2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)
12-13 September, RUET, Rajshahi, Bangladesh

2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON) | 979-8-3315-1798-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/PEEIACON63629.2024.10800349

Classification of X-ray Images for the Automated Severity Grading of Knee Osteoarthritis by Ensemble Learning through EfficientNet Architectures with Grad-CAM Visualization

1st Sajid Fardin Dipto, 2nd Md. Omaer Faruq Goni
[1,2]*Department of Electrical & Computer Engineering*
[1,2]*Rajshahi University of Engineering & Technology,*
Rajshahi-6204, Bangladesh
sajiddipto10@gmail.com[1], omaerfaruq@ece.ruet.ac.bd[2]

***Abstract*—Knee osteoarthritis (KOA) is a degenerative form of arthritis in the knee joints commonly occurring in old-aged individuals which can cause extreme pain, deterioration of bone joints, joint membranes or ligaments, and irregularity in the knee joints. Being affected by KOA, individuals experience limited social interaction and diminished quality of life. The KOA can be assessed using the Kellgren-Lawrence (KL) grading method. In this study, an ensemble model was developed to predict a KL grading reliably and precisely. Several EfficientNet architectures were implemented for the ensembling of this study. Among them, an ensemble of EfficientNetB0 and EfficientNetB4 pre-trained models gave satisfactory results compared to other deep learning or ensembling models. Here, the proposed model underwent evaluation utilizing the Osteoarthritis Initiative (OAI) dataset which contains X-ray radiographs. The dataset was split into training, testing, and validation sets with percentages of 75%, 17%, and 8% respectively. The overall multiclass classification accuracy was 96.03% with an F1-score of 93%, specificity of 98.56%, and recall of 93%. The proposed model achieved an optimistic result compared to the transfer learning models and state-of-the-art models. Gradient-Weighted Class Activation Mapping was also applied for the radiographs to implement Explainable AI for generating visual explanations for the predictions made by the deep learning networks which revealed the model's potential for grading the severity of KOA.**

***Index Terms*—Knee Osteoarthritis (KOA), Deep Learning, Osteoarthritis Initiative Dataset, Kellgren-Lawrence Grading, EfficientNet, Grad-CAM**

I. Introduction

KOA is a prevalent disease among elderly people which generally occurs due to the gradual deterioration of the cartilage of knee joints. Cartilage is a robust and flexible connective tissue present in healthy knee joints. Cartilage improves bone resilience and gives support to bony areas [1]. The degeneration of the joints that destroys cartilage is known as osteoarthritis. The KL grading method, officially recognized by the World Health Organization (WHO), is utilized to evaluate KOA based on the severity of the condition. Here, grades range from 0 (normal) to 4 (severe) where grades 1,2,3 are doubtful, mild, and moderate, respectively [2]. KL grading is directly related to osteophytes and sclerosis. Osteophytes are bony projections and sclerosis is the abnormal hardening or thickening of bone tissues. Most commonly KOA is associated with aging, but obesity, bone density, work-related injuries, lack of exercise, and gender factors can be also responsible [3].

The global occurrence of osteoarthritis symptoms stands at 18% among women over 60 years old and among men, it is 9.6%. Around 80% of the KOA patients experienced limitations in movement, with 25% unable to carry out their essential daily tasks [4]. Globally, the occurrence of KOA affects 16% of individuals who are 15 years old and above with a higher occurrence of 22.9% observed among 40 years old and above. Approximately 654,100,000 patients with KOA were 40 years old and above in the year 2020 worldwide [5].

The accuracy of KL grading can vary and can be low depending upon the expertise of the clinician as well as also experience and time. Manual diagnosis in the hospital, segmentation of the knee, and annotation of the knee images represent one of the most frequently utilized procedures by clinicians. These methods are time-consuming and also highly dependent upon the patient variation. To overcome these limitations, machine learning (ML) and deep learning (DL) approaches have been applied in various works. However, there exist some limitations in the works like the inability to evaluate non-image clinical data with image data, multi-modal fusion data, implementing ensemble learning, applying explainability for the models for visual interpretations, and so on. The outcomes of this study can be described as,

1) Severity of KOA was assessed through ensembling DL models to speed up the diagnosis process.
2) Necessary preprocessing techniques such as Data Augmentation, Resize, Histogram Equalization, Bilateral Filtering, Rescaling, etc. were applied to enhance model performance.
3) The results of the developed model were compared with transfer learning (TL) models and state of arts.
4) The model consists of fewer parameters compared to some of the individual TL models.

Fig. 1. KL Grading System of KOA [6]

II. Methodology

In this paper, Data partitioning was applied for splitting the data. After observing the performance of each model in the EfficientNet family individually, this study also examined other well-known TL and ensemble models. All these results were compared to identify the optimal approach, which was the ensemble of EfficientNetB0 and EfficientNetB4. The proposed ensemble model's results were explained visually using an Explainable AI (XAI) technique called Grad-CAM. Fig. 2 shows the proposed methodology of this study graphically with block diagrams which includes taking X-ray images and applying necessary preprocessing techniques. After preprocessing, the training and validation images were given to the EfficientNet architectures for feature extraction. Then, the extracted features were used in a set of dense layers with tuned hyperparameters, and finally, the test data were evaluated for KOA severity grading with explainability.

A. Dataset Description

The OAI dataset was taken from the Mendeley Data website for this study [6]. The main challenge in this dataset was having many overexposed images. Only after removing them, the proposed model in this work performed well. The dataset contains 9786 X-ray radiographs. Among them, 3857 are grade 0 (healthy) images. The number of images for grades 1, 2, 3, and 4 are the remaining 5929 images. As for data partitioning, the whole dataset was divided into training, testing, and validation sets with percentages of 75%, 17%, and 8% respectively for the best performance. Some samples from the OAI dataset have been presented in Fig. 1.

B. Preprocessing

Data partitioning was applied taking data from the OAI dataset at first and then they were converted into a shape of (224, 224, 3) to match the input shape of EfficientNet architectures. Histogram equalization for contrast adjustment and bilateral filtering for smoothness were applied initially. Afterward, for the proposed model of this study, only vertical and horizontal flipping were applied to increase the dataset. Data augmentation was applied only for the training data to enhance the model generalization but to preserve the integrity of validation and test datasets. This way, the accuracy can be improved. The total images for training were 7304 which increased to a total of 29,216 images after applying data augmentation. So, training images were 4 times the initial amount after data augmentation keeping the number of test and validation data unchanged. Grayscale images were converted to RGB format because EfficientNet architectures expect images to be in RGB format, that is, 3-channel format. Rescaling and some built-in preprocessing of the EfficientNet architecture were also applied before inputting the images into the proposed ensemble model.

C. Architecture of Proposed Model

The proposed hybrid model is represented graphically in Fig. 3 which includes the ensemble model that was developed using the stacking method. This method [7] involved combining two EfficientNet models by taking their outputs and feeding them into a custom neural network with dense layers to make a final prediction. The custom layers consist of several Dense layers, each followed by Batch Normalization and interspersed with Dropout layers in two cases. Specifically, the sequence includes a Dense layer with 512 neurons and a following Dropout rate of 40%, then a 256 neurons Dense layer with a 30% Dropout which progresses to layers with 128 and 64 neurons. The inclusion of L2 regularization, batch normalization, and Dropout ensures that the model remains generalizable and resistant to overfitting [8].

D. Performance Metrics

Various statistical parameters serve to assess the performance of classifiers. A confusion matrix along with metrics like Accuracy (Acc), F1-score (F1), Precision (Pre), Recall (Rec), Specificity (Spec), and Cohen's kappa value were used in this work as performance metrics. The positive class is accurately classified by a model if it is True Positives (TP). The True Negatives (TN) occur when the negative class is correctly classified. False Positives (FP) occur when a model erroneously classifies the negative class as positive, while False Negatives (FN) occur when a model fails to identify the positive class. The mathematical expressions of these are given below.

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$Pre = \frac{TP}{(TP + FP)} \tag{2}$$

$$Rec = \frac{TP}{(TP + FN)} \tag{3}$$

$$F1 = \frac{2(Pre)(Rec)}{(Pre + Rec)} \tag{4}$$

$$Spec = \frac{TN}{(TN + FP)} \tag{5}$$

$$Cohen's\ Kappa = \frac{(P_0 - P_e)}{(1 - P_e)} \tag{6}$$

Where,

$$P_0 = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{7}$$

$$P_e = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + TN + FP + FN)^2} \tag{8}$$

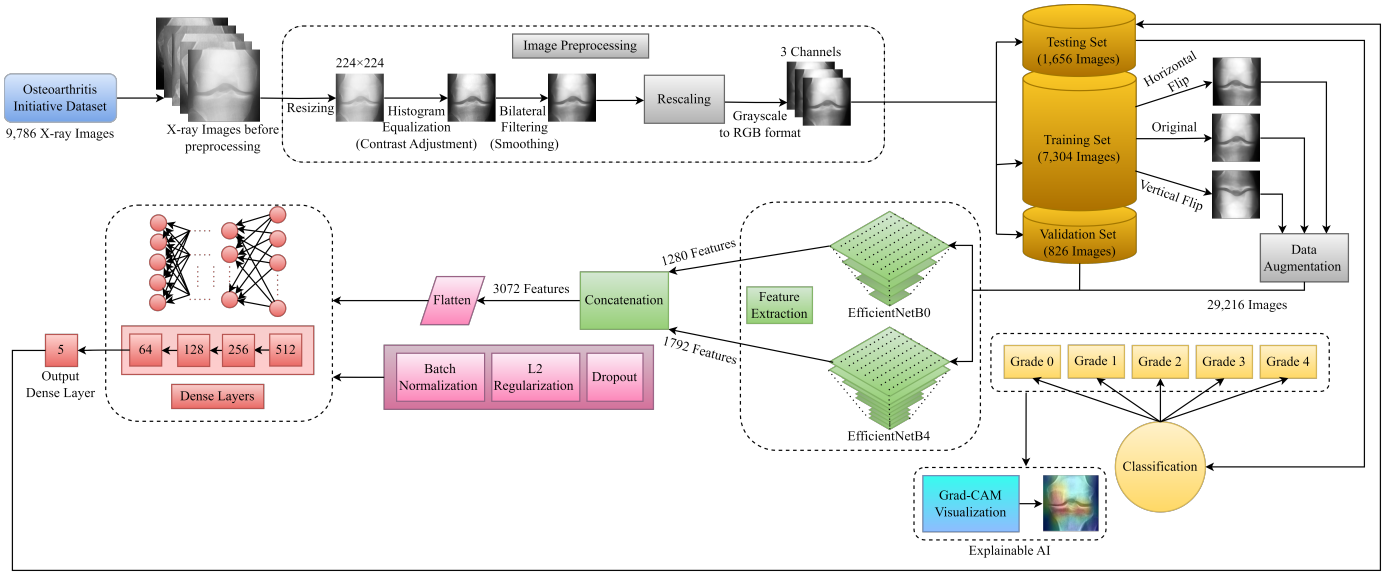

Fig. 2. Proposed Methodology of the Ensemble Model

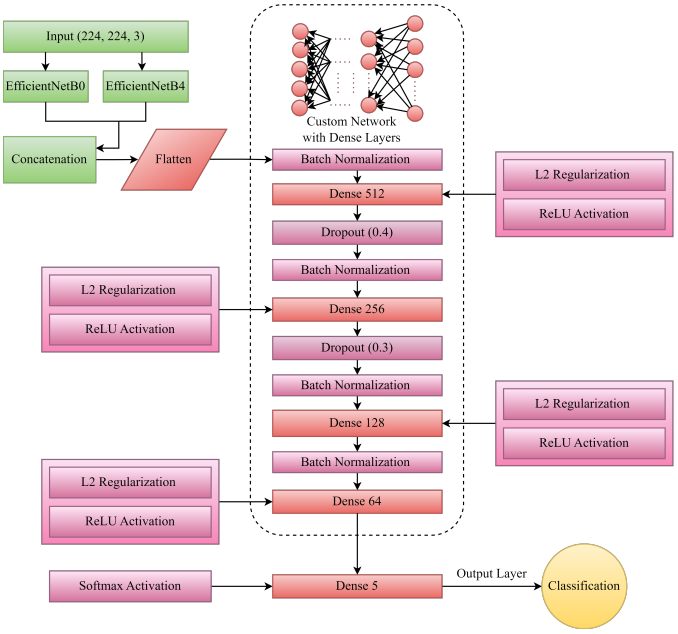

Fig. 3. Network Architecture of the Developed Ensemble Model

E. Explainable AI

XAI refers to methods in the application of artificial intelligence technology that allow humans to comprehend and trust the results and output created by ML algorithms [9]. For this particular task, the visualization technique used in this study was Grad-CAM visualization. It is a technique that aims to increase the interpretability of the convolutional neural network (CNN) through visualization [10]. This technique aids in depicting the exact parts in an image influencing the decision-making of the ML model by pointing out crucial sections responsible for the output of the model.

III. Results and Discussion

We performed this study on Kaggle to take advantage of its computational resources and the availability of datasets. In the following sections, the results and key findings are discussed.

A. Comparative Analysis of Proposed Model with Other TL Models

The hyperparameters of the proposed ensemble model were tuned by the trial-and-error method. L2 regularization, batch normalization, and dropout rate were used for a more robust structure. The Adam optimizer was used for the proposed model, and the total epochs were 150. The learning rate scheduler and ReduceLROnPlateau were used for handling the learning rate with an initial value of 0.001. The loss function used for the model was categorical cross-entropy. The developed ensemble model achieved a multiclass accuracy of 96.03%. The model was also evaluated on other metrics.

Table II represents the comparative analysis of the proposed model with the others. It achieved an average F1-score of 93% with both precision and recall being 93% on average. The ensemble model achieved the highest area under the curve (AUC) of 1.00 for class 3 and 4. The proposed model achieved Cohen's kappa of 90.95%. The confusion matrix is shown in Fig. 4. Fig. 5 shows the ROC curves of the five classes. The proposed ensemble model showed better performance compared to other models because of multiple preprocessing techniques that boosted its effectiveness. The proposed model consists of 23.485M parameters. Some of the individual TL models used in the state of arts have larger parameters than the proposed model which is shown in Table I. After comparing the mentioned TL models with the proposed model in this study, it is clear that the accuracy and other performance metrics of the proposed model are very satisfactory, in fact, better than that of all the approaches here.

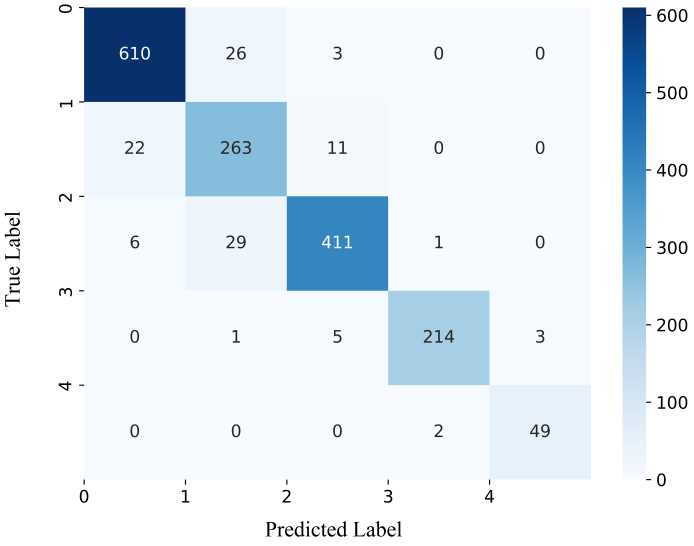

Fig. 4. Confusion Matrix of the Proposed Ensemble Model

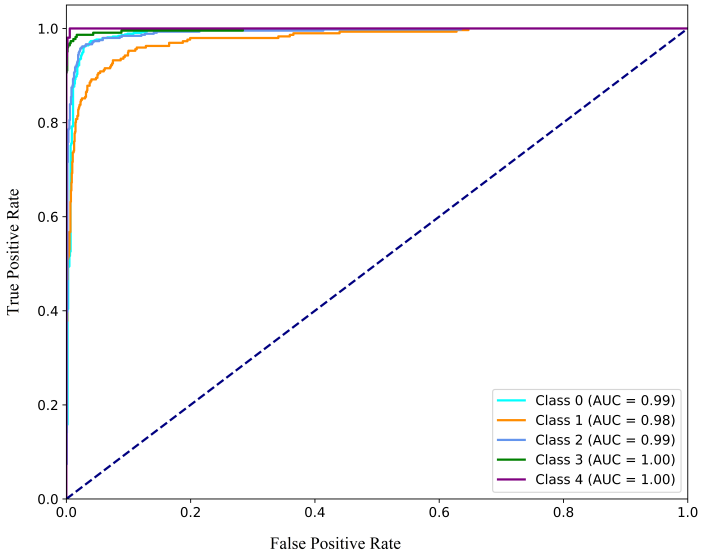

Fig. 5. ROC Curves of the Proposed Ensemble Model

B. XAI: Grad-CAM Visualization

Grad-CAM was applied to enhance the explainability of the model. Grad-CAM played an important role by revealing the specific regions of the images that helped in guiding toward the final decision. It emphasized the crucial areas that affected the model's output. Fig. 6 shows the Grad-CAM visualizations of knee osteoarthritic X-ray images. From these visualizations, it is evident that the middle portion of a knee image is highlighted in red on the heatmap. This indicates that this specific region is crucial for distinguishing between a healthy knee and an osteoarthritic knee.

TABLE I
PARAMETERS COMPARISON OF THE PROPOSED MODEL WITH OTHER TL MODELS

Models	Parameters
ResNet101	44.50M
DenseNet161	28.70M
EfficientNetB5	30.40M
VGG16	138.60M
VGG19	143.70M
ResNet50	25.64M
InceptionV3	23.90M
InceptionResNetV2	55.80M
Proposed Model	23.49M

TABLE II
COMPARATIVE ANALYSIS OF PROPOSED MODEL WITH OTHER TL MODELS

Model	Acc	Pre	Rec	F1	Spec
EfficientNetB0	93.22	90.00	90.00	90.00	97.30
EfficientNetB4	88.06	85.00	84.00	85.00	95.78
EfficientNetB7	89.00	89.00	89.00	89.00	96.74
MobileNetV2	90.94	88.00	87.00	87.11	96.45
ResNet50	87.52	83.00	82.00	82.00	95.14
DenseNet201	87.75	85.38	84.31	85.34	95.85
VGG16	75.29	15.04	39.10	21.00	80.00
InceptionV3+MobileNetV2	90.17	88.00	87.00	87.00	96.49
ResNet50+DenseNet201	88.70	86.00	84.00	85.01	95.90
Eff.NetB0+DenseNet201	93.93	91.00	91.00	91.00	97.46
Eff.NetB0+InceptionV3	93.76	91.00	91.00	91.00	97.33
Proposed Model	96.03	93.00	93.00	93.00	98.56

C. Comparative Analysis of Proposed Model with Previous Works

The proposed ensemble model showed better performance compared to previous models, as detailed in Table III. As the proposed model in this work was evaluated using five classes, this model has also been compared with other works based on five classes.

Models developed by Upadhyay et al. [11] and Abd El-Ghany et al. [12] previously held the highest metrics in this area, as shown in the comparative analysis in Table III. However, our proposed model outperforms these, establishing it as a more effective solution for KOA severity grading. Pi et al. [13] applied ensemble learning in their study and also implemented grad-CAM visualization as XAI. They used six-model-based and eight-model-based networks which are very complex and have high computational costs. But in this study, the ensemble model is not that complex, rather it is far simpler. Also, their accuracy and other performance metrics are comparatively lower than the proposed model. In case of parameters, the proposed model of this study has 23.485M parameters. But in the paper of Pi et al. [13], their eight-model-based network had a total parameter of 265.8M and the optimized six-model-based network had 168.2M parameters. Also, from the comparative analysis, it is clear that the proposed model achieved higher accuracy, precision, recall, and F1-score compared to other state of arts along with visual explainability.

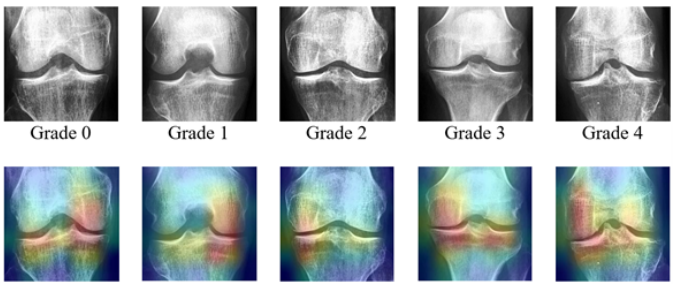

Fig. 6. Grad-CAM Visualization of Image Samples from OAI Dataset

TABLE III
COMPARATIVE ANALYSIS OF PROPOSED MODEL WITH PREVIOUS WORKS

Ref.	Year	Models	Acc	Pre	Rec	F1
Cueva et al. [14]	2022	CADx, ResNet34	61.00	65.80	61.60	-
Ahmed et al. [15]	2022	CNN, PCA, SVM	74.57	76.80	75.40	74.80
Navale et al. [16]	2022	CAD, CNN	81.41	81.00	83.00	82.00
Ribas et al. [17]	2022	Complex Network	81.69	-	77.60	-
Yunus et al. [18]	2022	YOLOV2, ONNX	90.60	85.00	91.00	88.00
Sharma et al. [19]	2023	EfficientNetB5	93.84	87.80	85.70	86.20
Upadhyay et al. [11]	2023	Deep CNN	95.00	-	-	-
Abd El-Ghany et al. [12]	2023	Fine-tuned DenseNet169	95.93	85.80	88.80	87.08
Mohammed et al. [20]	2023	ResNet101	69.00	67.00	67.00	65.00
Pi et al. [13]	2023	Ensemble of 6 TL Models	76.93	78.80	75.30	76.65
Proposed Model	2024	EfficientNetB0, EfficientNetB4	96.03	93.00	93.00	93.00

IV. CONCLUSION

In this study, a hybrid ensemble model is introduced using EfficienNet architectures which was designed to classify the severity grading of KOA. The OAI dataset was used for this study because it is the most popular and available dataset online and most of the state-of-arts used this dataset for evaluating KL grading of KOA. The hybrid model unfolded in two key steps. First, the input images were fed to the EfficientNet architectures and then features were extracted for concatenation. After that, a custom dense layer model was developed for further improvements of the predictions made by the algorithm. The model consists of 23.485M parameters. Even when considering some of the individual TL models, the ensemble model of this study has fewer parameters. However, recognizing that the ensemble of multiple classifiers can yield superior results, the proposed ensemble model achieved a multiclass classification accuracy of 96.03% with precision, recall, F1 score, specificity, and Cohen's kappa value of 93%, 93%, 93%, 98.56%, and 90.95% respectively, for KOA severity grading considering the KL grading system which was satisfactory. This ensemble method excels in classification tasks by combining the deep pattern recognition of EfficientNetB4 with the efficiency and overfitting resistance of EfficientNetB0, outperforming other models. Additionally, this model was introduced with Grad-CAM visualization for better explainability. Therefore, the proposed ensemble model proves to be a reliable method for the accurate grading of KOA with visual explainability.

The noisy dataset had a big impact on the model's performance and its ability to make accurate predictions. The future plan of this study includes working with experts in medical imaging who will help this study obtain a refined and improved dataset. Additionally, collaborating with healthcare institutions will enable the collection of a more diverse and cleaner dataset. The future plan also includes solving the issues of dataset imbalances and working with various types of radiographs like MRI and CT scans in addition to X-ray radiographs through multimodal fusion as well as clinical non-image data. Developing the model into an Android application, a web-based application, or a hardware application as an embedded system will be the ultimate step for this study.

REFERENCES

[1] "Overview of cartilage, available link: https://my.clevelandclinic.org/health/body/23173-cartilage, last access: [16-04-2024]."
[2] D. Schiphof, M. Boers, and S. M. Bierma-Zeinstra, "Differences in descriptions of kellgren and lawrence grades of knee osteoarthritis," *Annals of the rheumatic diseases*, vol. 67, no. 7, pp. 1034–1036, 2008.
[3] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 84–92, 2019.
[4] T. Neogi, "The epidemiology and impact of pain in osteoarthritis," *Osteoarthritis and cartilage*, vol. 21, no. 9, pp. 1145–1153, 2013.
[5] A. Cui, H. Li, D. Wang, J. Zhong, Y. Chen, and H. Lu, "Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies," *EClinicalMedicine*, vol. 29, 2020.
[6] "Knee osteoarthritis severity grading dataset, available link: https://data.mendeley.com/datasets/56rmx5bjcr/1, last access: [16-04-2024]."
[7] "Stacking to improve model performance, available link: https://shorturl.at/gtNP2, last access: [16-04-2024]."
[8] "Regularization: Batch-normalization and drop out, available link: https://shorturl.at/ejCG6, last access: [16-04-2024]."
[9] "What is explainable ai?, available link: https://shorturl.at/iqvBV, last access: [16-04-2024]."
[10] "Visualizing model insights: A guide to grad-cam in deep learning, available link: https://shorturl.at/zEL08, last access: [16-04-2024]."
[11] A. Upadhyay, O. Sawant, and P. Choudhary, "Detection of knee osteoarthritis stages using convolutional neural network," *SN Computer Science*, vol. 4, no. 3, p. 257, 2023.
[12] S. Abd El-Ghany, M. Elmogy, and A. Abd El-Aziz, "A fully automatic fine tuned deep learning model for knee osteoarthritis detection and progression analysis," *Egyptian Informatics Journal*, vol. 24, no. 2, pp. 229–240, 2023.
[13] S.-W. Pi, B.-D. Lee, M. S. Lee, and H. J. Lee, "Ensemble deep-learning networks for automated osteoarthritis grading in knee x-ray images," *Scientific Reports*, vol. 13, no. 1, p. 22887, 2023.
[14] J. H. Cueva, D. Castillo, H. Espinós-Morató, D. Durán, P. Díaz, and V. Lakshminarayanan, "Detection and classification of knee osteoarthritis," *Diagnostics*, vol. 12, no. 10, p. 2362, 2022.
[15] S. M. Ahmed and R. J. Mstafa, "Identifying severity grading of knee osteoarthritis from x-ray images using an efficient mixture of deep learning and machine learning models," *Diagnostics*, vol. 12, no. 12, p. 2939, 2022.

[16] D. I. Navale, D. D. Ruikar, D. D. Sawat, P. M. Kamble, K. V. Houde, and R. S. Hegadi, "Automatic knee osteoarthritis stages identification," in *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pp. 53–60, Springer, 2021.
[17] L. C. Ribas, R. Riad, R. Jennane, and O. M. Bruno, "A complex network based approach for knee osteoarthritis detection: Data from the osteoarthritis initiative," *Biomedical Signal Processing and Control*, vol. 71, p. 103133, 2022.
[18] U. Yunus, J. Amin, M. Sharif, M. Yasmin, S. Kadry, and S. Krishnamoorthy, "Recognition of knee osteoarthritis (koa) using yolov2 and classification based on convolutional neural network," *Life*, vol. 12, no. 8, p. 1126, 2022.
[19] G. Sharma, V. Anand, and V. Kumar, "Classification of osteo-arthritis with the help of deep learning and transfer learning," in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 446–452, IEEE, 2023.
[20] A. S. Mohammed, A. A. Hasanaath, G. Latif, and A. Bashar, "Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed x-ray images," *Diagnostics*, vol. 13, no. 8, p. 1380, 2023.