Stacked Ensemble of Fine-Tuned DCNNs for Knee Osteo - Arthritis Severity Grading

COURSE PROJECT
DA221M

Tanvi Doshi*
t.doshi@iitg.ac.in

Adarsh Gupta*
adarsh.gupta@iitg.ac.in

Japleen Kaur*
japleen@iitg.ac.in

* Equal Contribution. Listing Order is Random

Abstract

Knee osteoarthritis (OA) is one of the most prevalent musculoskeletal conditions affecting the lives of a substantial portion of the global population. It can cause significant limitations and impairments in daily activities, especially among older individuals. To assess the severity of knee OA, doctors usually examine X-ray images of the affected knee and assign a grade using the Kellgren-Lawrence (KL) grading system which divides KOA severity into five grades ranging from 0 to 4. However, this approach requires a high level of expertise, and time, and is susceptible to subjective interpretation, thereby introducing potential diagnostic inaccuracies. To address this problem, we developed a Deep Learning model capable of efficiently diagnosing and classifying knee osteoarthritis severity based on the knee X-ray images. Our approach encompasses two distinct classification tasks: A binary classifier for distinguishing between the presence and absence of KOA, and a multiclass classifier for precise grading across the KL spectrum. Our proposed methodology is based on the utilization of a stacked ensemble of deep learning models, leveraging transfer learning on a diverse set of pre-trained state-of-the-art architectures including MobileNetV2, You Only Look Once (YOLOv8), DenseNet, EfficientNet, Convolutional Vision Transformer (CvT), and ResNet50. To address the issue of imbalanced class sizes, we used a weighted class strategy. We achieved test accuracy of 71.1% in multiclass classification and 88% in binary classification which is higher than previous works in extant literature.

1 Introduction

Osteoarthritis is a degenerative joint disease in which the cartilage wears away over time. It is characterized by joint pain, swelling, and in extreme cases mobility limitations. All joints in the body are sensitive to such wear and tear, however, due to their weight-bearing nature, the hip and the knee are most likely to be affected. Approximately 13% of women and 10% of men aged above 60 years have symptomatic knee osteoarthritis.[1] Above 70 years of age, the likelihood is as high as 40%. Ranked in the top 50 most common diseases, it is estimated that by 2050, 130 million individuals worldwide will be affected by KOA. Common causes of Knee Osteoarthritis include age, heredity, obesity, and recurring trauma to the knee joint.

Diagnosis of Knee Osteoarthritis can be done using an MRI that reflects the 3D structure of the knee joint revealing two hallmark features - joint space narrowing (JSN) and Osteophyte formation. However, due to its high cost and low availability MRI scans are rarely used. Instead, X-ray scans are used due to their safety and cost-effectiveness. The Kellgren and Lawrence(KL) grading system, which is accepted by the World Health Organization, is used to categorize KOA into 5 grades, where grade 0 signifies a healthy joint, grade 1 signifies doubtful cases, grade 2 reveals the presence of mild OA with the presence of osteophytes and possible JSN, grade 3 is moderate OA with noticeable osteophytes and JSN and lastly, grade 4 denotes severe OA.[2] Currently, the accuracy of detection and classification of KOA into these classes largely depends on the physician's ability and precision. Since these grades have very minute differences and grading may vary from physician to physician, dependability on manual methods is low. Moreover, this process is time-consuming.

---

[1] Hunter H., Ryan M.S. Knee Osteoarthritis-Statpearls-NCBI Bookshelf. (4 August 2019) [(accessed on 2 February 2023)]; Available online: https://www.ncbi.nlm.nih.gov/books/NBK507884/ [Ref list]

[2] Schiphof D., Boers M., Bierma-Zeinstra S.M. Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. *Ann. Rheum. Dis.* 2008;67:1034–1036. doi: 10.1136/ard.2007.079020. [PubMed] [CrossRef] [Google Scholar] [Ref list]

Due to these reasons, and the rising prevalence of KOA, there is an urgent need for automated and efficient approaches to classification. AI-based systems can help achieve objective and reproducible results quickly.

Currently, the only available treatments for KOA are behavioral interventions such as weight loss, exercise, and strengthening of muscles, which may slow the course of the disease. Thus, the best way to tackle KOA and prevent future disability is early detection. This makes the use of machine learning based diagnosis extremely crucial to aid medical professionals make sound decisions.

In recent years, deep learning based disease diagnosis models are gaining popularity. Some common diseases include Alzheimer's disease [3], breast cancer [4], pneumonia [5], and heart failure [6]. Image classification using pre-trained models such as MobileNetv2, VGGNet, ResNet, and several others has been proven to be quite powerful.

This paper aims to build on previous works[7] in an attempt to increase accuracy using an ensemble-based learning technique on the following state-of-the-art models - MobileNetv2, ResNet50, EfficientNet, YOLOv8, DenseNet, and CvT for multiclass classification as well as binary classification. To the best of our knowledge, stacking ensemble learning and class-weighted cross-entropy loss have not been used for this task in earlier works.

The dataset used is the OAI dataset. Analyzing the dataset reveals that the data is imbalanced with more data samples of grade 0 and grade 1 KOA and lesser samples of grade 4. This imbalance leads to undesirable training of the models. Models learn to output certain classes rather than learning the actual features. To resolve this, a class-weighted loss function has been employed in this paper, achieving highest test accuracy of 67.3% for multiclass classification and 86.5% for binary classification among all models.

To improve accuracy further, stacking ensemble techniques using the following meta-learners: Random Forest, CatBoost, KNN, Logistic regression, XGBoost, LightGBM, and TabNet have been used. The highest test accuracy achieved with this is 71.1% for multiclass classification and 87.9% for binary classification. Moreover, the highest balanced test accuracy achieved is 73% for multiclass classification and 87.5% for binary classification. There is a clear improvement in accuracy in comparison to the previous works.

The rest of the paper has been organized as follows- Section 2: Related work, Section 3: Theoretical background, Section 4: Materials and methods, Section 5: Results, Section 6: Conclusion.

---

[3] A. Farooq, S. Anwar, M. Awais and S. Rehman, "A deep CNN based multi-class classification of Alzheimer's disease using MRI," 2017 IEEE International Conference on Imaging Systems and Techniques (IST), Beijing, China, 2017, pp. 1-6, doi: 10.1109/IST.2017.8261460.

[4] Heang-Ping Chan, Ravi K. Samala, Lubomir M. Hadjiiski, CAD and AI for breast cancer—recent development and challenges, *British Journal of Radiology*, Volume 93, Issue 1108, 1 April 2020, 20190580, https://doi.org/10.1259/bjr.20190580

[5] Sharma, A., et al. "Detection of Pneumonia using ML & DL in Python." *IOP Conference Series: Materials Science and Engineering*. Vol. 1022. No. 1. IOP Publishing, 2021.

[6] Ogunpola, A.; Saeed, F.; Basurra, S.; Albarrak, A.M.; Qasem, S.N. Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics* **2024**, *14*, 144. https://doi.org/10.3390/diagnostics14020144

[7] Mohammed AS, Hasanaath AA, Latif G, Bashar A. Knee Osteoarthritis Detection and Severity Classification Using Residual Neural Networks on Preprocessed X-ray Images. Diagnostics (Basel). 2023 Apr 10;13(8):1380. doi: 10.3390/diagnostics13081380. PMID: 37189481; PMCID: PMC10137589.

2 Related Works

Previous work done on Knee Osteoarthritis is highlighted as follows. In 2016, Antony et al.[8] used a fully convolutional neural network to quantify knee OA severity using KL grades as input. Norman et al.[9] proposed an ensemble based technique using state of the art models. Tiulpin et al.[10] in 2018 presented a technique based on deep Siamese CNNs, which employs a concept called Contrastive Loss to gauge the similarity between pairs of images within a dataset.

Chen et al.[11] developed a model based on two CNNs and the best performing one was chosen for classification. A specialized YOLOv2 network was employed to detect the X-ray images. Moustakidis et al.[12] worked on deep neural networks, followed by Thomas et al.[13] proposing newer networks with increased accuracy. Tiulpin et al.[14] have developed DeepCNN that leverages an ensemble network of 50 layers. Brahmin et al.[15] presented a computer-aided diagnostic method using various ML techniques such as ICA, random forest and Naive Bayes. Wang et al.[16] proposed a fully automatic scheme based on a pre-trained YOLO model. Yadav et al.[17] suggested a highly effective SFNet model and Lau et al.[18] developed a method based on ImageNet, the Xception model and a dataset of X-ray images. Finally, work on ML and DL techniques together showed excellent results for binary-class classification. However, they did not prove to be very accurate for multiclass classification.

Mohammed et al.[19] trained six models (VGG16,VGG19,ResNet101,MobileNetv2,InceptionResNetv2 and DenseNet) on the Osteoarthritis Initiative (OAI) Dataset, consisting of a total of 9786. They experimented on how the number of classes affects the classification accuracy, by creating 3 datasets from the original dataset, binary classification of KOA diagnosis, three-class classification of KOA severity, and five-class classification based on KL scale. They achieved a highest test accuracy of 69% in multiclass classification, 83% in binary classification, and 89% in 3 classes classification.

---

[8] Antony, J.; McGuinness, K.; Moran, K.; O'Connor, N.E. Automatic Detection of Knee Joints and Quantification of Knee Osteoarthritis Severity Using Convolutional Neural Networks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 376–390. [Google Scholar]

[9] Norman, B.; Pedoia, V.; Noworolski, A.; Link, T.M.; Majumdar, S. Applying Densely Connected Convolutional Neural Networks for Staging Osteoarthritis Severity from Plain Radiographs. *J. Digit. Imaging* **2019**, *32*, 471–477. [Google Scholar] [CrossRef] [PubMed]

[10] Tiulpin, A.; Thevenot, J.; Rahtu, E.; Lehenkari, P.; Saarakkala, S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. *Sci. Rep.* **2018**, *8*, 1727. [Google Scholar] [CrossRef] [Green Version]

[11] Chen, P.; Gao, L.; Shi, X.; Allen, K.; Yang, L. Fully Automatic Knee Osteoarthritis Severity Grading Using Deep Neural Networks with a Novel Ordinal Loss. *Comput. Med. Imaging Graph.* **2019**, *75*, 84–92. [Google Scholar] [CrossRef]

[12] Moustakidis, S.; Papandrianos, N.I.; Christodolou, E.; Papageorgiou, E.; Tsaopoulos, D. Dense Neural Networks in Knee Osteoarthritis Classification: A Study on Accuracy and Fairness. *Neural Comput. Appl.* **2020**, *5*, 1–13. [Google Scholar] [CrossRef]

[13] Thomas, K.A.; Kidziński, Ł.; Halilaj, E.; Fleming, S.L.; Venkataraman, G.R.; Oei, E.H.G.; Gold, G.E.; Delp, S.L. Automated Classification of Radiographic Knee Osteoarthritis Severity Using Deep Neural Networks. *Radiol. Artif. Intell.* **2020**, *2*, e190065. [Google Scholar] [CrossRef] [PubMed]

[14] Tiulpin, A.; Saarakkala, S. Automatic Grading of Individual Knee Osteoarthritis Features in Plain Radiographs Using Deep Convolutional Neural Networks. *Diagnostics* **2020**, *10*, 932. [Google Scholar] [CrossRef]

[15] Brahim, A.; Jennane, R.; Riad, R.; Janvier, T.; Khedher, L.; Toumi, H.; Lespessailles, E. A Decision Support Tool for Early Detection of Knee OsteoArthritis Using X-Ray Imaging and Machine Learning: Data from the OsteoArthritis Initiative. Comput. Med. Imaging Graph. 2019, 73, 11–18. [Google Scholar] [CrossRef] [PubMed]

[16] Wang, Y.; Wang, X.; Gao, T.; Du, L.; Liu, W. An Automatic Knee Osteoarthritis Diagnosis Method Based on Deep Learning: Data from the Osteoarthritis Initiative. *J. Healthc. Eng.* **2021**, *2021*, 5586529. [Google Scholar] [CrossRef] [PubMed]

[17] Yadav, D.P.; Sharma, A.; Athithan, S.; Bhola, A.; Sharma, B.; Dhaou, I. Ben. Hybrid SFNet Model for Bone Fracture Detection and Classification Using ML/DL. *Sensors* **2022**, *22*, 5823. [Google Scholar] [CrossRef]

[18] Lau, L.C.M.; Chui, E.C.S.; Man, G.C.W.; Xin, Y.; Ho, K.K.W.; Mak, K.K.K.; Ong, M.T.Y.; Law, S.W.; Cheung, W.H.; Yung, P.S.H. A Novel Image-Based Machine Learning Model with Superior Accuracy and Predictability for Knee Arthroplasty Loosening Detection and Clinical Decision Making. *J. Orthop. Transl.* **2022**, *36*, 177–183. [Google Scholar] [CrossRef]

[19] Mohammed AS, Hasanaath AA, Latif G, Bashar A. Knee Osteoarthritis Detection and Severity Classification Using Residual Neural Networks on Preprocessed X-ray Images. Diagnostics (Basel). 2023 Apr 10;13(8):1380. doi: 10.3390/diagnostics13081380. PMID: 37189481; PMCID: PMC10137589.

3 Theoretical Background

3.1 Convolutional Neural Networks

CNNs are deep learning algorithms designed and trained to enable computers to view the world as humans do. They are most commonly used for image classification and computer vision tasks. In the context of image classification, CNNs learn the distinguishable features of the input images in order to make predictions. CNNs mainly comprise convolutional layer, activation layer, pooling layer and fully-connected layer.

Convolutional layer - It is the main part of the CNN that activates certain features of the image using convolutional filters called kernels. These filters convolve with the image through matrix operations and create a feature map. This information is used to extract features from the image.

Activation layer - The output of the convolutional layers is activated in order to introduce non-linearity to the system. Common activation functions used are ReLU, Soft max(used for multiclass classification by assigning probabilities) and Sigmoid(mainly used for binary classification). The activated features are passed onto the next layers.

Pooling layer - Also known as downsampling, this layer conducts dimensionality reduction, thereby reducing the number of parameters. This helps minimize overfitting and the number of extracted features. The pooling layer also sweeps a filter across the input layer, applying an aggregate function. Common methods are max pooling and average pooling.

Fully connected layer - A fully connected layer connects every node in the output layer to a node in the previous layer. This layer performs classification on the basis of the features extracted in the previous layers.

3.2 Transfer Learning for CNN

Often in machine learning and deep learning, building a model from scratch can be time consuming and impractical. Transfer learning works by using pre-trained CNN models as a starting point and using it for a related problem. It uses the knowledge gained from the original model and adjusts the weights in accordance with the new dataset/domain. This technique works because in the early layers of the network, the model learns basic features like color, edges,etc that are common among many different tasks. The critical part of transfer learning is to fine-tune the model appropriately for the new task.

3.3 Transformers

Transformers is a deep learning technique that establishes a relationship between the input sequence and output sequence. It mainly has two components: an encoder and a decoder. What sets transformers apart is their self-attention mechanism. Instead of looking at the data in sequence, this mechanism enables the model to look at different parts of the sequence all at once and determine which parts are most important. Thus, transformers are able to establish long-range dependencies that cannot be detected by basic ML models.

3.4 Ensemble Learning

Ensemble learning is an approach in which two or more models are fitted to the same data and the predictions of the respective models are combined in an attempt to achieve better performance by reducing variance.

Common ensemble learning techniques include bagging, boosting and stacking[20]. Our paper uses stacking ensemble learning algorithms.
Although the concept of stacking was originally developed in 1992[21], the theoretical guarantees for stacking were not proven until the publication of a paper titled, “Super Learner”, in 2007[22]. In this paper, it was shown that the Super Learner ensemble represents an asymptotically optimal system for learning.
Stacking typically uses information from various models in terms of probabilities and class labels and combines it to generate a new model. Stacking methods may perform better than the underlying models. The meta-learners used for stacking by us as are follows:

Random Forest - A multitude of decision trees are created during the training phase, the final result is predicted by taking the mode of the predictions outputted by each of these trees. In order to ensure that the predictions of the trees are unique, during training a random subset of features are selected for each tree.

KNN - It is a popular method that relies on the fact that similar data points tend to have similar labels. KNN finds the K-nearest neighbors to a given data point based on metrics like Euclidean distance and determines its label based on the mode or average of the K-neighbors. It is non-parametric and makes no assumptions about the data.

Logistic Regression - A classification model that takes the input and outputs a probability of whether it belongs to a class or not based on a threshold(0.5). The function used to implement this is a sigmoid function.

CatBoost - Catboost is a variant of gradient boosting that can handle both categorical and numerical features. It works by iteratively building decision trees to minimize errors and improve predictions.

LightGBM - LightGBM is an open-source, distributed, high-performance gradient boosting framework developed by Microsoft. It incorporates several novel techniques, including Gradient-based One-Side Sampling (GOSS), and histogram-based algorithms for efficient tree construction.

TabNet - TabNet provides a high-performance and interpretable tabular data deep learning architecture. It uses a method called sequential attention mechanism to enable which feature to choose to cause high interpretability and efficient training.

XGBoost - XGBoost is an implementation of Gradient Boosted decision trees, in which decision trees are created in a sequential manner.

3.5 Class Weights

Class imbalance is a common problem faced in classification tasks. It occurs when certain classes have a large number of data samples while others have very few. This poses a challenge to predict accurate results as the model does not have enough data to properly learn the features of the minority classes and gets biased towards the majority class. A way to tackle this is to assign class weights to each model. Minority classes are given more weight while majority classes are given less weight. This way the model penalizes misclassification of

[20] https://www.sciencedirect.com/topics/computer-science/ensemble-learning
[21] https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.1533
[22] van der Laan, Mark J., Polley, Eric C and Hubbard, Alan E.. "Super Learner" *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, 2007. https://doi.org/10.2202/1544-6115.1309

minority classes much more than that of majority classes. This minimizes the bias of the model towards certain classes.

$$W_j = \frac{number\ of\ samples}{(samples\ of\ class)_j}$$

3.6 Loss Function

Loss function is a measure of how well the model is predicting the expected outcome. Different loss functions are used depending on the type of data being classified and the problem that is being solved. During the training, the model adjusts its weights in order to minimize the loss function. Mean squared error, cross-entropy loss, hinge loss and mean absolute error are some examples.[23]

We have used class-weighted cross entropy loss. Cross-entropy loss, also known log loss, is a common loss function used for classification models whose prediction output is between 0 and 1.The class-weighted cross-entropy takes into account the weights of each class while calculating the loss.

When the number of classes is 2 (binary classification):

$$L = -[w_1 y \log p + w_0 (1 - y) \log(1 - p)]$$

When the number of classes is more than two(multi-class classification):

$$L = \sum_i -\alpha_i y_i \log(p_i)$$

($\boldsymbol{\alpha}$ stands for the class-weight)

3.7 Models Used

In recent years, deep learning methods have been successfully applied to image classification tasks. Certain convolutional neural network (CNN) architectures are renowned for their remarkable performance across various benchmarks. The CNN architectures used in our approach are as follows:

ResNet - ResNet50's[24] architecture solves the issue of multiple nonlinear layers that fail to learn identity maps and suffer from degradation issues. It does this by using stacked residual units that incorporate convolution and pooling layers.

YOLO - YOLO[25] excels in object detection with its state-of-the-art performance, speed, and accuracy. It frames the task as a regression problem and employs a single CNN to predict spatially separated bounding boxes and class probabilities within a grid.

MobileNetv2 -MobileNetv2[26] is a lightweight neural model, which employs depthwise convolutions to significantly reduce the model parameters while maintaining efficiency and prediction accuracy.

---

[23] https://builtin.com/machine-learning/common-loss-functions

[24] arXiv:1512.03385 **[cs.CV]** (or arXiv:1512.03385v1 **[cs.CV]** for this version) https://doi.org/10.48550/arXiv.1512.03385

[25] Ogunpola, A.; Saeed, F.; Basurra, S.; Albarrak, A.M.; Qasem, S.N. Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics* **2024**, *14*, 144. https://doi.org/10.3390/diagnostics14020144

[26] Sandler, Mark et al. "MobileNetV2: Inverted Residuals and Linear Bottlenecks." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 4510-4520.

DenseNet- DenseNet[27] is a convolutional neural network that improves accuracy by solving the problem of vanishing gradients in high-level neural networks. It achieves this by fostering deep connections between layers, resulting in improved accuracy.

EfficientNet -EfficientNet[28] architecture scales neural network models by uniformly enhancing depth, width and resolution dimensions, justified by the need for increased layers and channels in larger input images.

CvT - Convolutional vision Transformer[29] (CvT) improves performance and efficiency by introducing convolutions into ViT[30], leveraging a hierarchical structure and convolutional Transformer blocks to combine the strengths of CNNs and Transformers.

3.8 Evaluation Metrics

Accuracy: Accuracy is the ratio of the total correct predictions to the total number of samples in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

'TP' : true positive
'TN' : true negative
'FP' : false positive
'FN' : false negative

Balanced Accuracy: To obtain meaningful inferences about the model from the accuracy, the dataset must be balanced. This is because a high classification accuracy on an unbalanced dataset could be the result of a high rate of correct predictions in the class with a larger number of samples. The classes with fewer samples hold less weight in the final accuracy.

ROC: The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR(True Positive Rate) against the FPR(False Positive Rate) at various threshold values

AUC Score: AUC is the Area Under the ROC curve. AUC calculates the two-dimensional area under the entire ROC curve ranging from (0,0) to (1,1). It is the measure of the ability of a classifier to distinguish between classes.The value of AUC ranges from 0 to 1, which means an excellent model will have AUC near 1, and hence it will show a good measure of Separability.

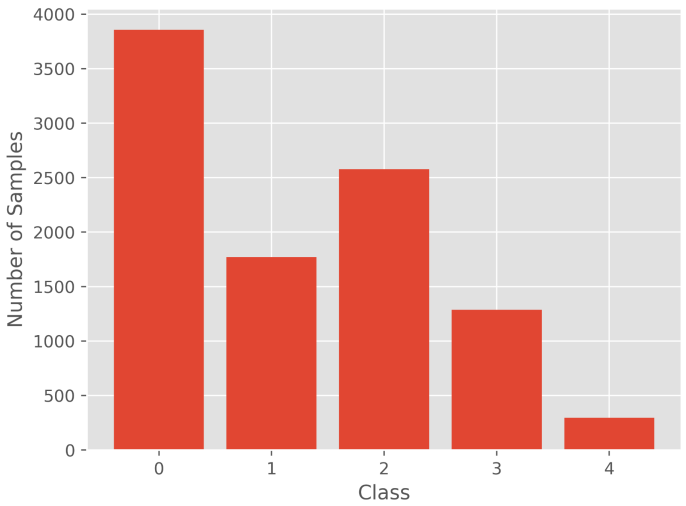
---

[27] Y. Zhu and S. Newsam, "DenseNet for dense flow," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 790-794, doi: 10.1109/ICIP.2017.8296389.
[28] Mingxing Tan, & Quoc V. Le. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.
[29] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, & Lei Zhang. (2021). CvT: Introducing Convolutions to Vision Transformers.
[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, & Neil Houlsby. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

4 Materials and Methods

4.1 Dataset

The source of the knee X-ray images is the OsteoArthritis Initiative(OAI) dataset.[31] Evidence from previous works reveals that this is the most extensively used dataset for KOA prediction. Totally, this dataset has 9786 X-ray images which was later partitioned into train dataset, validation dataset and test dataset in the ratio 7:2:1.



The dataset has 5 classes of images based on the KL grading - 0(healthy), 1(doubtful), 2(mild), 3(moderate) and 4(severe). The dataset was used as such for multi-class classification into the above 5 classes. Binary classification was also performed for which data of class 0 and class 1 was combined to create class 0 (No disease) and class 2,3 and 4 were combined to create class 1(Disease).

4.2 Data Preprocessing:

The original images were not suitable in terms of clarity and localization to be used for deep learning. First segmentation(cropping) was applied to the images so that they capture the part of the X-ray that reveals the most important features required for disease identification.

Next, the quality of the image was improved by applying equalization. Before equalization, the images needed to be converted to grayscale images. Adaptive histogram equalization is a common method that adaptively computes histograms in distinct sections of the image and uses them to redistribute the lightness values of the images.[32] It enhances the definitions of edges in regions of an image. In this paper, CLAHE (Contrasted Limited AHE) is used with a clipping limit of 3. CLAHE operates on smaller regions of the images and then combines neighboring regions by binary interpolation. This helps reduce noise amplification that AHE can lead to. Following this, before the data was fed into the pre-trained models, it was augmented to increase size and diversity of the dataset. Random flip(horizontal) and random zoom(0.1) were performed. Additionally, the images were resized appropriately to be used by the CNNs.

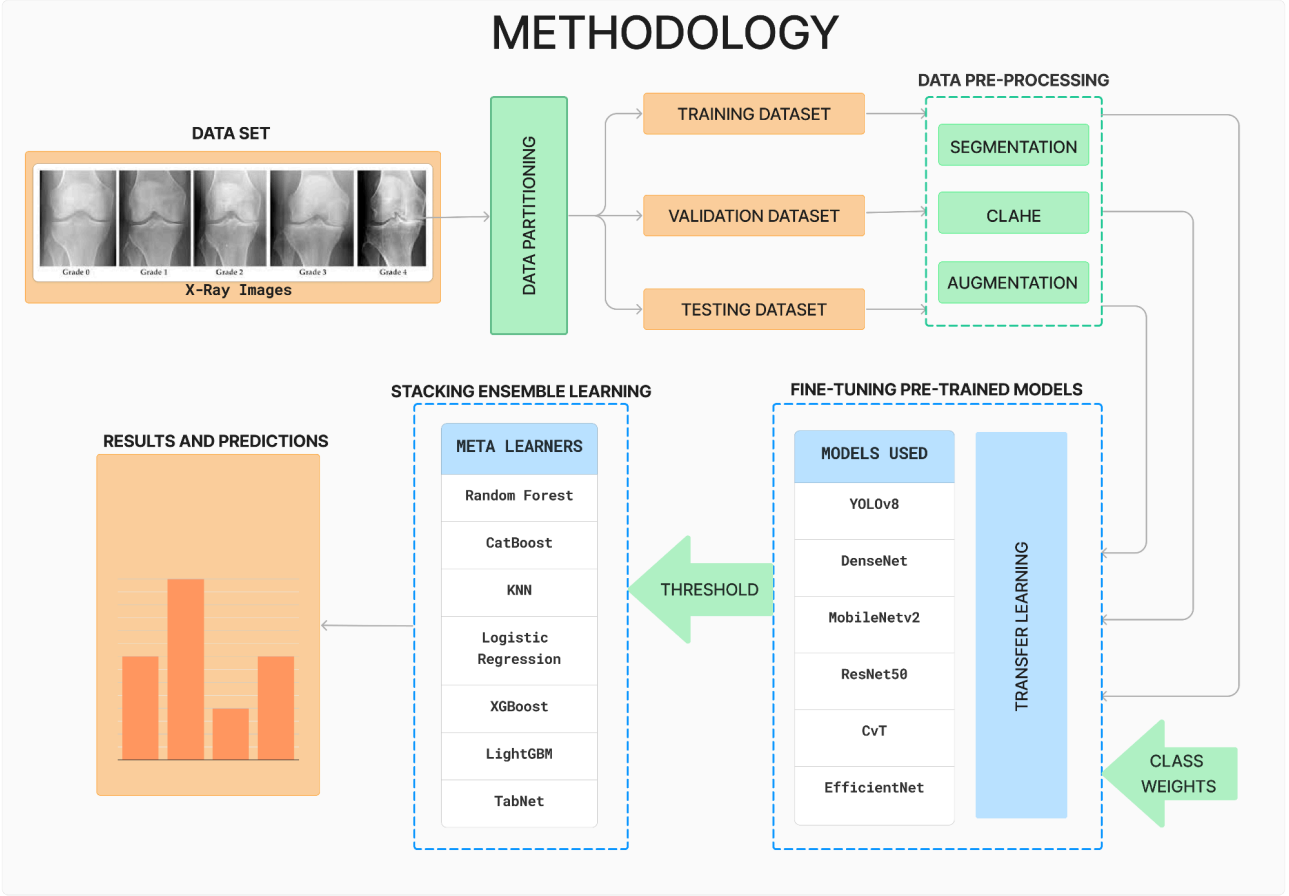
---

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

[32] https://towardsdatascience.com/histogram-equalization-5d1013626e64

4.3 Proposed Framework

4.3.1 Proposed Approach Outline



4.3.2 Fine Tuning Models

For our classification tasks (multiclass and binary), we employed State-of-the-art architectures, namely MobileNetv2, YOLOv8, EfficientNet, DenseNet, CvT and ResNet pretrained on ImageNet1k dataset. This dataset consists of 1000 classes with over a million image samples.

For all models except YOLOv8, we introduced a dense layer with 320 neurons with ReLU activation. Further, we added a dropout layer with dropout probability 0.2 which drops 20% of the layers during training to prevent overfitting. Finally, a dense output layer with 5 neurons is added with softmax activation function.

In DenseNet architecture, an additional global average pooling 2D layer was added before the newly added layers to enhance feature extraction.

We utilized the SGD optimizer with a learning rate of 0.001 and momentum of 0.9. Each architecture was run for varied amount of epochs: MobileNetv2 (40 epochs), YOLOv8 (150 epochs), EfficientNet (10 epochs), DenseNet (9 epochs), CvT (3 epochs), and ResNet (5 epochs).

In multiclass classification, we employed the Categorical Loss Entropy loss function whereas in Binary Classification, Binary Cross Entropy loss function was used. In the last layer of the Binary Classifier, the softmax activation was replaced by sigmoid activation function.

4.3.3 Stacking Ensemble

After fine-tuning the convolutional neural networks, the models that had good accuracies on the validation set were selected as the base learners for stacking ensemble learning. The selection threshold was set to 0.5 for

multiclass classification and 0.7 for binary classification. The final models used as base learners in the stacked ensemble were YOLOv8, DenseNet, and MobileNetv2 for both multiclass classification and binary classification.
To make the process of training the ensemble easier, we ran the entire train, validation, and test set through the models for both multiclass and binary classification to get their class probability prediction. This made the process of fine-tuning the hyperparameters of meta-learners significantly easier.
GridSearchCV and RandomisedSearchCV were used to fine-tune the hyperparameters of the meta-learners. Refer to the tables below to see the final set of hyperparameters obtained for all the meta-learners.

Metalearners	**Hyperparameters**
Random Forest	min_samples_split=16, n_estimators=227
CatBoost	depth=10, iterations=100, learning_rate=0.1
LightGBM	n_estimators=123, learning_rate=0.009
TabNet	optimizer=Adam, step_size=50, gamma=0.9, lr=2e-3
KNN	k=6
Logistic Regression	C=0.01

Table 1: Multiclass Classification Ensemble Meta-learners' best hyperparameters

Metalearners	**Hyperparameters**
Random Forest	criterion='log_loss', min_samples_split=6, n_estimators=180
CatBoost	depth=15, iterations=100, learning_rate=5e-5
LightGBM	n_estimators=76, learning_rate=0.092
TabNet	max_epochs=200
KNN	k=4
XGBoost	learning_rate=0.0024, max_depth=5, n_estimators=96

Table 2: Binary Classification Ensemble Meta-learners' best hyperparameters

After hyperparameter tuning, the best classifier was trained on the train set and used to generate the accuracy score, balanced accuracy score and AUC values on the train, test and val sets.

5 Experimental Results

This section summarizes the results obtained by various CNN and meta learner models used in this paper.

5.1 Fine Tuning

Multiclass	Train acc.	Val acc.
DenseNet	**0.832**	**0.582**
YOLOv8	0.687	0.567
MobileNetv2	0.613	0.556
EfficientNet	0.426	0.327
CvT	0.19	0.03
ResNet50	0.219	0.258

Table 3: Initial Fine Tuning Results for Multiclass classification

Multiclass	Train acc.	Train AUC	Val acc.	Val AUC	Test acc.	Test AUC
DenseNet	**0.832**	**0.967**	**0.582**	**0.849**	**0.673**	**0.898**
YOLOv8	0.687	0.908	0.567	0.823	0.631	0.876
MobileNetv2	0.613	0.84	0.556	0.8	0.575	0.817

Table 4: Final Fine Tuning Results for Multiclass classification

In multi-class classification,DenseNet performed the best with a testing accuracy of 67.3% and a 0.898 AUC of the test dataset. The YOLOv8 and MobileNetv2 also performed reasonably well with testing accuracies of 63.1% and 57.5%. Out of the 6 CNN models trained, CvT had the poorest performance. From this we can infer that transformers didn't work well on the given dataset.

Binary	Train acc.	Val acc.
DenseNet	0.847	0.798
YOLOv8	0.813	0.781
MobileNetv2	**0.94**	**0.821**
EfficientNet	0.572	0.582
CvT	0.615	0.517
ResNet50	0.447	0.464

Table 5: Initial Fine Tuning Results for Binary Classification

Binary	Train acc.	Train AUC	Val acc.	Val AUC	Test acc.	Test AUC
DenseNet	0.847	0.936	0.798	0.86	0.796	0.881
YOLOv8	0.813	0.899	0.781	0.841	0.793	0.876
MobileNetv2	**0.94**	**0.982**	**0.821**	**0.892**	**0.865**	**0.945**

Table 6: Final Fine Tuning Results for Binary classification

In binary classification MobileNetv2 performed the best with a testing accuracy of 86.5% and a test AUC of 0.945. Followed by DenseNet and YOLOv8 with testing accuracies of 79.6% and 79.3% respectively. ResNet50 performed the poorest among all models.
In order to obtain the best results a threshold of 50% testing accuracy for multi-class classification and 70% testing accuracy for binary classification was chosen to decide which models to use with the meta learners. Thus, DenseNet, YOLOv8 and MobileNetv2 were used in the ensemble.

5.2 Stacking results

Multiclass Metalearner	Train acc.	Train bal acc.	Train AUC	Val acc.	Val bal acc.	Val AUC	Test acc.	Test bal acc.	Test AUC
Random Forest	**0.941**	**0.948**	**0.997**	**0.738**	**0.731**	**0.964**	0.707	0.724	0.909
CatBoost	0.912	0.921	0.988	0.701	0.699	0.917	**0.711**	0.72	**0.912**
KNN	0.879	0.89	0.982	0.657	0.658	0.919	0.709	**0.73**	0.875
Logistic Regression	0.852	0.859	0.972	0.632	0.637	0.854	0.71	0.714	0.91
LightGBM	0.863	0.879	0.973	0.613	0.629	0.86	0.705	0.712	0.906
TabNet	0.873	0.891	0.978	0.607	0.626	0.811	0.708	0.723	0.889

Table 7: Results for Multiclass Classification with Stacking

Binary Classification	Train acc.	Train bal acc.	Train AUC	Val acc.	Val bal acc.	Val AUC	Test acc.	Test bal acc.	Test AUC
Random Forest	**0.991**	**0.99**	**1**	**0.958**	**0.956**	**0.996**	0.877	0.872	0.939
CatBoost	0.959	0.955	0.986	0.828	0.819	0.904	**0.879**	**0.875**	**0.945**
KNN	0.959	0.954	0.995	0.855	0.843	0.953	**0.879**	0.871	0.913
XGBoost	0.949	0.941	0.986	0.838	0.82	0.902	0.873	0.861	0.941
LightGBM	0.96	0.957	0.987	0.828	0.82	0.906	0.874	0.87	**0.945**
TabNet	0.865	0.88	0.984	0.757	0.778	0.902	0.806	0.823	**0.945**

Table 8: Results for Binary Classification with Stacking

In Multiclass classification, CatBoost yielded the highest testing accuracy of 71.1% and the highest testing AUC of 0.912. CatBoost achieved a balanced accuracy of 72% on the test set. However, the KNN classifier surpassed this with a Balanced Accuracy of 73% on the test set. All the meta-learners performed reasonably well, attaining test accuracies greater than 70% and Test AUC greater than 0.87.

In Binary classification, both CatBoost and KNN achieved the highest Test Accuracy of 87.9% showcasing strong performance. CatBoost also achieved the highest balanced accuracy on the test set of 87.5%. TabNet got the lowest Test Accuracy of 80.6%. Notably, all the meta learners performed exceptionally well by attaining Test Accuracies greater than 80% and Test AUCs greater than 0.91.

6 Conclusion

In this paper, we introduced a stacked ensemble model for the severity grading of Knee Ortho-Arthritis on the OAI dataset. The ensemble combined the fine-tuned DCNN models to derive a more powerful image classification scheme than individual CNNs. As shown in Section 5, the ensembles used were able to correctly classify the majority of images in the OAI dataset consisting of 9786 images in 5 imbalanced classes, achieving higher accuracy than previous works in the extant literature. We achieved the highest testing accuracy of 71.1% in multiclass classification and 88% in binary classification.
In the future transformers can be incorporated by appropriately fine-tuning with better computational resources, and a larger set of meta-learners can be explored as well.