Enhanced Osteoarthritis Classification through Transfer Learning and Hyperparameter-Tuned Multi-Layer Ensemble Models

Abstract—**Knee osteoarthritis (OA) significantly limits activity and causes physical disability in older adults. Classifying OA early is crucial to slowing down its progression. This paper introduces a method for OA classification using a transfer learning fusion network with parameter tuning and Multi-Layer Ensemble. The approach starts by balancing the dataset through data augmentation. Then combines a pre-trained model with a customized CNN model. Various model performances are combined through Multi-Layer Ensemble, resulting in superior performance with 75.73% accuracy. Our proposed method produced a good-performing model in OA classification and overcoming prior constraints, emphasizing the importance of automated knee OA classification and providing an effective solution.**

Index Terms—**Knee Osteoarthritis (OA) Classification, Transfer Learning Fusion (TLF), Machine Learning (ML) and Multi-Layer Ensemble (MLE).**

I. Introduction

Research on osteoarthritis (OA) pathology has been crucial due to its substantial economic impact, causing disability, and pain, and affecting the patient's lifestyle. OA involves more than just anatomical and physiological changes, extending to cellular stress and the degradation of the cartilage matrix [1]. Generally, OA is associated with aging. However, there are other risk factors namely obesity, lack of exercise, genetic predisposition, bone density, occupational injury, trauma, and gender [2], [3]. OA affects nearly 240 million people worldwide [4]. According to the World Health Organization (WHO), by 2050, approximately 20% of the world's population will be over 60 years old. Of that percentage, 15% will have symptomatic OA, and one-third of these people will be severely disabled [1] to [4]. Knee Osteoarthritis (KOA) is the most prevalent type, especially in older adults, leading to symptoms like chronic pain, stiffness, muscle weakness, and difficulty in daily activities [5]. Current OA diagnosis relies on physical examination and imaging techniques like X-rays, MRI scans, and arthroscopy. Efficient automatic knee OA detection is crucial to address these challenges.

Our paper structure is as follows: Section II reviews the literature. Section III provides an overview of the dataset, and Sections IV-A to IV-C detail our research approach, data preprocessing, Proposed Transfer Learning (TL) Architecture, and Justification of Our Procedural Architecture. Section V analyzes experimental results, while Sections V-B to V-C delve into Evaluation Metrics, Experimental Setup, Result Assessment. Sections VI and VII contain discussions, comparisons, and considerations of potential result validity threats. Finally, Section VIII summarizes findings, and in Section IX, we suggest future research directions, followed by references.

II. Literature Review

Several research endeavors have employed a variety of models in their investigations. For example, Christos Kokkotis et al [6] incorporated Support Vector Machines in their research. Specific studies utilized Convolutional Neural Network (CNN) models referenced by [3], [7], and [2]. Meanwhile, Joseph Humberto Cueva et al. [8] and Bofei Zhang et al. [9] applied the ResNet model. Pingjun Chen et al. [3] employed a combination of VGG-19, CNN, YOLOv2, Inceptionv3, ResNet, and DenseNet. Conversely, Jean Baptiste Schiratti et al. [10] chose EfcientNet-B0.Christos Kokkotis et al [6] reported a 74.07% accuracy in knee OA detection using SVM. Generally ML model has lower computational requirements compared to deep learning. Deep learning, with its ability to automatically learn representations from data, has achieved remarkable success in various complex tasks. Pingjun Chen et al. [3] attained a 69.7% accuracy employing VGG-19. Other models, like CNN, YOLOv2, Inceptionv3, ResNet and DenseNet. Good interpretability but comparatively low accuracy. Kevin A. Thomas et al. [7] achieved an accuracy of 71.00% using CNN. Provide good performance but used complex model and less no of data. Aleksei Tiulpin et al. [2] employed CNN to achieve a 66.71% accuracy in knee OA. Comparatively low accuracy and training their proposed model only for the right knee persists as a limitation. Joseph Humberto Cueva et al. [8] achieved 61.00% accuracy using ResNet. Comparatively low performance and unable to emphasize the important region are the main drawbacks. Bofei Zhang et al. Jean Baptiste Schiratti et al. [10] employed the EfcientNet-B0 model, with achieving classification rates of 72.00%. Cumulatively, these investigations provide valuable perspectives and approaches for the identification and classification of knee OA, making substantial contributions to progress in environmental management. Literature review briefly explain in Table IV

III. Dataset Description

The summary of our ”Knee Osteoarthritis” is conveniently presented in Table I and II, which can be easily accessed on Kaggle [11]. Furthermore, Figure 1 provides a visual representation of the dataset's distribution across various classes. The dataset exhibits an imbalance in its distribution of data. This dataset contains knee X-ray data for both knee joint detection and knee KL grading. The Grade descriptions are as follows:

TABLE I
SUMMARY OF THE DATASET

No of Images	**Format**	**No of Classes**	**Source**
9786	PNG	5	kaggle.com

TABLE II
NUMBER OF IMAGES PER CLASS

Classes	No Of Images	Classes	No Of Images
Grade 0	2286	Grade 3	757
Grade 1	1046	Grade 4	173
Grade 2	1516		

- Grade 0: Healthy knee image.
- Grade 1 (Doubtful): Doubtful joint narrowing with possible osteophytic lipping.
- Grade 2 (Minimal): Definite presence of osteophytes and possible joint space narrowing.
- Grade 3 (Moderate): Multiple osteophytes, definite joint space narrowing, with mild sclerosis.
- Grade 4 (Severe): Large osteophytes, significant joint narrowing, and severe sclerosis.

IV. METHODOLOGY

The process begins by obtaining the dataset and performing the necessary preprocessing. "Transfer Learning Fusion (TLF)" includes feeding images into pre-trained models, and adjusting tensors. The structure consists of four convolution blocks with different kernel sizes (5x5, 3x3, 1x1), filters (32, 64, 128, and 256), Dropout layer, 'BatchNormalization,' MaxPooling2D with mish activation. The final output, obtained through max-pooling, is flattened and passes through three dense layers with Softmax activation for predicting class probabilities. The addition of a 'Multi-Layer Ensemble (MLE)' for predictions involves Majority Voting, Softmax Averaging, and Weighted Averaging. The process is outlined in Figure 3.

The workflow of our research is illustrated in Figure 2.

A. Data Preprocessing

Images are cleaned, resized, and normalized during preprocessing to ensure uniformity and lower noise. There are existing training, testing, and validation sets within the dataset. In order to address unequal class distribution and ensure efficient training and model evaluation, data augmentation is employed. By adding more training examples, it grows the dataset and facilitates the model's comprehension of new data. Nevertheless, there are drawbacks, such as the possibility that the model will become overly particular to added data and higher processing demands.

B. Architecture of Proposed Transfer Learning (TL)

Using pre-trained models, our technique begins with the input of images, then reshaping tensors for fine-tuning. We use four convolution blocks (CBs) with 32, 64, 128 and 256 filters, using 'BatchNormalization' layers and different kernel sizes (5x5, 3x3, and 1x1). A MaxPooling2D layer with additional activation functions at the end of each CB solves the vanishing gradient issue. After being flattened, the output from the final max-pooling layer is processed by three dense layers with different activation functions (256, 128 and 5 neurons). Softmax activation is used in the last layer to forecast multiclass probabilities. Then, in the first prediction stage, 'Multi-Layer Ensemble (MLE)' is presented with Majority Voting (MV), Softmax Averaging (SAvg), and Weighted Averaging (WAvg). The most productive group moves on to higher tiers [12]. In Figure 4, the Multi-Layer Ensemble (MLE) procedure is illustrated visually. Using various architecture types, we deliberately train a subset of pre-trained models, such as DenseNet, MobileNet, and VGG16. Figure 3 shows the architectural overview.

C. Justification of Our Procedural Architecture

In CNN, batch normalization addresses internal covariate shifts, maintaining a consistent input distribution for each layer during training. This expedites training, promotes stable convergence, and enhances overall performance. Our model performs better than the previous task due to its architecture and the Multi-Layer Ensemble approach amalgamates architectural strengths, resulting in heightened prediction accuracy.

V. PERFORMANCE ANALYSIS

A. Evaluation Measures

We utilized diverse metrics, such as accuracy, precision, recall (sensitivity), f1-score, specificity, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), to assess the efficacy of the model. The mathematical formulations for these metrics are provided below [13], [14].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

The model's capacity to minimize false positives is measured by precision, recall examines how well it captures true positives, F1-score combines precision and recall, and specificity analyzes how well it can decrease false negatives. Accuracy checks overall accuracy.

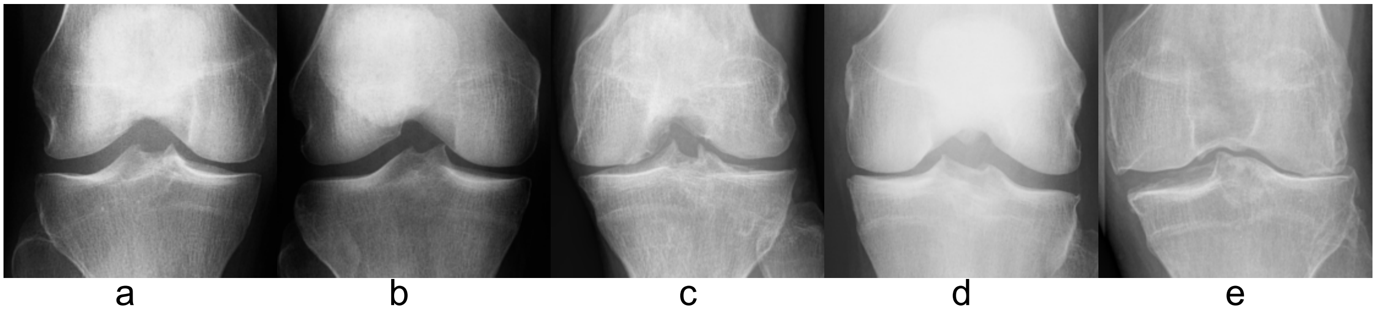

Fig. 1. Sample images for each class

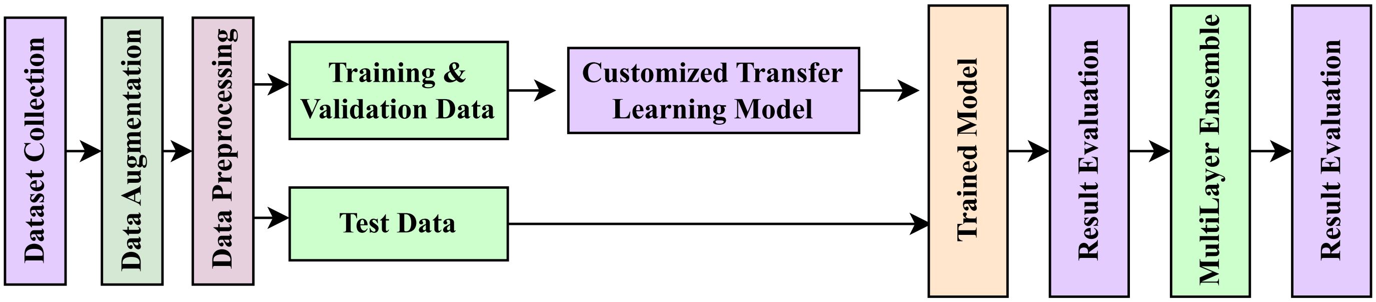

Fig. 2. Sequential workflow of the proposed methodology

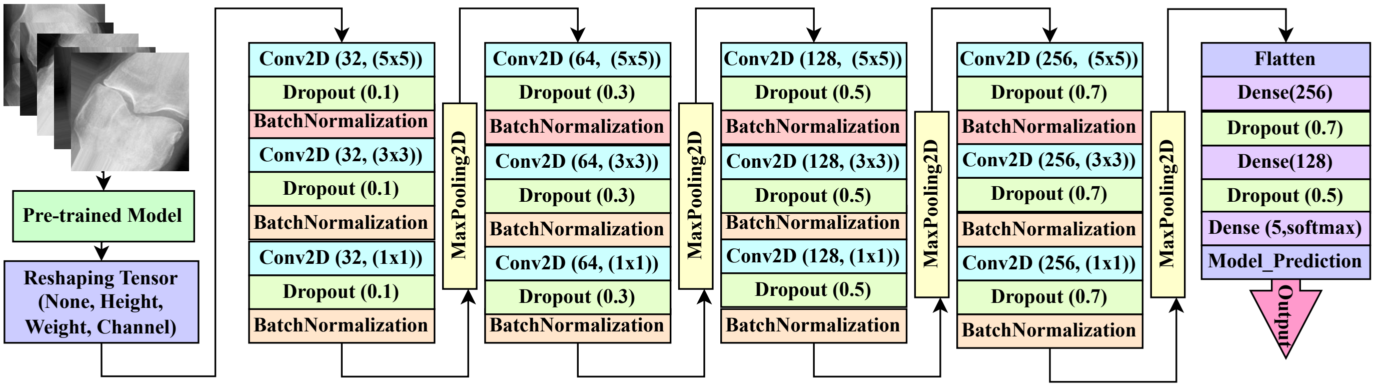

Fig. 3. Architectural view of the proposed methodology

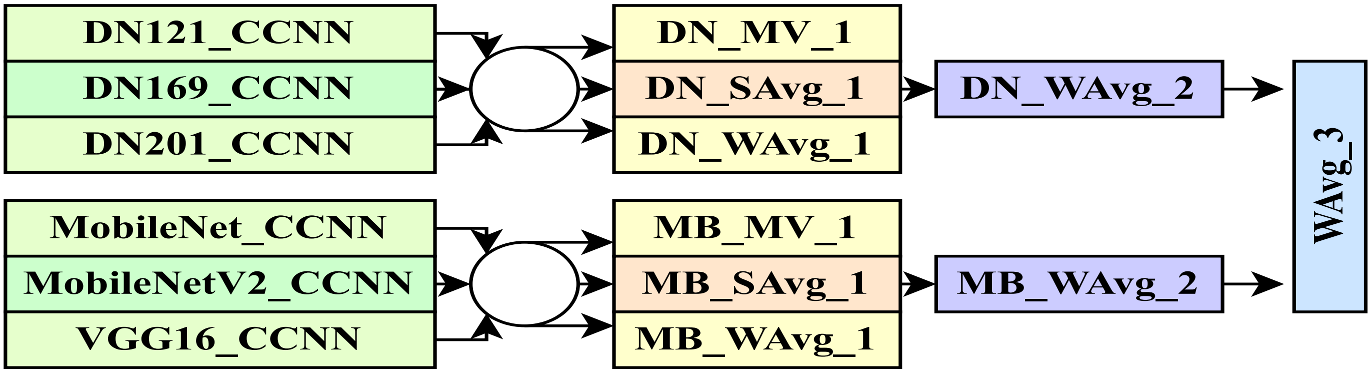

Fig. 4. Multi-Layer Ensemble Approach

TABLE III
PERFORMANCE ANALYSIS ON TEST DATA

Algorithm	Accuracy %	Precission%	Recall%	F1-score%
DN121-CCNN	73.07	73.29	73.07	72.61
DN169-CCNN	72.19	71.92	72.19	71.69
DN201-CCNN	71.54	70.98	71.54	70.46
DN_MV_1	74.72	74.45	74.72	74.03
DN_Savg_1	74.72	74.45	74.72	74.03
DN_Wavg_1	74.95	74.92	74.95	74.31
DN_Wavg-2	74.95	74.92	74.95	74.31
MobileNet _CCNN	72.42	72.86	72.42	71.62
MobileNetV2 _CCNN	71.51	72.33	71.51	70.52
Vgg16_CCNN	71.12	72.23	71.12	70.16
MB_MV_1	73.84	74.81	73.84	73.05
MB_Savg_1	73.84	74.81	73.84	73.05
MB_Wavg_1	74.14	75.03	74.14	73.36
MB_Wavg-2	74.14	75.03	74.14	73.36
Wavg-3	75.73	75.88	75.73	74.87

Note: Gradually 1,2 and 3 means 1st, 2nd and 3rd layer Enssemble.

B. Experimental setup

An Intel Xeon CPU with two cores (690 ms/step) and a GPU P100 were used for architectural operations on Kaggle. Images entered were measured (224,224,3). In order to meet our requirements and get rid of overfitting problems, models go through various amounts of epochs. batch size 16) using the Adam optimizer (categorical cross-entropy loss, lr: 0.005). Reduce was used to stop early on the Plateau (patience: 10).

C. Results Assessment

Table III showcases the results of our TLF architectures and MLE methods. Notably, the inclusion of the TLF model significantly boosts performance across various metrics.

In this work, we used three basic architectures: DenseNet, MobileNet, and VGG16. Two MobileNet variations, MobileNet and MobileNetV2, as well as three DenseNet variants, DenseNet121, DenseNet169, and DenseNet201, were examined. Since VGG16 was unique, there were no other variations. A customized CNN model was smoothly integrated with each architecture. DenseNet-based models have performed exceptionally well, with accuracy rates of 71.42% for MobileNet and 73.07% for DenseNet121. However, out of all these suggested models, VGG16 performs the lowest, at 71.12%. However, the MobileNet-based model outperforms the VGG16. The DenseNet-based model's first-layer ensemble has increased accuracy to 74.72% for MV, 74.72% for SAvg, and 74.95% for WAvg. Following the first layer ensemble, the second layer ensemble's WAvg increased to 74.95%. While a second-layer ensemble has increased it to 74.14%, the first layer ensemble's performance for the MobileNet-based model was worse than that of the DenseNet-based model. The accuracy reached 75.73% with a third-layer ensemble including the two top WAvg models. Figure 5 shows the final model's confusion matrix, which is in line with the findings.

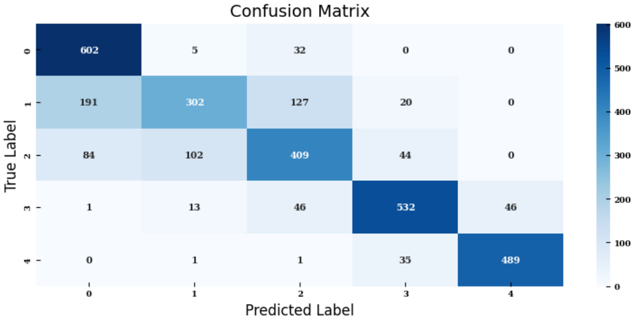

Fig. 5. Confusion Matrix of the proposed model

Nonetheless, the optimal situation is illustrated by the ROC-AUC curve in Figure 6, which graphically evaluates the architecture's convergence. The true positive rate and false positive rate exhibit low volatility, forming almost straight lines, which suggests that our model is steadily convergent.

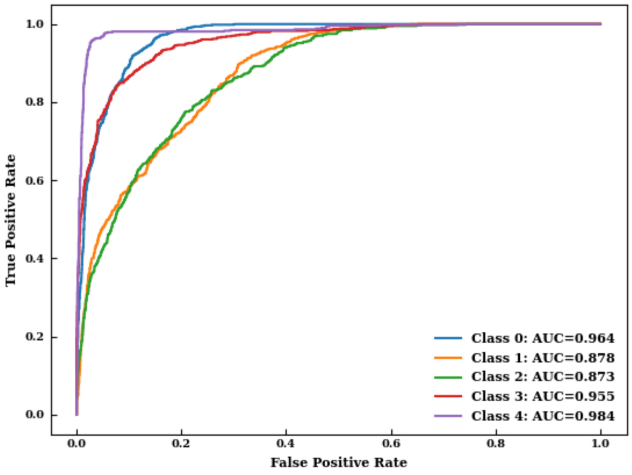

Fig. 6. ROC-AUC Curve of the proposed model

VI. DISCUSSION AND EXTENDED COMPARISON

Table IV now shows a detailed comparison, highlighting that our recent results surpass those of previous studies. It's essential to mention that although Rima Tri Wahyuningrum et al. [9] achieved a higher accuracy of 74.81% in their earlier work, they used only a small subset of the dataset and employed complex algorithms. In contrast, we worked with a dataset of 9,786 images, confirming the superiority of our model in this context. Our model, while being less complex, is advanced and user-friendly, proving its effectiveness.

VII. THREATS TO VALIDITY

Although our research uses augmentation to address dataset imbalance, it is not without limits. We intend to investigate new approaches to data balance in the future. The model's deployment on mobile devices presents difficulties and could affect performance because of normalization and resizing

TABLE IV
PERFORMANCE COMPARISON

Article	Accuracy	Precision	Recall	F1-score	Specificity
[6]	74.07	-	-	-	-
[3]	69.70	-	-	-	-
[7]	71.00	-	-	-	-
[9]	74.81	-	-	-	-
[2]	66.71	-	-	-	-
[8]	61.00	-	-	-	-
[10]	72.00	-	-	-	-
Ours	**75.73**	**75.88**	**75.73**	**74.87**	**93.73**

processes. Additionally, make an effort to use improved techniques that can eliminate overfitting problems.

VIII. CONCLUSION

This paper focuses on knee OA classification using a multi-layer ensemble and hyperparameter-tuned Transfer Learning Fusion Network. In order for these automated methods of knee OA classification and grading to be utilized for clinical research or medical diagnosis, this paper examines the several approaches and problems that may be investigated and resolved. In order to increase the performance of the model, we want to integrate the pre-trained model with a customized CNN model and a Multi-Layer Ensemble. Following the third layer ensemble, the accuracy of our suggested technique is 75.73%. Our model's performance in the study was in line with our initial predictions, emphasizing the critical role that automated knee OA identification plays and offering an improved solution.

IX. FUTURE SCOPE

For tasks like pre-processing, feature extraction, and classification, numerous approaches have been studied, but the outcomes are still far from being applied in real-world situations. Therefore, investigating cutting-edge methods may result in increased knee OA classification accuracy. We intend to extend our study in the future to incorporate automatic knee OA diagnosis and grading techniques that integrate many imaging modalities, such X-ray with ultrasound or MRI datasets.

REFERENCES

[1] A. E. Nelson, "Osteoarthritis year in review 2017: clinical," *Osteoarthritis and cartilage*, vol. 26, no. 3, pp. 319–325, 2018.

[2] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach," *Scientific reports*, vol. 8, no. 1, p. 1727, 2018.

[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] J. Abedin, J. Antony, K. McGuinness, K. Moran, N. E. O'Connor, D. Rebholz-Schuhmann, and J. Newell, "Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain x-ray images," *Scientific reports*, vol. 9, no. 1, p. 5761, 2019.

[5] K. Kalo, D. Niederer, M. Schmitt, and L. Vogt, "Acute effects of a single bout of exercise therapy on knee acoustic emissions in patients with osteoarthritis: A double-blinded, randomized controlled crossover trial," *BMC musculoskeletal disorders*, vol. 23, no. 1, pp. 1–12, 2022.

[6] C. Kokkotis, S. Moustakidis, G. Giakas, and D. Tsaopoulos, "Identification of risk factors and machine learning-based prediction models for knee osteoarthritis patients," *Applied Sciences*, vol. 10, no. 19, p. 6797, 2020.

[7] K. A. Thomas, Ł. Kidziński, E. Halilaj, S. L. Fleming, G. R. Venkataraman, E. H. Oei, G. E. Gold, and S. L. Delp, "Automated classification of radiographic knee osteoarthritis severity using deep neural networks," *Radiology: Artificial Intelligence*, vol. 2, no. 2, p. e190065, 2020.

[8] 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.

[9] B. Zhang, J. Tan, K. Cho, G. Chang, and C. M. Deniz, "Attention-based cnn for kl grade classification: Data from the osteoarthritis initiative," in *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. IEEE, 2020, pp. 731–735.

[10] J.-B. Schiratti, R. Dubois, P. Herent, D. Cahané, J. Dachary, T. Clozel, G. Wainrib, F. Keime-Guibert, A. Lalande, M. Pueyo *et al.*, "A deep learning method for predicting knee osteoarthritis radiographic progression from mri," *Arthritis Research & Therapy*, vol. 23, pp. 1–10, 2021.

[11] "knee osteoarthritis dataset," https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity, (Accessed on 12/10/2023).

[12] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.

[13] N. Haque, A. H. Efat, S. M. Hasan, N. Jannat, M. Oishe, and M. Mitu, "Revolutionizing pest detection for sustainable agriculture: A transfer learning fusion network with attention-triplet and multi-layer ensemble," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2023, pp. 1–6.

[14] N. Haque, R. Toufiq, M. Z. Islam, and M. A. A. T. Shoukhin, "Garbage classification using a transfer learning with parameter tuning," in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*. IEEE, 2024, pp. 1252–1256.