classification and reported an F1 score of 61%. They concluded that correctly classifying the conditions severity grades 0, 1 and 2 is challenging due to the small variations in the disease progression. In contrast, [6] Wahyuningrum, R. T. et al. (2019) used a cropped version of the X-Ray dataset and trained VGG-16 Residual Net (ResNet) and DenseNet. They used CNN to extract features and LSTM to classify the severity and reported an accuracy of 75.28 %. My work is closely related to both these research papers. I used selective augmentation to minimize the impact of class imbalance and by transfer learning on Resnet152 with custom fine tuning layers I got an F1 score of 62.3% on the test dataset and 61.5% on the train dataset. For a more efficient fine tuning, iterative model training is used.

The reported metrics in all these studies including mine suggests that there are still challenges in correctly classifying the severity grades of 0, 1, and 2 due to the small variations in disease progression. These findings highlight the need for continued research in this field to develop more accurate and reliable models for the early detection and diagnosis of knee osteoarthritis.

III. Methodology

The first step in this research would be to collect a dataset of knee X-ray images from the Osteoarthritis Initiative (OAI) database. The dataset needs to be labeled based on the severity grading. This has already been done by [2] Chen, Pingjun (2018). Each X-Ray is classified into 5 classes based on the severity of Osteoarthritis. The [1] Kellgren-Lawrence grading system is a widely used system for grading the severity of osteoarthritis based on the degree of joint space narrowing, osteophytes, and other radiographic features. It classifies osteoarthritis into five grades: Grade 0 - Normal/Healthy, Grade 1 - Doubtful, Grade 2 - Mild, Grade 3 - Moderate, and Grade 4 - Severe. The OARSI atlas is another grading system that uses similar criteria to classify osteoarthritis into five grades: 0 - Normal, 1 - Doubtful, 2 - Minimal, 3 - Moderate, and 4 - Severe.

The dataset is divided into training, validation, and testing sets, ensuring that there is no overlap between them, 70-10-20 split is used for this. Keras Image Augmentation techniques such as rotation, zoom, flip, brightness, and contrast adjustment is used to increase the size of the dataset and improve the generalization of the model. This is done using dataset generators provided in tensorflow.keras. As our baseline models, I choose pre-trained VGG16 and ResNet without fine tuning layers. After analyzing the initial results, ResNet152 is used as the transfer learning model.

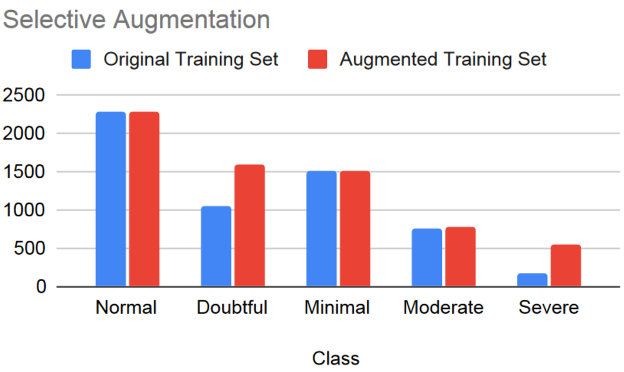


Fig 2: Selective Augmentation Technique.

To deal with class imbalance issues, selective augmentation is implemented. Using this, images in under represented classes are augmented and added to the training data. In reality the distribution of X-Rays in each class will not be equal since the majority of X-Rays will be normal and distribution decreases as the severity increases. Considering this a threshold of 0.7 is used, which means each class will have at least 70% of the image distribution of the previous severity class.

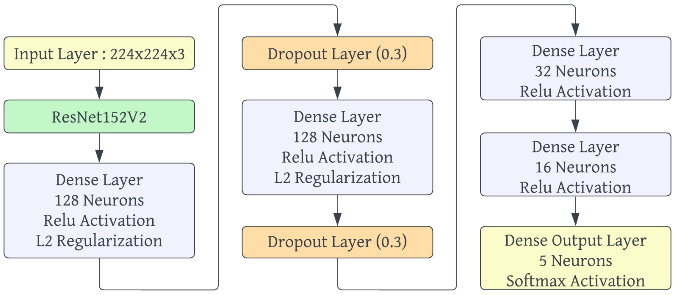


Fig 3: High Level Model Architecture.

IV. Experimental Setup

A. Dataset descriptions

The dataset contains 8260 X-Ray images of size 224x224. The target distribution pie chart is given in Fig 4. Class imbalance is observed, especially for Severe and Moderate, it makes it more difficult for the model to classify them correctly. I have used 5778 X-Ray images for our training purpose, 826 images for validation and 1656 images for testing. i.e. 70% for training, 10% for validation and 20% for testing.

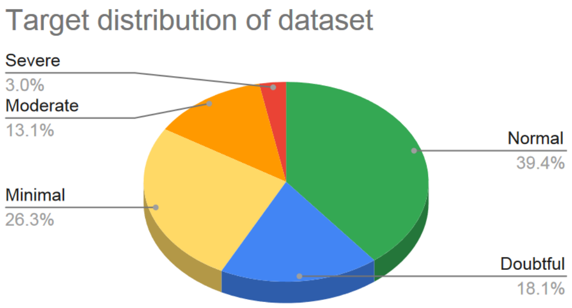
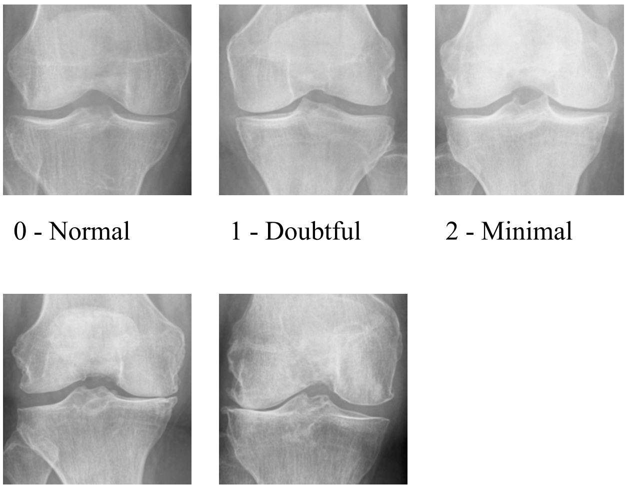


Fig 4: Target Distribution in Full Dataset.



0 - Normal 1 - Doubtful 2 - Minimal

3 - Moderate 4 - Severe

Fig 5: Samples from each class.

B. *Metrics*

Having observed class imbalance, the confusion matrix is drawn to understand the predictions. Metrics like Accuracy, Precision, Recall and F1-score can be calculated with this. Both precision and recall are important in medical diagnosis, but their importance may depend on the specific context and goal of the diagnosis. Precision is the proportion of true positive cases among all cases predicted as positive, while recall is the proportion of true positive cases among all actual positive cases. In medical diagnosis, precision is important to ensure that a positive diagnosis is accurate and trustworthy, while recall is important to ensure that no positive cases are missed, which can have serious consequences. Therefore, a balance between precision and recall is often necessary.

C. *Implementation*

In this project, I have used ResNet, MobileNet and VGG16 architectures for transfer learning. For baseline metrics these models are trained without any added custom layers. These results are further improved by adding more custom layers and implementing Selective Augmentation.

After selectively augmenting the data, images and labels are saved in data frames and Keras flow from dataframe is used for creating batches that go into the training and validation process, in this process images are further augmented using techniques like rotations, shifting and flipping. Fig 3 represents the high level architecture of the model. Models are compiled using adam optimizer, categorical cross entropy loss function and accuracy metric. Fine tuning layers contain a combination of Dense and Dropout layers. Kernel regularization is used in these layers to control overfitting. Finally a Dense layer with softmax activation is used as the output layer. The model is trained for 100 epochs and with an initial learning rate of 0.001. Keras ReduceLROnPlateau and EarlyStopping ensured that the model ran for enough epochs without overfitting or underfitting which can be seen in Fig 6. A batch size of 64 is used after experimentation with different values. In several experiments the model ran for around 70 epochs before stopping early and the learning rate was reduced multiple times and finally reached 0.000008.

To enhance the fine-tuning procedure, an iterative training approach is implemented. Initially, all layers of the ResNet152 model are frozen, meaning their trainable parameter is set to false. From this point, a fixed number of layers, determined as a threshold (in this case, 5), are selected to be unfrozen. The model is then trained as a whole for 10 epochs. Following this, one layer at a time is iteratively released by unfreezing it, and the entire model is trained again for another 10 epochs. This iterative process continues until all layers as per the threshold have been unfrozen and trained. Consequently, the overall model undergoes a total of 50 epochs to complete the training process.

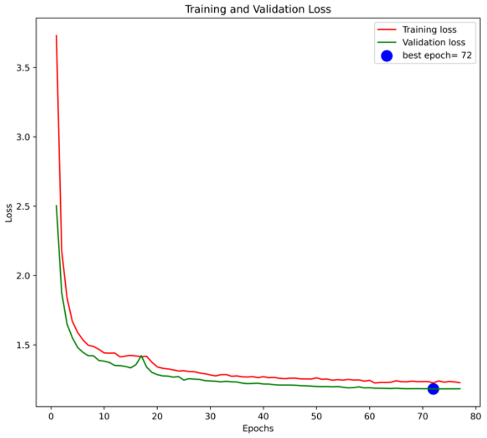


Fig 6: Training and Validation Loss across epochs

To train the deep learning model, categorical cross entropy is used. It is a loss function used in machine learning to measure the difference between the predicted probability distribution and the actual probability distribution of a multi-class classification problem. It is important in multi-class classification as it helps to optimize the model by penalizing incorrect predictions and updating the model's weights to improve its performance. It is used with the softmax activation function, which outputs a probability distribution over the classes. The categorical cross entropy loss function then compares the predicted distribution with the true distribution, and calculates the error or loss, which is minimized during training.

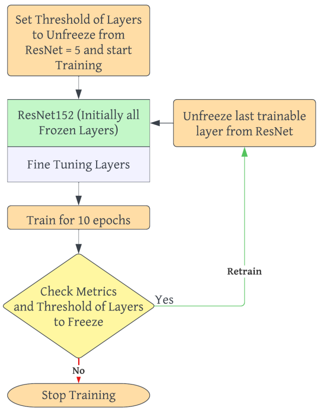
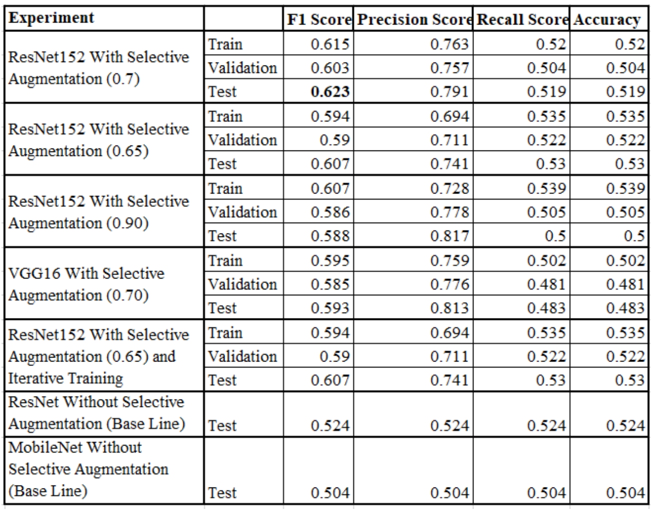

Set Threshold of Layers to Unfreeze from ResNet = 5 and start Training
ResNet152 (Initially all Frozen Layers)
Fine Tuning Layers
Unfreeze last trainable layer from ResNet
Train for 10 epochs
Retrain
Check Metrics and Threshold of Layers to Freeze
Yes
No
Stop Training

Fig 7: Iterative Model Training

V. Results

Evaluation metrics are given in the table below. Without any fine tuning layers, both ResNet and MobileNet gave average results. Fig 8 summarizes all the experiments done in this research. Selective Augmentation technique improved the results by a good margin. But over augmenting the dataset with a higher threshold of 0.9 decreased the f1 score by around 3.5%. This might be because in reality there won't be an equal distribution of the disease progression.

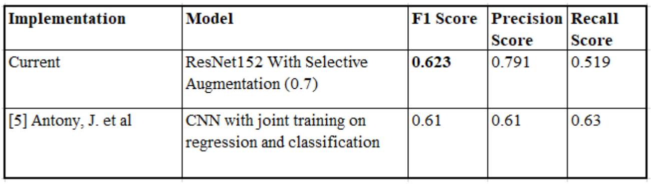
Iterative training process did not improve the scores but the model is trained just for 10 epochs per iteration. More experiments with different architectures, epochs, batch sizes are needed in this area.



Experiment		F1 Score	Precision Score	Recall Score	Accuracy
ResNet152 With Selective Augmentation (0.7)	Train	0.615	0.763	0.52	0.52
	Validation	0.603	0.757	0.504	0.504
	Test	**0.623**	0.791	0.519	0.519
ResNet152 With Selective Augmentation (0.65)	Train	0.594	0.694	0.535	0.535
	Validation	0.59	0.711	0.522	0.522
	Test	0.607	0.741	0.53	0.53
ResNet152 With Selective Augmentation (0.90)	Train	0.607	0.728	0.539	0.539
	Validation	0.586	0.778	0.505	0.505
	Test	0.588	0.817	0.5	0.5
VGG16 With Selective Augmentation (0.70)	Train	0.595	0.759	0.502	0.502
	Validation	0.585	0.776	0.481	0.481
	Test	0.593	0.813	0.483	0.483
ResNet152 With Selective Augmentation (0.65) and Iterative Training	Train	0.594	0.694	0.535	0.535
	Validation	0.59	0.711	0.522	0.522
	Test	0.607	0.741	0.53	0.53
ResNet Without Selective Augmentation (Base Line)	Test	0.524	0.524	0.524	0.524
MobileNet Without Selective Augmentation (Base Line)	Test	0.504	0.504	0.504	0.504

Fig 8: Results of all Experiments.

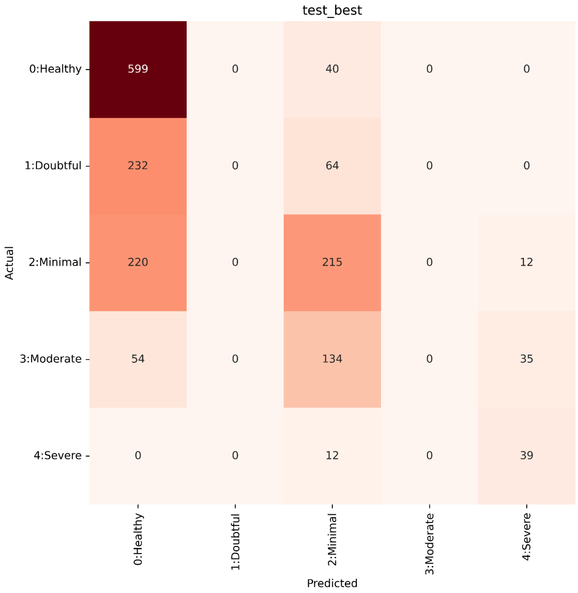
In comparison with [5] Antony, J. et al, my implementation gave a slightly better F1 Score of 62.3% and Precision of 79.1% and lower Recall Score of 51.9%. My model compromised Recall for a better Precision score.



Implementation	Model	F1 Score	Precision Score	Recall Score
Current	ResNet152 With Selective Augmentation (0.7)	**0.623**	0.791	0.519
[5] Antony, J. et al	CNN with joint training on regression and classification	0.61	0.61	0.63

Fig 9: Comparison with related research paper.

Confusion matrices based on the predictions on 1656 test images are also given in Fig 10. It is evident that most of the misclassification occurs on target grades 1 and 3. The models are not capable enough to learn the subtle differences between these severity grades. Even with selective augmentation techniques this issue has not been resolved. The model might give better results with 3 classes. Upon further investigation of the dataset it is found that some images are mislabelled, relabeling this with expert supervision might be needed to improve model metrics.



	0:Healthy	1:Doubtful	2:Minimal	3:Moderate	4:Severe
0:Healthy	599	0	40	0	0
1:Doubtful	232	0	64	0	0
2:Minimal	220	0	215	0	12
3:Moderate	54	0	134	0	35
4:Severe	0	0	12	0	39

Fig 10: Confusion Matrix for Test Predictions

VI. References

[1] Kellgren-Lawrence Scale for radiographic classification of osteoarthritis - mcmaster textbook of internal medicine. (n.d.). Retrieved March 6, 2023, from https://empendium.com/mcmtextbook/table/031_0728

[2] Chen, Pingjun (2018), “Knee Osteoarthritis Severity Grading Dataset”, Mendeley Data, V1, doi: 10.17632/56rmx5bjcr.1

[3] *Osteoarthritis - orthoinfo - aaos*. OrthoInfo. (n.d.). Retrieved April 28, 2023, from https://orthoinfo.aaos.org/en/diseases--conditions/osteoarthritis/

[4] Kokkotis, Christos et al. (2020) “A Machine Learning workflow for Diagnosis of Knee Osteoarthritis with a focus on post-hoc explainability”, en 2020 11th International Conference on Information, Intelligence, Systems and Applications IISA. IEEE.

[5] Antony, J. et al. (2017) “Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks,” arXiv [cs.CV]. Available at: http://arxiv.org/abs/1703.09856

[6] Wahyuningrum, R. T. et al. (2019). A New Approach to Classify Knee Osteoarthritis Severity from Radiographic Images based on CNN-LSTM Method. In 2019 IEEE 10th International Conference on Awareness Science and Technology, iCAST 2019 - Proceedings [8923284] (2019 IEEE 10th International Conference on Awareness Science and Technology, iCAST 2019 - Proceedings). Institute of Electrical and Electronics Engineers Inc.

[7] Mayo Foundation for Medical Education and Research. (2021, June 16). *Osteoarthritis*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/osteoarthritis/symptoms-causes/syc-20351925

[8] Schiratti, J.-B. (2021, October 18). *A deep learning method for predicting knee osteoarthritis radiographic progression from MRI - Arthritis Research & Therapy*. BioMed Central. Retrieved March 6, 2023, from https://arthritis-research.biomedcentral.com/articles/10.1186/s13075-021-02634-4