Knee Osteoarthritis Severity Detection Using Deep Leaning

Shaik Gouse Mastan Vali
Department of Computer Science
Sam Houston State University, Huntsville, TX 77341, USA
Contact: gxs085@shsu.edu

***Abstract* --- Knee osteoarthritis (Knee OA or KOA) is a common and challenging medical condition. Manual diagnosis of KOA can be time-consuming and prone to errors. Since 60% of Knee OA effected to the old-aged people, it is very important to have accurate and error free system. More over Manual Diagnosis need expert physicians with proper experience. Deep neural network (DNN) models can be used to identify and classify Knee OA images. Deep neural networks excel at capturing complex relationships within images, allowing for a more nuanced analysis of Knee Osteoarthritis (KOA) images. In this study, we propose a method for diagnosing Knee OA using images obtained from the Mendeley Data Platform [1] and Kaggle Dataset [17]. We will implement couple of DNN models, Custom CNN, ResNet101, EfficientNetV2, and ConvNextBase and evaluated their performance on a held-out test set.**

The proposed method has the potential to revolutionize the diagnosis of Knee OA. The method could be used to improve the accuracy of diagnosis, and it could also be used to provide early diagnosis of Knee OA. This could lead to earlier treatment, which could help to prevent further damage to the knee joint.

***Index Terms* — Knee osteoarthritis (KOA), X-ray Images, Recurrent Neural Network, Severity Classification**

I. MOTIVATION

Knee osteoarthritis (KOA) is a prevalent disease primarily affecting older individuals. It is the most common joint disease in the United States, with a higher occurrence in women (13%) than in men (10%) above the age of 60 [12]. The Kellgren-Lawrence (KL) grading system is used in the medical field to assess the severity of KOA based on radiographs. Despite the introduction of other medical imaging technologies, radiographs remain the preferred method due to their accessibility and cost-effectiveness [13]. The KL grading system, consisting of five stages ranging from 0 (healthy) to 4 (severe), is used to classify the severity of KOA. However, the accuracy of the diagnosis relies heavily on the expertise of the physician, leading to subjective interpretations and variations in grading [14].

To address the limitations of traditional diagnosis methods, automated approaches utilizing deep learning have emerged. Deep learning, a subfield of machine learning, involves constructing neural networks inspired by the human brain to perform specific tasks such as classification and object detection. Neural networks process information through layered structures and can make observations and inferences from unstructured data without explicit training.

Extensive research in deep learning has resulted in the development of large datasets like ImageNet, MNIST, and MS-COCO, which serve as benchmarks for evaluating deep learning models. Pre-trained models, trained on these datasets, learn generic features that can be applied to related problems in various disciplines. In the context of image classification, popular pre-trained models include VGGNet, MobileNet, ResNet, InceptionResNet, EfficientNet, DenseNet and ConvNextBase (state-of-art). These models have found applications in medical diagnosis, factory quality control, security systems, and other industries.

In the medical field, deep learning techniques have been employed to automate the diagnosis process. For example, in the diagnosis of leukemia, a deep learning model using pre-trained convolutional neural networks (CNNs) achieved high accuracy by distinguishing between cancer cells and healthy cells [15].

In the specific domain of KOA, deep learning techniques have been applied to detect the presence of the disease and classify its severity according to the KL grading system. This automated approach aims to provide accurate and reliable severity classification systems, supporting medical specialists in making more informed diagnoses. The efficiency of these algorithms is calculated by obtaining information retrieval metrics – Precision and Recall.

The rest of this paper is structured as follows. Related works are described in next section. The Timeline for the project is presented in section IV.

II. BACKGROUND

Recent years have seen a lot of activity in Knee OA detection, with numerous studies investigating different machine learning, deep learning, and to handle the issue. This section reviews some of the most related work in the field.

The paper [11] aimed to develop an automated model for detecting the severity of knee osteoarthritis (KOA) using radiographs. The model's performance was evaluated by comparing it with the assessments of musculoskeletal radiologists. They used a large dataset of over 40,000 images from the Osteoarthritis Initiative (OAI) and reported an average accuracy of 71% and an F1 score of 70% for the full test dataset. However, they did not compare their model's performance with other related work in the field.

In their work [6], the authors introduce a model that utilizes joint space width (JSW) for osteoarthritis recognition. Their methodology involves image preprocessing, region of interest extraction, edge computation, and JSW calculation, followed by classification based on JSW. The dataset used consists of 140 labeled images, with severity levels determined by two radiologists and two orthopedic surgeons. The proposed method achieves impressive results, with a 98.4% F1 score and 97.14% accuracy for KOA classification. However, it is worth noting that the small size of their dataset may impact the model's accuracy in real-world scenarios despite its high performance in the study.

In their study [5], the authors propose a computer-aided diagnosis system for detecting knee osteoarthritis (KOA) using machine learning (ML) algorithms. Their approach involves preprocessing X-ray images and applying multivariate linear regression normalization to reduce irregularities between healthy knees and those with osteoarthritis. Feature extraction is performed using independent component analysis, followed by classification using random forest and Naïve Bayes models. The dataset used in their study consists of 1024 knee X-ray images from the OAI. The proposed method achieves an accuracy of 82.98%, specificity of 80.65%, and sensitivity of 87.15%. Notably, their approach relies on pixel intensities for feature extraction rather than texture analysis, demonstrating potential for high accuracies in real-world scenarios.

In their research [7], the authors employed a deep neural network (DNN) to detect osteoarthritis by leveraging statistical data related to patients' health behaviors and medical utilization. The study utilized data from 5749 subjects sourced from the 2015-2016 KNHANES, a nationwide cross-sectional study conducted in Korea. Their proposed method involved automatically extracting features from the data using scaled principal component analysis and DNNs to identify risk factors associated with KOA. The classification model developed achieved an accuracy of 71.97% and a sensitivity of 66.67%. Although this study did not rely on X-ray images, it showcased the potential of utilizing health behaviors and medical utilization as viable alternatives that could be integrated with other methods for detecting osteoarthritis.

In addressing the challenges of overfitting and the non-Euclidean nature of the shape space, the author of [8] proposed an integration of the Graph Convolutional Network (GCN) with the concept of reducing intrinsic dimension. They conducted a comparative analysis between their classifier and an alternative extrinsic approach. The dataset utilized in their study was obtained from the OAI and consisted of 201 randomly selected samples. The images were categorized into grades zero, one, and two, representing the absence, minimal, and definite presence of osteophytes, respectively. The intrinsic model achieved an accuracy of 64.64%, outperforming the Euclidean method, which only achieved an accuracy of 58.62%. It is worth noting that this paper employed a grading system different from the widely used Kellgren–Lawrence grading system, which might explain the relatively lower prediction accuracy observed in their model.

In their research [10], the authors investigated the effectiveness of machine learning (ML) techniques for early detection of knee osteoarthritis (KOA) using a limited dataset. They compared ML methods like ANN, SVM, and RF with their own proposed approach, which included the integration of a group method. The dataset comprised only 31 images, with severity labels based on the Kellgren–Lawrence grading system. By utilizing Zernike-based texture features, their framework achieved an accuracy of 85.0%, resulting in an impressive 11% improvement in diagnostic accuracy. While their approach showed promising results in radiology tests, further studies are needed to validate its effectiveness on a larger scale involving more patient cases.

In their study [9], researchers proposed an automated knee osteoarthritis image classification method using deep neural networks (DNNs). The approach involved two key steps: image preprocessing and classification. Initially, a VGG network was used to extract the knee joint center, followed by classification using a ResNet-50 network. The authors also employed a dataset rebalancing technique to address class imbalances. By combining these techniques, they achieved an 81.41% classification accuracy. This work demonstrates the potential of DNNs and preprocessing methods in accurately classifying knee osteoarthritis images, offering promising implications for automated diagnosis systems.

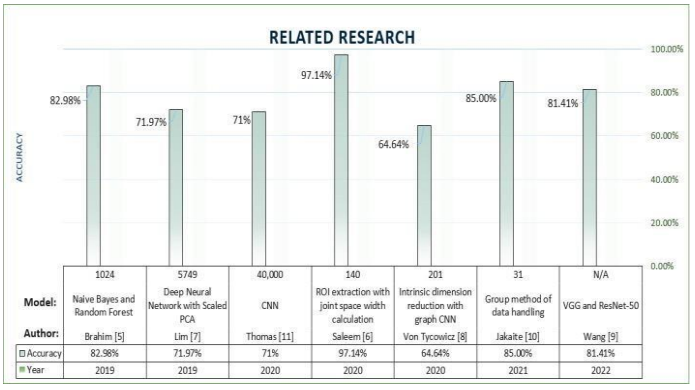


Fig (1): Related Research

Our work focuses on knee osteoarthritis (OA) and utilizes deep learning Convolutional Neural Network (CNN) such as ResNet101, VGG16, Custom CNN, EfficientNet, ConvNextBase and few other models. These models are employed to analyze and classify knee OA images, enabling accurate identification and diagnosis of the condition. By leveraging the capabilities of these advanced convolutional neural networks, we aim to enhance the efficiency and effectiveness of knee OA detection and classification.

III. DATASET

In this Project we used two datasets

Dataset-1:

We utilized the Knee OA dataset, which consists of More than 8,200 X-ray images in. This dataset1 [1] consists of 5 levels from 0(healthy) to 4(severity) images, with labels indicating their authenticity. The dataset is divided into three sections: training, testing, and validation, with 70% in the training set, 10% in the testing set, and 20% in the validation set. The dataset required to pre-process to ensuring the images is clean and suitable for model training. Since it is medical data, the dataset is imbalance, so we need to perform data augmentation to balance the data.

Dataset-2:

We utilized the Expert-I Knee OA Dataset, which comprises over 1200 training X-ray images and approximately 100 test images. This dataset, developed by a group of medical experts, denoted as Dataset2 [17], includes images classified into 5 levels ranging from 0 (healthy) to 4 (severity), with labels indicating their authenticity. Dataset2 required preprocessing and augmentation to ensure the images are clean, balanced, and suitable for training

IV. PROPOSED METHODOLOGY

In this section, we provide a detailed explanation of our proposed approach, depicted in Figure 2. The approach consists of four key stages: data acquisition, dataset preprocessing, model training, and classification. Firstly, we obtained the datasets of knee osteoarthritis (KOA) X-ray images from the Mendeley Data Platform [1], which is also available on Kaggle [16] and Dastset2, which is available in Kaggle as well. These datasets consist of images classified into five different categories: 0 (healthy), 1 (doubtful), 2 (minimal), 3 (moderate), and 4 (severe).

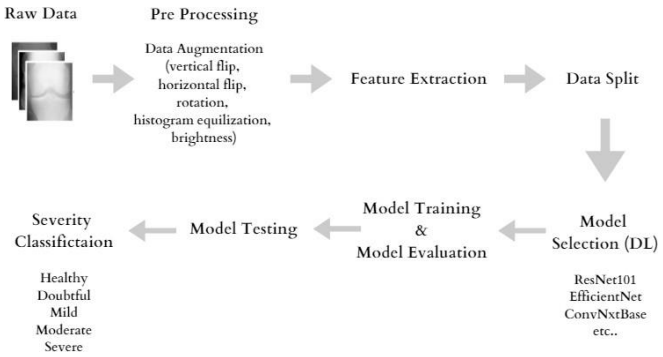


Fig (2): Proposed Methodology

Pre-Processing:

We've addressed the limitations of our knee joint X-ray dataset by implementing robust data preprocessing. This process included precise *image segmentation* to focus on the knee area and enhance image clarity. We also improved contrast and visibility through *histogram equalization.*

$$s_k = T(r_k) = (L-1)\sum_{j=0}^{k} p_r(r_j) \quad (1)$$

$$p_r(r_j) = \frac{n_j}{MN} \quad (2)$$

Formula Represents: histogram equalization

The formula which adjusts the pixel values in an image to enhance contrast and improve visibility. Equation (1) represents the histogram equalization transformation, where 'r' and 's' denote input and output pixel values, respectively. 'L' represents the maximum pixel value in the image. Equation (2) calculates the probability of intensity level 'rj' occurrence, given 'MN' as the total number of pixels and 'nj' as the count of pixels with intensity 'rj'.

Additionally, we employed **Data Augmentation techniques** like *flips, rotations, brightness adjustments, image resizing (240x240 pixels), Normalization, Gray Scale Conversion (pre-processing) and Zoom augmentation.* The goal of these techniques is to prepare and diversify the image datasets, making it suitable for training deep learning models. Augmented Image datasets is saved in a different folder apart from original datasets folder. These steps substantially enhanced the dataset's suitability for deep learning models, reinforcing the quality and efficacy of our project. Along with these we also experimented few noises removal, resizing techniques, and other techniques we implemented and experimented.

Dataset1: Initially, our dataset exhibited significant class imbalance class 0 with 2286 images, class 1 with 1046 images and class 2 with 1516 images, class 3 with 757 images and class 4 with 173 (for training dataset). However, through the strategic application of data augmentation techniques, we have successfully rebalanced the dataset. Each class in our training data now comprises 1400 images, resulting in a total of 7000 images distributed across the five classes. Likewise, our testing and validation datasets each contain 1500 images, collectively contributing to an augmented dataset comprising a total of 8500 images. With this balanced dataset in hand, we are poised to proceed with our tasks.

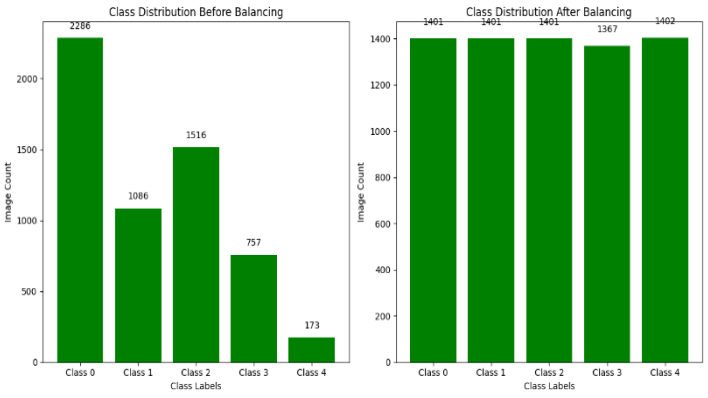


Fig. Train data Distribution of Dataset1.

Dataset2: Initially, our dataset exhibited significant class imbalance class 0 with 432 images, class 1 with 423 images and class 2 with 173 images, class 3 with 157 images and class 4 with 188 (for training dataset). However, through the strategic application of data augmentation techniques, we have successfully rebalanced the dataset.

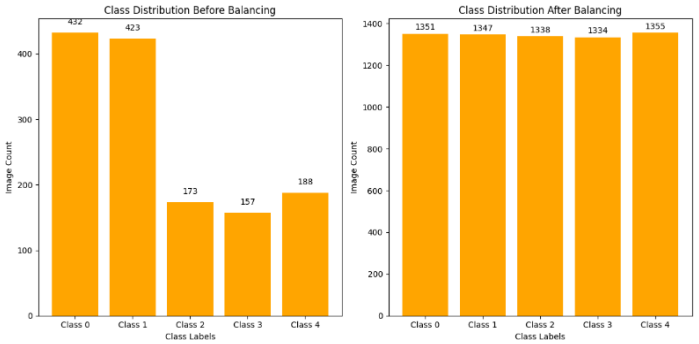


Fig. Train data Distribution of Dataset2.

Each class in our training data now comprises around 1000 images, resulting in a total of 5000 images distributed across the five classes. Likewise, our testing and validation datasets contain 100 and 1750 images, collectively contributing to an augmented dataset comprising a total of nearly 7000 images. With this balanced dataset in hand, we are poised to proceed with our tasks. But we found that quality of images in dataset2 is really good, so we used them in our final model

Datasets was then split into training, testing, and validation sets with percentage of 70%, 10%, 20%, respectively. We will experiment with six pre-trained CNN models (VGG16, ResNet101, EfficientNet, ConvxNextBase, Custom CNN, and other models) using Google Colab's computational resources.

This approach allowed us to effectively preprocess the X-ray images and create different dataset for specific classification tasks. The subsequent sections provide more details on each stage of our proposed methodology.

(i) Deep Learning Models: This section describes the deep learning models implemented for knee osteoarthritis severity detection. Our system consists of several components: preprocessing, feature extraction, model training, and evaluation.

Convolution Neural Network (CNN) In the field of computer vision, the convolutional neural network (CNN) has emerged as a successful deep learning algorithm. CNNs excel at extracting features from images by applying filters that capture spatial dependencies. They transform images into a computationally manageable form while preserving important features. CNNs consist of convolution and pooling layers, responsible for feature extraction and dimensionality reduction. These extracted features are then used for classification with a regular neural network.

Transfer learning is an effective technique that utilizes pre- trained CNN architectures on large datasets. By leveraging the learned features from these architectures, transfer learning enables better performance on new tasks. Popular CNN architectures include MobileNet, VGGNet, ResNet, InceptionResNetV2, EfficientNet, ConvNextBase and DenseNet121.

ResNet101 (Residual Network 101) Architecture

In deep learning, deeper convolutional neural networks (CNNs) often perform better, but they can encounter the vanishing gradient problem, where gradients become too small for effective learning. To address this, ResNet models use residual blocks with skip connections that transmit activations from earlier to later layers, preserving gradient information and aiding optimization. ResNet101, a variant with 101 layers, leverages skip connections as regularization, allowing it to build very deep networks without vanishing gradient issues. These skip connections improve the model's focus on critical features, enhancing overall performance.

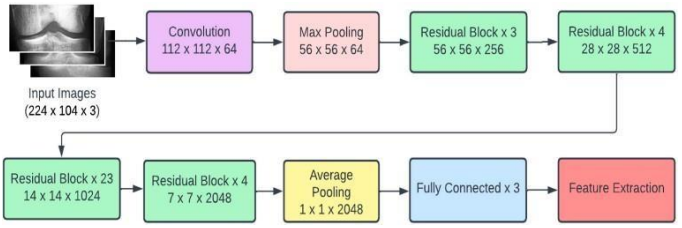


Fig: Resnet101 Architecture Illustration

EfficientNet Architecture (EfficientNet-B0 model)

EfficientNet, a renowned family of convolutional neural network architectures, is valued for its efficiency and high performance in computer vision tasks. Its strong accuracy, computational efficiency, and adaptability make it an excellent tool for knee osteoarthritis (KOA) detection. Our project successfully implemented EfficientNet-B0, achieving accurate KOA diagnosis with reduced computational resources. Ongoing efforts involve fine-tuning and exploration of different model variants for further enhancements

VGG16 Architecture (Visual Geometry Group16)

Another Algorithm we would to implement is VGG16, a deep convolutional neural network (CNN), can extract meaningful features from knee osteoarthritis (OA) images. It captures changes in joint space, bone density, and osteophyte presence. By training on a large dataset of knee X-ray images, VGG16 learns to recognize disease-related features for accurate classification. These features include shape, texture, and spatial relationships of knee structures. VGG16 helps in early detection and monitoring of knee OA by distinguishing between healthy and OA-affected knees. The model requires labeled knee X-ray images, with a sufficient number of samples representing both healthy and OA cases. Additional metadata like patient age and sex can provide valuable insights. Preprocessing techniques like image cropping, resizing, and histogram equalization may enhance image quality for improved performance.

ConvNextBase Architecture

ConvNextBase is a convolutional neural network (CNN) architecture designed for image classification tasks. It is based on the ConvNeXt architecture, a highly efficient and accurate CNN architecture that has achieved state-of-the-art performance on various image classification benchmarks. ConvNextBase builds upon the strengths of ConvNeXt by introducing several modifications that further enhance its accuracy and efficiency with the help of key features like Large Kernal Size, Depthwise Seperable Convoultions, Layer Scaling, Inverted Bottleneck Blocks and etc. In practical application, I have implemented the ConvNextBase architecture as part of my final model, achieving an impressive accuracy of 84%.

Dataset 1	Dataset 2
Inception-V3	Custom CNN
VIT-B16	ResNet-101
VGG-16	VGG-16
ResNet-101	EfficientNet-V2B0
DenseNet-121	ConvNextBase
EfficientNet-V1B0	
ConvNextBase	

Table: Deep Learning Models used for Classification

V. PERFORMANCE EVALUATION MATRIX

In this section, we will outline the evaluation criteria that we will use to rate the effectiveness of our Knee Osteoarthritis Severity Detection system. We will use a variety of metrics, including accuracy, recall, precision, and F1-score, to ensure the assessment is extensive. These metrics will give us a clear picture of the system's performance, allowing us to identify areas for improvement and optimize the models for better results. By employing these evaluation criteria, we aim to assess the system's capability in distinguishing between real and fake news articles accurately and efficiently.

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Formula Represents: Performance Evaluation Metrics

where ‘TP’ is the true positive, ‘TN’ is the true negative, ‘FP’ is the false positive, and ‘FN’ is the false negative.

1. **Accuracy** measures the overall correctness of a model by calculating the ratio of correctly predicted instances to the total instances.

2. **Precision:** Precision assesses the accuracy of positive predictions, focusing on the ratio of true positives to the sum of true positives and false positives.

3. **Recall (Sensitivity or True Positive Rate):** Recall evaluates a model's ability to capture all relevant positive instances, calculating the ratio of true positives to the sum of true positives and false negatives.

4. **F1 Score:** F1 Score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives.

VI. RESULTS

Dataset 1:

Experiment 1:

1. Data Collection: The dataset was sourced from the Mendeley data platform, providing a substantial volume for model training. However, the dataset exhibited an imbalance in class distribution. Notably, the available x-ray images of knee joints were limited in open-source repositories. Additionally, the acquired data had undergone augmentation procedures to enhance diversity and improve the model's ability to generalize.

2. Data Preprocessing: we implemented diverse data augmentation techniques on knee X-ray images for training dataset enrichment. Rotation, brightness adjustment, horizontal and vertical flips, and zoom are applied to introduce variability. A data balancing function addresses class imbalance by generating augmented images for each class. For contrast enhancement, histogram equalization is employed in the preprocessing step. The loaded datasets undergo shuffling and prefetching, optimizing data loading for efficient model training.

3. Training and Evaluation: The preprocessed dataset was then split into three subsets: training, validation, and testing. Following the preprocessing of knee X-ray images, we constructed a Convolutional Neural Network (CNN) using the ConvNeXtBase architecture with added layers, such as batch normalization and dropout, to enhance its classification capabilities.

To optimize training, a custom learning rate schedule was implemented for both dataset 1 and dataset 2, featuring a warm-up phase and cyclic patterns, promoting efficient convergence during the training process, after trying with many loss functions. Essential callbacks employed to improve training outcomes and prevent overfitting. The compiled model, utilizing the Adam optimizer, and experimented with different loss functions. Before and after augmentation we tried with weighted CE (Cross Entropy), Focal CE, Class Balanced CE, Class Balanced Focal CE, Categorical CE loss functions, and accuracy metric, underwent training on an augmented dataset of knee X-ray images.

A meticulous fine-tuning process was conducted, further optimizing the model to attain the highest accuracy. The training process spanned multiple epochs with multiple models, out of them ConvNextBase achieved highest accuracy of 70% with Weighted CE Loss function on original dataset. We have the diagram of custom learning rate below (we implemented same custom LR for both datasets, since it adapted well to our model)

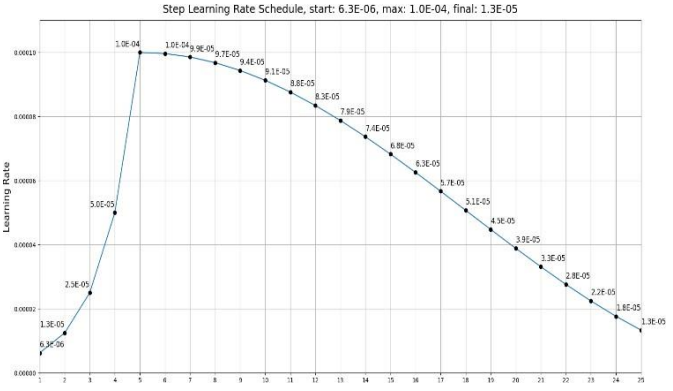


Fig: Custom Learning Rate Curve

Results of Dataset 1

Model	Accuracy	Precision	Recall	F1 score
DenseNet121	57	54	57	52
VGG16	62	62.6	62	61
InceptionV3	46	46.2	46	45.7
VIT-B16	48	48	48	48
EfficientNet-V1B0	65	65	65	65
ResNet101	64	64	64	64
ConvNextBase	**66**	**66**	**66**	**66**

Table: Dataset 1 (Before Augmentation i.e., Original Data)

Model	Accuracy	Precision	Recall	F1 score
DenseNet121	62	67	62	63
VGG16	63	66	66	66
InceptionV3	47.2	47.5	47.2	47.2
VIT-B16	48	48.3	48	48.2
EfficientNet-V1B0	65	64	63	62
ResNet101	63	66	66	66
ConvNextBase	**66.3**	**66.1**	**64.2**	**65.7**

Table: Dataset 1 (After Augmentation)

Model	Loss Function	Accuracy	Precision	Recall	F1 score
ConvNextBase	**Weighted CE**	**70**	**70**	**71**	**70**
ConvNextBase	Weighted Focal CE	69	68	69	67
ConvNextBase	Categorical CE	66	66	66	66

Table: Dataset 1 with different Loss Functions (Original Data)

From the results, we achieved the highest accuracy of 70% with ConvNextBase Model with Weighted CE, below data gives more details.

The classification report is as follows:

	precision	recall	f1-score	support
0	0.74	0.84	0.79	639
1	0.42	0.23	0.30	296
2	0.68	0.71	0.69	447
3	0.79	0.82	0.80	223
4	0.89	0.92	0.90	51
accuracy			0.70	1656
macro avg	0.70	0.71	0.70	1656
weighted avg	0.67	0.70	0.68	1656

Dataset 2:

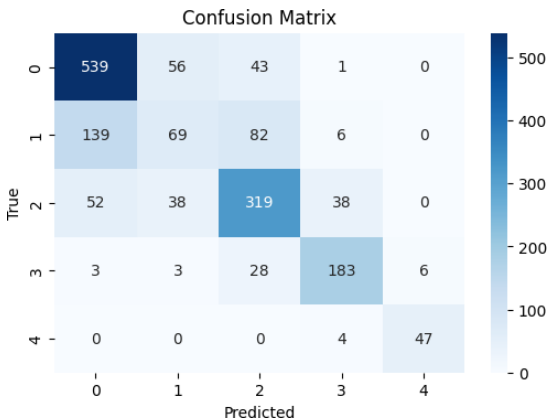
Experiment 2:

1. Data Collection: The dataset2 was meticulously curated from the Kaggle data platform, representing a valuable resource for model training. Despite its smaller size, this dataset is characterized by superior image quality, curated by experts in the medical field. Notably, the x-ray images of knee joints within this dataset are unique, with no augmentation applied, ensuring authenticity and reliability. However, the dataset presented an imbalance in class distribution, prompting us to address this challenge during model development. The discovery of this dataset was a result of extensive research, including thorough examination of articles and research papers in the medical domain. The deliberate selection of dataset2 enhances the diversity of our training data and contributes to the robustness of our model's performance.

2. Data Preprocessing: In our efforts to enhance the knee X-ray image dataset for model training, we applied various advanced data augmentation techniques. This included rotation, brightness adjustment, flips, and zoom to introduce diversity. To address class imbalance, we implemented a data balancing function, generating augmented images for each class. For refined contrast, we used histogram equalization during preprocessing.

In our experimentation, we tested Gaussian filters, resizing, grayscale conversion, and normalization. However, histogram equalization emerged as the most effective choice for our dataset. This meticulous preprocessing ensures dataset augmentation and optimizes the training process, aligning with deep learning best practices.

3. Training and Evaluation: After the dataset preprocessing, we strategically divided it into training, validation, and testing subsets. Leveraging the ConvNeXtBase architecture, our Convolutional Neural Network (CNN) incorporated additional layers for improved



classification, including batch normalization and dropout.

To enhance training efficiency, a custom learning rate schedule and various loss functions were meticulously tested. Despite initial attempts with default settings, our model's performance significantly benefited from these tailored adjustments.

Throughout the experimentation process, we explored different transfer learning models and diverse loss functions, ultimately achieving optimal results with ConvNextBase, attaining a peak accuracy of 84%. Despite initial exploration of custom architectures (CNN) and YOLOv8 for joint extraction, their performance fell short of expectations, leading us to prioritize the original dataset and employ state-of-the-art transfer learning models mainly ConvxNextBase. The below figures give more information about dataset 2 accuracies (we have used only augmented dataset for our implementation.)

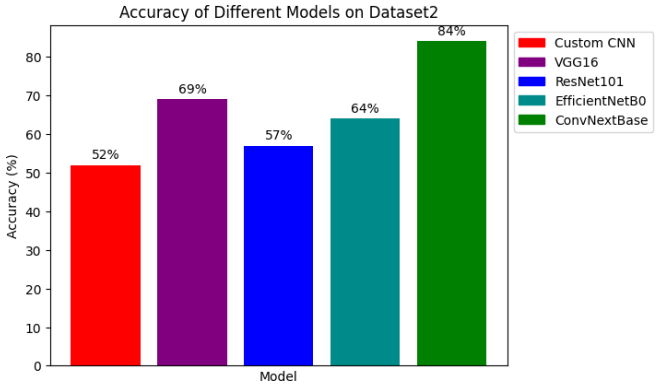
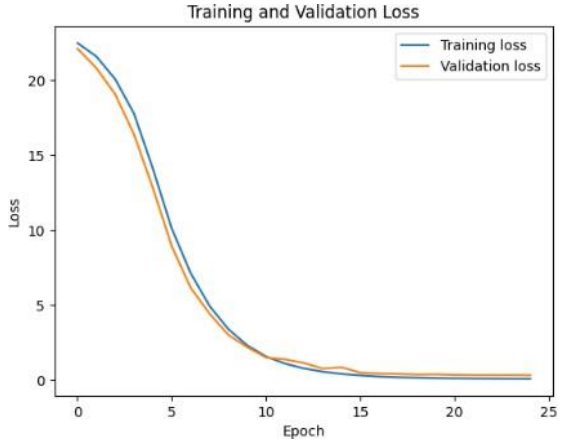


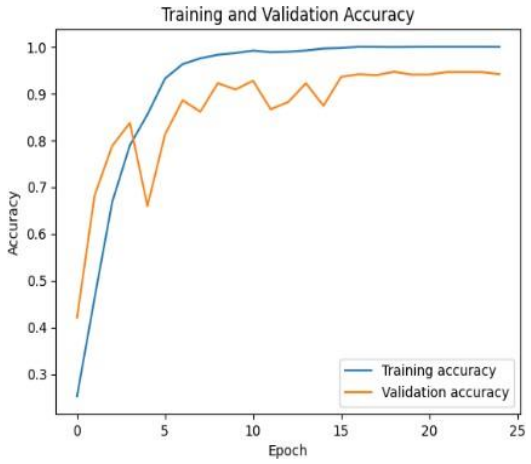
Fig: Dataset 2 accuracies

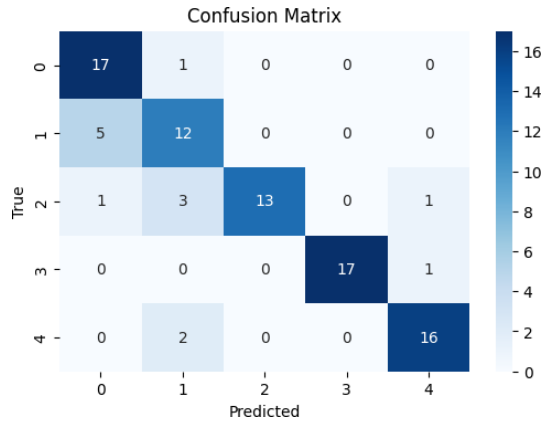
Model	Loss Function	Accuracy	Precision	Recall	F1 score
ConvNextBase	Weighted CE	76	76	75.3	76
ConvNextBase	Weighted Focal CE	83	85	83	83
ConvNextBase	Categorical CE	84	86	84	84
ConvNextBase	Focal Loss	84	85	84	84

Table: Dataset 2 with different loss functions

From the results, we achieved the highest accuracy of 84% with ConvNextBase Model. below data gives more details.







The classification report is as follows:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.94      0.83        18
           1       0.67      0.71      0.69        17
           2       1.00      0.72      0.84        18
           3       1.00      0.94      0.97        18
           4       0.89      0.89      0.89        18

    accuracy                           0.84        89
   macro avg       0.86      0.84      0.84        89
weighted avg       0.86      0.84      0.84        89
```

Our experimental results affirm the efficacy of the implemented deep learning models for knee osteoarthritis (OA) classification. The ConvNeXtBase model, achieving a notable accuracy of 84%, stands out as a robust performer, emphasizing the potential of advanced convolutional neural networks for accurate knee OA image classification.

Moreover, it's noteworthy that the ConvNeXtBase model achieved this high accuracy with both focal loss and categorical cross-entropy (CE) loss functions, demonstrating its versatility and effectiveness in handling different loss functions

To validate the accuracy of our model, we conducted K-fold cross-validation on the training data (5-fold). Successfully training the model on each fold, we achieved an impressive average accuracy of 87% across all folds. This approach provides a more reliable estimate of the model's performance.

VII. LIMITATIONS AND FUTURE WORK

1. Dataset Size and Imbalance: The relatively small size of the dataset, particularly in Dataset 2, may impact the model's ability to generalize effectively. Additionally, addressing class imbalance, especially in Dataset 1, could be challenging despite data augmentation efforts.

2. Model Complexity: While state-of-the-art models like ConvNeXtBase demonstrated high accuracy, their inherent complexity might pose challenges in terms of interpretability and computational requirements.

3. Joint Extraction Attempts: The exploration of YOLOv8 for joint extraction, though intriguing, did not yield the expected improvements. Further investigations into advanced joint localization methods might be warranted.

4. Generalization to Diverse Populations: The dataset's source and composition might limit the model's generalizability across diverse demographic groups or medical conditions not well-represented in the current data.

5. Fine-Tuning Sensitivity: While fine-tuning significantly improved model performance, its sensitivity to hyperparameter tuning and potential overfitting could be a concern. A more extensive hyperparameter search or regularization techniques might be explored.

6. Clinical Interpretability: The challenge of interpreting complex CNN models in a clinical context may limit their direct application without thorough validation and collaboration with medical experts

To address these limitations, future work could focus on the following:

- Augment Dataset 2 with diverse samples and collaborate for high-quality knee X-ray images
- Explore lightweight CNNs and distillation for interpretability while maintaining high performance.
- Implement advanced joint localization beyond YOLOv8, exploring models like RetinaNet or EfficientDet.
- Collaborate for diverse knee X-ray images, ensuring representation of demographics and medical conditions.
- Systematically fine-tune hyperparameters, employing grid or random search, and implement advanced regularization.
- Collaborate with medical professionals, seeking feedback on decisions, and integrate explainability techniques like attention maps.

VIII. CONCLUSION

This paper presents an innovative knee osteoarthritis (OA) classification system using deep learning and data augmentation techniques. The ConvNeXtBase model emerged as the most accurate, achieving 84% accuracy, demonstrating promising outcomes. Despite these achievements, challenges persist, particularly in dataset imbalance and model interpretability. Future efforts will focus on addressing these limitations, aiming to enhance the system's accuracy and applicability.

In terms of the software process model, our project adopted a hybrid approach combining Agile and Rapid Application Development (RAD) methodologies. This allowed for flexibility, collaboration, and efficient iteration, resulting in the successful development of a robust knee OA classification system that caters to evolving project needs and delivers significant value to users.

IX. REFERENCES

[1] https://data.mendeley.com/datasets/56rmx5bjcr/1

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7704420/

[3] https://www.mdpi.com/2075-4418/13/8/1380

[4] https://arxiv.org/ftp/arxiv/papers/2207/2207.12521.pdf

[5] 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]

[6] Saleem, M.; Farid, M.S.; Saleem, S.; Khan, M.H. X-ray image analysis for automated knee osteoarthritis detection. Signal Image Video Process. 2020, 14, 1079–1087. [Google Scholar] [CrossRef]

[7] Lim, J.; Kim, J.; Cheon, S. A deep neural network-based method for early detection of osteoarthritis using statistical data. Int. J. Environ. Res. Public Health 2019, 16, 1281. [Google Scholar] [CrossRef][Green Version]

[8] von Tycowicz, C. Towards shape-based knee osteoarthritis classification using graph convolutional networks. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 750–753. [Google Scholar]

[9] Wang, Y.; Li, S.; Zhao, B.; Zhang, J.; Yang, Y.; Li, B. A ResNet-based approach for accurate radiographic diagnosis of knee osteoarthritis. CAAI Trans. Intell. Technol. 2022, 7, 512–521. [Google Scholar] [CrossRef]

[10] Jakaite, L.; Schetinin, V.; Hladůvka, J.; Minaev, S.; Ambia, A.; Krzanowski, W. Deep learning for early detection of pathological changes in x-ray bone microstructures: Case of osteoarthritis. Sci. Rep. 2021, 11, 1–9. [Google Scholar] [CrossRef]

[11] Thomas, K.A.; Kidziński, Ł.; Halilaj, E.; Fleming, S.L.; Venkataraman, G.R.; Oei, E.H.; Delp, S.L. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. Radiology. Artif. Intell. 2020, 2, e190065. [Google Scholar]

[12] Hunter, H.; Ryan, M.S. Knee Osteoarthritis-Statpearls-NCBI Bookshelf. (4 August 2019). Available online: https://www.ncbi.nlm.nih.gov/books/NBK507884/

[13] Kellgren, J.H.; Lawrence, J.S. Radiological Assessment of Osteo-Arthrosis. Ann. Rheum. Dis. 1957, 16, 494–502. [Google Scholar] [CrossRef] [PubMed] [Green Version]

[14] 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] [PubMed]

[15] Kasani, P.H.; Park, S.W.; Jang, J.W. An aggregated-based deep learning method for leukemic B-lymphoblast classification. Diagnostics 2020, 10, 1064. [Google Scholar] [CrossRef] [PubMed]

[16] https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity

[17] https://www.kaggle.com/datasets/tommyngx/digital-knee-xray/data?select=MedicalExpert-I