KNEE OSTEOARTHRITIS DETECTION AND SEVERITY CLASSIFICATION USING RESIDUAL NEURAL NETWORKS ON PREPROCESSED X-RAY IMAGES

[1]Thippushree,[2]Chandana J K,[3]Anup Ritti,[4]Bibi Ayesha,[5]Nandini N
[1,2]PG Student,[3] Assistant Professor,[4,5]Research Scholars
[1,2]Department of Computer Studies in Computer Applications,
[1]Davangere University, Davangere, Karnataka,India

Abstract : *One of the most common and challenging medical conditions to deal with in old-aged people is the occurrence of knee osteoarthritis (KOA). Manual diagnosis of this disease involves observing X-ray images of the knee area and classifying it under five grades using the Kellgren–Lawrence (KL) system. This requires the physician's expertise, considerable experience, and a great deal of time, and even then the diagnosis can be prone to errors. Therefore, researchers in the ML/DL domain have employed the capabilities of deep neural network (DNN) models to identify and classify KOA images in an automated, faster, and more accurate manner. To this end, we propose the application of six pretrained DNN models, namely VGG16, VGG19, ResNet101, MobileNetV2, InceptionResNetV2, and DenseNet121, for KOA diagnosis using images obtained from the Osteoarthritis Initiative (OAI) dataset. More specifically, we perform two types of classification: a binary classification, which detects the presence or absence of KOA, and a three-class classification, which classifies the severity of KOA. For comparative analysis, we experiment on three datasets (Dataset I, Dataset II, and Dataset III) with five, two, and three classes of KOA images, respectively. We achieved maximum classification accuracies of 69%, 83%, and 89%, respectively, with the ResNet101 DNN model. Our results show improved performance over the existing work in the literature.*

IndexTerms* - *knee osteoarthritis, residual neural networks, X-ray images, severity classification

I. INTRODUCTION

Knee osteoarthritis (KOA) is a disease that is most common in older people and results from the wearing of the articular cartilage between the knee joints. It is the most common joint disease in the United States alone, occurring in 13% of women and 10% of men aged above 60 years. The disease affects more than 250,000 individuals worldwide and ranks among the 50 most common diseases [1]. In the medical field, practitioners use the Kellgren–Lawrence (KL) grading system [2] as the standard to classify the severity of KOA from radiographs. Radiographs continue to be used for imaging due to their accessibility and cost-efficient nature, despite the introduction of other medical imaging technologies [3]. The KL grading system, which was accepted by the World Health Organization (WHO) as the standard in 1961 [2], splits the severity into five progression levels: 0 (healthy), 1 (doubtful), 2 (minimal), 3 (moderate), and 4 (severe). The accuracy of the severity diagnosis heavily depends on the carefulness and experience of the physician. The low dependability of the physician's grading is believed to result from the very minute differences between radiographs of adjacent grades [4]. In addition, every physician may form a different opinion on the severity grade of a radiograph based on their experience and understanding. KOA is also a disease that is very hard to detect in its early stages, when the differences between grades 0 and 1 are minimal. Because of these shortcomings in traditional diagnosis methods, automated and efficient approaches have been introduced. Many fields in recent years have seen the implementation of artificially intelligent systems, which ease and streamline tasks that were previously performed manually. The increased use of Machine Learning (ML) techniques in the medical field is assisting experts by fully or partially automating the diagnosis process. More specifically, supervised Machine Learning techniques can be used to support medical practitioners in making more sound clinical decisions [5].

Deep learning (DL) is a subfield of ML in which methods revolve around building layered models that enable a computer to autonomously perform specific tasks such as classification and object detection. Neural networks make up the backbone of DL and process information in a way that is inspired by the human brain, in that neurons receive, process, and output information. The neurons are arranged in a layered structure to mimic the structure of the human brain. What makes neural networks so valuable is that they are able to make observations from unstructured data and draw inferences without direct, explicit training. Extensive research has been conducted in the field of DL. To benchmark this research, large corpuses of datasets were built, such as ImageNet, MNIST, and MS-COCO, which became standard methods for benchmarking DL models. This gave rise to the concept of pretrained models. Having been trained on large datasets, the features learnt by pretrained models can be applied to a variety of related problems: if a model is trained on a dataset that is representative of the problem being solved, the observations made by the network can be considered a generic model of the visual world. Deep learning models are versatile in that they can be applied across many different disciplines, one example being the analysis of time-series information for forecasting.

In the realm of image classification, some popular pretrained models include VGGNet, MobileNet, ResNet, InceptionResNet, and DenseNet. Neural networks have seen use across various industries, such as medical diagnosis through image classification, automated quality control in factories, and security applications such as face recognition.

Over the past few years, the medical field has increasingly used DL techniques to streamline and automate the diagnosis process. In [6], the authors explore the use of computer-aided systems for the diagnosis of leukemia, proposing an aggregation-based deep learning model for the classification of leukemic B-lymphoblast cells that distinguishes between healthy and cancer cells. The proposed model employed pretrained, fine-tuned CNNs for feature extraction, and the extracted features were fed into an ensemble deep neural network to produce the final prediction. The authors demonstrated that their aggregation-based deep learning model performed much better than individual state-of-the-art CNNs, achieving an overall accuracy of 96.58%. The authors of [7] employed deep learning for the classification of glioma tumors, building a multiclass model that used deep learning techniques for feature extraction and SVMs for classification, achieving a classification accuracy of 96.19%.

The detection and diagnosis of KOA is one of the areas where DL techniques have been applied. After training, data are fed to the model, which predicts the severity of KOA according to the KL grading system. The high prevalence of KOA calls for accurate, reliable, and automated severity classification systems, and DL is one such solution.

II. RELATED WORK

The authors of [8] attempted to devise an automated model for detecting the severity of KOA from radiographs. To test the model's performance, it was compared with the opinions of musculoskeletal radiologists. The radiograph images were first augmented automatically before being used as input to a CNN model. The dataset used in this study was from the Osteoarthritis Initiative (OAI) and comprised more than 40,000 images. The authors reported an average accuracy of 71% and an F1 score of 70% on the full test dataset. Although they used an immensely large dataset, they compared their model's performance with the opinions of trained radiologists rather than with other related work in the field.

To treat the KL grading system as an ordinal regression problem, the authors of [9] created an ordinal regression module (ORM) for neural networks and evaluated its performance against various existing neural network solutions. The dataset used was from the OAI and consisted of 8260 knee radiographs, divided in a 1:2:7 ratio into validation, test, and training sets, respectively. They reported an accuracy of 88.09% using DenseNet-161 trained with the ORM approach, and a Quadratic Weighted Kappa score of 0.8609. Although their approach yielded good results overall, the model made a number of misclassifications when distinguishing KOA images with KL grades of zero and one.

Considering the problem of overfitting and the non-Euclidean pattern of the shape space, the author of [10] integrated a graph convolutional network (GCN) with the concept of reducing intrinsic dimension, comparing the performance of this classifier against an alternative extrinsic approach. The dataset, created by the OAI, consisted of 201 randomly selected samples graded zero, one, and two, corresponding to none, minute, and definite presence of osteophytes, respectively. The intrinsic model achieved an accuracy of 64.64%, compared with 58.62% for the Euclidean method. This study used a different grading system from the standard Kellgren–Lawrence system, which may explain the comparatively lower prediction accuracy.

In [11], the authors developed an automated classification model for KOA by extending their earlier study. An SVM was used for classification on top of a model from their previous work, which used Inception-Net-V2. Under the agreement of the Institutional Review Board of a hospital in Seoul, a dataset of 728 knee images from 364 patients was collected from their database. The work achieved an F1 score, sensitivity, and precision of 0.71, 0.70, and 0.76, respectively, and demonstrated a direct relationship between the severity of radiographic KOA images and extracted gait features, suggesting that a more precise automatic classification model could be designed using gait-image features.

Using Hu's invariant moments to study geometric alterations in knee X-ray images, the authors of [12] detected KOA at an early stage. Their method involved image preprocessing to remove unwanted distortions, identification and extraction of the cartilage region, computation of Hu's invariant moments, and finally classification using K-NN and decision tree models. Their dataset of 2000 images was manually annotated by two medical experts in accordance with the KL grading framework, and the overall accuracy of both experts' data was around 99.23%. Their high accuracy stems from segmenting regions of interest on the basis of pixel density.

The authors of [13] present a review focused on how ML techniques can be used to diagnose and predict KOA. Based on papers published between 2006 and 2019, the survey is divided into four areas: segmentation, optimal post-treatment planning, classification, and prediction/regression. Their results show that the accuracies of most diagnostic models for KOA ranged from 76.1% to 92%, and the work highlights how ML has played a significant role in creating new, automated pre- and post-treatment solutions for KOA.

In [14], the authors proposed a model based on joint space width for the recognition of osteoarthritis. Their methodology consisted of image preprocessing, extraction of the region of interest, edge computation, and calculation of the joint space width (JSW), with classification performed in the final step. The dataset of 140 images was labeled according to severity by two radiologists and two orthopedic surgeons. Their method achieved a 98.4% F1 score and 97.14% accuracy; however, a dataset as small as theirs may limit the reliability of the model in real-world deployment.

The authors of [15] present a KOA-detecting computer-aided diagnosis system using ML algorithms. Their approach preprocessed the X-ray images and then applied a normalization method using multivariate linear regression to minimize irregularity between healthy and osteoarthritic knees. Independent component analysis was used during feature extraction, followed by random forest and Naïve Bayes models for classification. Using 1024 knee X-ray images from the OAI, the approach achieved an accuracy of 82.98%, a specificity of 80.65%, and a sensitivity of 87.15%. Because feature extraction relied only on pixel intensities rather than texture analysis, this approach may not generalize as well to real-world conditions.

In [16], the authors present a discriminative regularized autoencoder that captures both discriminative and meaningful properties to refine the detection process. Adding a discriminative loss term to the training criteria forces the learned model to retain discriminative information. Using 3900 knee radiographs from the OAI, the authors compared their results with other DL methods and reported an accuracy of 82.53%, outperforming other state-of-the-art DL techniques.

In [17], the authors explore how efficiently a small dataset can be used to detect KOA at an early stage using ML techniques such as ANN, SVM, and RF, compared against a proposed approach that integrates a group method. Using only 31 images labeled according to the Kellgren–Lawrence system and Zernike-based texture features, their proposed framework achieved 85.0% accuracy, an 11% improvement over the comparison methods. While this significantly improved diagnostic accuracy, further studies involving larger patient cohorts are needed.

The authors of [18] used a deep neural network (DNN) to detect osteoarthritis from patients' statistical data on health behaviors and medical utilization, using data from 5749 subjects taken from the 2015–2016 KNHANES, a nationwide cross-sectional study conducted in Korea. Risk factors for KOA were automatically extracted using scaled principal component analysis and DNNs, achieving a classification accuracy of 71.97% and a sensitivity of 66.67%. Although this study did not rely on X-ray images, the use of health-behavior and medical-utilization data is a promising alternative that could be used alongside other methods.

In a related study, two deep convolutional neural networks were successfully applied for the automatic detection of KOA and its severity, using baseline X-ray images from the OAI. The proposed approach first localized the knee joints using a custom YOLOv2 network, then classified the knee X-ray images into severity classes based on the KL grading system by fine-tuning variants of DenseNet, VGG, ResNet, and InceptionV3. Their knee-joint detection method obtained a mean Jaccard index of 0.858 and a recall of 92.2%, while their calibrated VGG-19 model achieved a 69.7% accuracy in detecting KOA severity.

The authors of [19] collected data from healthy individuals and re-imaged their knee X-rays three years later to show that some individuals developed symptomatic osteoarthritis, demonstrating that detection is possible at a reversible stage. Their method used a linear discriminant classifier based on 3D TBM features on a dataset of 86 subjects from the OAI, reporting a testing accuracy of 78% for detecting osteoarthritis three years before symptom onset. This comparison with other supervised learning approaches yielded promising results that could inform further study of pre-symptomatic detection.

In [20], 2D radiograph images are converted into 3D images and LBP features are extracted. Dark-net-53 and Alex-Net were used to extract deep features, and the images were classified with an accuracy of 90.6%. The proposed localization model, based on a combination of YOLOv2 and an Open Neural Network Exchange (ONNX) pipeline, achieved 0.98 mAP. Images were obtained from the OAI dataset and divided into training and test sets in a 3:1 ratio. The authors successfully fused handcrafted features with deep features on a publicly available dataset, providing a useful basis for further research in this field.

III. PROPOSED METHODOLOGY

In this section we elaborate on our proposed approach, which is illustrated in Figure 1. The approach is composed of four main stages, namely, data acquisition, dataset preprocessing, model training, and classification. To begin with, the dataset of KOA X-ray images was obtained from the Osteoarthritis Initiative (OAI), available on Kaggle. This dataset had 5 different classes of images, namely, 0 (healthy), 1 (doubtful), 2 (minimal), 3 (moderate), and 4 (severe).

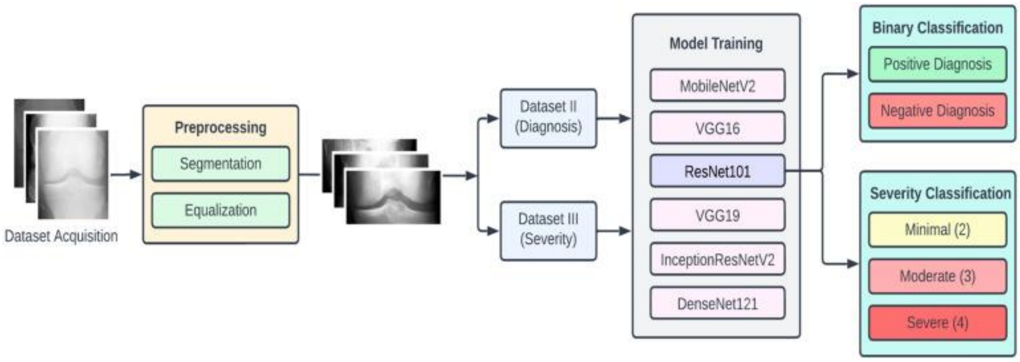

Figure 1. Proposed methodology workflow.

The dataset of X-ray images for the knee joints is not suitable, in terms of clarity and localization, for direct input to the DL models. This calls for a data preprocessing stage, in which the images are transformed so that they clearly capture the joint area where information about KOA is likely to exist. We first performed segmentation of the image, cropping it to the desired knee region so that unwanted regions were excluded. The next step was equalization of the image regions to enhance contrast for clearer visibility of the desired regions. These preprocessed images were labeled Dataset I.

Since we planned to perform two types of KOA classification, we also arranged the preprocessed images into two new datasets, namely, Dataset II and Dataset III. Dataset II consisted of two classes, negative (classes 0 and 1 combined) and positive (classes 2, 3, and 4 combined), used for binary classification (diagnosis of KOA). Dataset III consisted of three classes (original classes 2, 3, and 4), used for KOA severity classification.

For experimental purposes, we partitioned each dataset into training, testing, and validation sets in a ratio of 7:2:1, respectively. For a comprehensive study, we experimented with six pretrained CNN models, namely, VGG16, VGG19, ResNet101, MobileNetV2, InceptionResNetV2, and DenseNet121. The choice of these models was based on criteria such as availability, popularity, computational complexity, and classification accuracy. Computational resources for training these models were obtained from Google Colab. The following subsections provide a more detailed description of each stage of the proposed methodology.

3.1. Data Preprocessing

The images in all three datasets went through two preprocessing steps. The first step, segmentation, aimed to discard excess information in the image and highlight the knee joint; this was achieved by cropping the images by 60 pixels from both the top and bottom, bringing the image resolution down to 224 × 104. The second step, equalization, aimed to enhance the contrast of the images by modifying their intensity distribution; we performed histogram equalization on the images in the dataset to achieve this goal. Figure 2 illustrates the preprocessing steps.

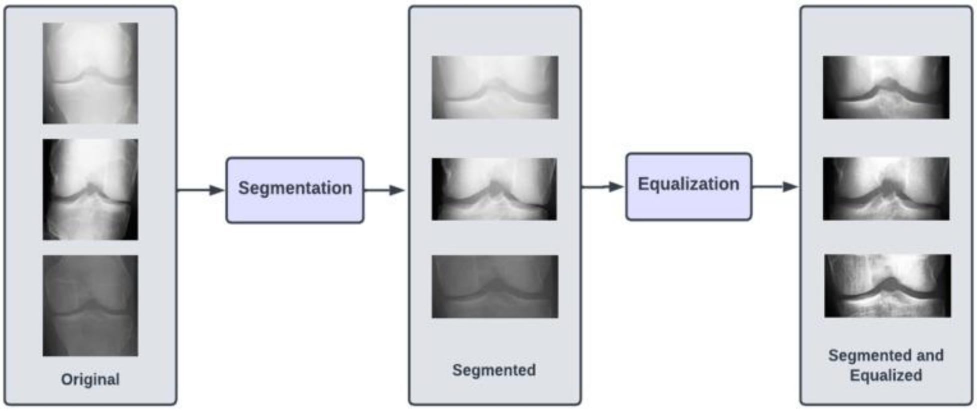

Figure 2. Preprocessing steps: segmentation and histogram equalization.

The formula applied to the images to perform histogram equalization is shown in Equation (1), where ‘r’ and ‘s’ are the input and output pixel values, respectively, and ‘L’ is the maximum pixel value in the image. The formula for the probability of intensity level rj occurring is shown in Equation (2), where ‘MN’ is the total number of pixels in the image and nj is the number of pixels with intensity rj.

$$s_k = T(r_k) = (L - 1)\, \Sigma_{kj=0}\, pr(r_j) \quad (1)$$

$$pr(r_j) = n_j / MN \quad (2)$$

3.2. Dataset Description

In this study, the knee X-ray images used to train the models are from the knee osteoarthritis severity grading dataset available on Kaggle, organized by the Osteoarthritis Initiative (OAI). There are 9786 knee images in total, divided into 5 severity levels based on the Kellgren–Lawrence (KL) grading system: 0 (healthy), 1 (doubtful), 2 (minimal), 3 (moderate), and 4 (severe). All images had a resolution of 224 × 224 pixels. Approximately 40% of the dataset images belonged to the healthy class, compared with around 18% for doubtful images, 26% for minimal images, 13% for moderate images, and just above 3% for severe images. A summary of the dataset, along with sample images, is shown in Table 1.

Table 1. Sample images of each class from the OAI dataset.

Class	**Total Images**	**Sample X-ray Images**		
0 (Healthy)	3857			
1 (Doubtful)	1770			
2 (Minimal)	2578			
3 (Moderate)	1286			
4 (Severe)	295			

To apply a multistep diagnosis approach, two more datasets were derived from the original one containing 5 classes (0–4). A binary dataset was created by combining classes 0 and 1 to represent a negative diagnosis of KOA, while classes 2, 3, and 4 were combined to represent a positive diagnosis. The second dataset was created to determine the severity of KOA and was made by removing classes 0 and 1, leaving classes 2, 3, and 4. The two derived datasets support a multistep classification approach: the first step detects the presence of KOA, and the second step diagnoses its severity. In later sections, we refer to the three datasets as follows: Dataset I is the original dataset; Dataset II is the binary dataset created by combining classes 0 and 1 as one class and 2, 3, and 4 as another; and Dataset III is the dataset created by removing classes 0 and 1 and retaining three classes corresponding to classes 2, 3, and 4.

3.3. Convolutional Neural Networks

The field of AI has witnessed rapid growth in recent years and has been applied in various domains, such as computer vision. The primary objective of computer vision is to enable computers to perceive the world in a manner similar to humans, so that relevant and pertinent information can be digitally extracted and processed from the environment. Various algorithms have been devised to achieve this goal; one, the convolutional neural network (CNN), has been particularly successful. In the context of images, a CNN is a deep learning algorithm that takes an image as input and assigns weights to various features of the image so that it is distinguishable from other images processed by the same algorithm.

Through the application of pertinent filters, CNNs have the unique ability to capture the spatial dependencies of the input image. CNNs transform images into a form that is computationally easier to process while retaining the critical features present in the image. The fundamental building block of a CNN is the convolution operation, which distinguishes CNNs from regular neural networks and is responsible for extracting high-level features from an image. Another important operation is pooling, which primarily reduces the dimensionality of an image to reduce the computational complexity of processing the input data. The convolution and pooling operations together form the i-th layer of a CNN and are primarily responsible for its feature-extraction process. The extracted features are then fed to a regular neural network for classification.

Transfer learning is a highly effective approach for tackling deep learning problems. Simply put, transfer learning makes use of the knowledge and features learnt by CNN architectures trained on large, comprehensive datasets. The idea is that if an architecture has been trained on a dataset representative enough of the real world, the features it has learnt can be treated as a generic model of the visual world. Over the years, certain architectures have performed particularly well on benchmarks and proved highly successful; MobileNetV2, VGGNet, and ResNet are a few popular examples. The CNN architectures used in our approach are as follows:

- ResNet101: a CNN containing 101 layers based on residual networks that make optimization easier, allowing increased depth and higher accuracy.
- InceptionResNetV2: a deep CNN with 164 layers based on the inception architecture that replaces filter concatenation with residual connections.
- VGG16: a CNN containing 16 layers, an upgrade over prior-art configurations, known for its uniform architecture.
- VGG19: similar to VGG16 but with 3 extra convolutional layers at the end of the network.
- DenseNet121: a CNN that simplifies the connectivity pattern between layers found in other models by incorporating several dense blocks.
- MobileNetV2: a CNN containing 53 layers, commonly used in mobile applications for its lightweight, fast, and efficient nature.

Owing to its architecture, ResNet101 can be considered the strongest CNN model for detecting and classifying KOA. With the help of regularization in its residual blocks, any layer that would reduce model performance is effectively skipped. The next subsection further describes the architecture of ResNet101.

3.4. ResNet101

Recent research in deep learning has generally affirmed that, for CNNs, a deeper model tends to perform better. However, this assumption is vulnerable to the vanishing gradient problem: once a neural network becomes too deep, the loss-function gradients shrink toward zero after repeated application of the chain rule, and once the gradients vanish, the model weights stop updating and further learning cannot occur. ResNet architectures solve the vanishing gradient problem using residual blocks, which contain skip connections that connect activations from one layer to later layers, skipping the layers in between. ResNet models are built by stacking multiple residual blocks; the skip connections provide a form of regularization, effectively skipping any layer that would reduce performance, which allows for very deep neural networks without the vanishing gradient problem. The ResNet101 model uses 101 layers, as shown in Figure 3.

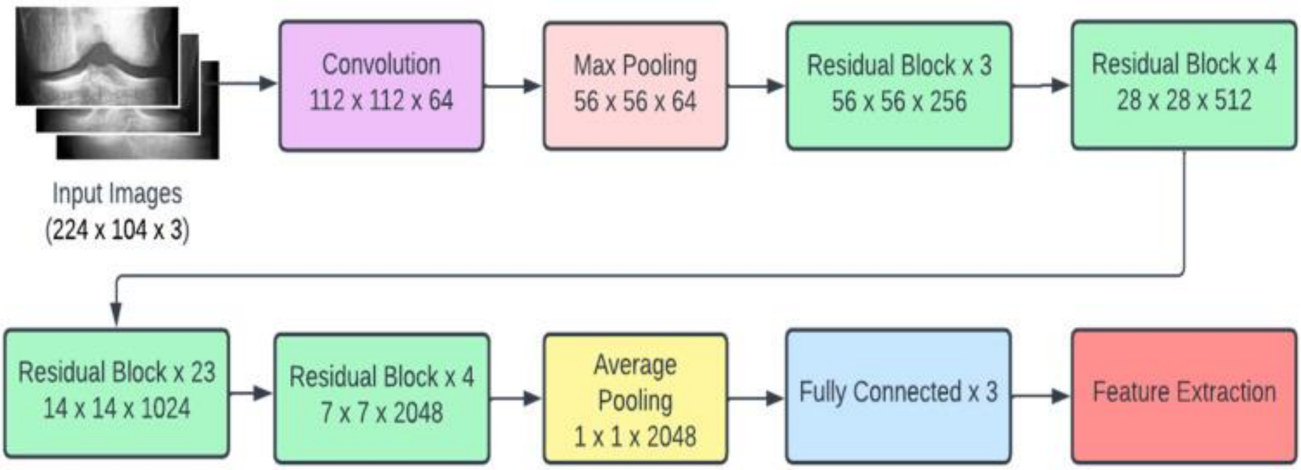

Figure 3. Architecture of the ResNet101 model.

3.5. Performance Metrics

In this subsection, we briefly discuss the performance metrics used to evaluate our classifiers. Prior to deployment, evaluating the performance of Machine Learning models is essential. By convention, classification accuracy and F1 scores are used to evaluate classifiers. Classification accuracy is the ratio of total correct predictions to the total number of samples in the dataset; meaningful inferences from this metric require the dataset to be reasonably balanced, since high accuracy on an unbalanced dataset can simply reflect a high rate of correct predictions in the majority class, with classes containing fewer samples holding less weight in the final accuracy. Another way to evaluate model performance is the F1 score, the harmonic mean of precision and recall. Precision is the ratio of correctly classified true positives to the total number of samples classified as positive, and high precision demonstrates the model's reliability in classifying samples as positive. Recall is the ratio of true positives to the total number of positive samples in the dataset, and high recall demonstrates the model's ability to correctly classify positive samples as positive. The formulae for these performance metrics are shown in Equations (3)–(6), where 'TP' is the true positive count, 'TN' is the true negative count, 'FP' is the false positive count, and 'FN' is the false negative count.

$$\text{Accuracy} = (TP + TN) / (TP + FN + TN + FP) \quad (3)$$

$$\text{Precision} = TP / (TP + FP) \quad (4)$$

$$\text{Recall} = TP / (TP + FN) \quad (5)$$

$$\text{F1 Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (6)$$

IV. EXPERIMENTAL RESULTS

The experiments were conducted on three datasets: the original dataset and two derived datasets, named Dataset I, Dataset II, and Dataset III. As discussed above, Dataset II is a binary dataset in which classes 0 and 1 were combined to represent a negative diagnosis of KOA and classes 2–4 were combined to represent a positive diagnosis. Dataset III classifies the severity of KOA and was derived by omitting classes 0 and 1, both of which represent the absence of KOA, from the original dataset. Each dataset was split into training, testing, and validation sets in a 7:2:1 ratio. Our experiments were performed using the Python programming language and its available modules. We used Google Colab, which provided a Jupyter Notebook environment and powerful hardware resources for our experiments; in its free version, Google Colab uses the Nvidia K80 GPU, which has 12 GB of memory and delivers 4.1 TFLOPS of performance.

The derived datasets were designed to support a multistep diagnosis system: the first step detects KOA, and the second step diagnoses its severity. In this section, we illustrate and discuss the experimental results obtained from our proposed methodology. We ran the deep learning models for 80 epochs with callbacks and early stopping, and the model weights were selected based on the epoch with the highest validation accuracy.

The results of the experiments performed on Dataset I are shown in Table 2. ResNet101 yielded the highest testing accuracy, at 69%. Although ResNet101 obtained the highest classification accuracy on the test dataset, the highest F1 score, 0.67, was achieved by MobileNetV2. InceptionResNetV2 had the poorest performance of the six CNNs, with the lowest testing accuracy, at 63%. The experimental results for Dataset II, shown in Table 3, indicate that ResNet101 achieved the highest classification accuracy, at 83%. The highest F1 score, 0.83, was obtained by VGG16, while the poorest-performing models, with the lowest classification accuracy of 81%, were MobileNetV2, VGG19, and InceptionResNetV2. Finally, the results for Dataset III show that the highest classification accuracy, 89%, was once again achieved by ResNet101; however, the highest F1 score, 0.89, was obtained by VGG16. The poorest-performing CNN was DenseNet121, with an accuracy of only 80%, notably lower than that of the other CNNs.

Table 2. Results of different classifiers on Dataset I containing 5 classes.

Model	**Training Accuracy**	**Training Loss**	**Validation Accuracy**	**Validation Loss**	**Testing Accuracy**	**Precision**	**Recall**	**F1 Score**
MobileNetV2	0.9951	0.0409	0.5363	2.0755	0.67	0.69	0.69	0.67
ResNet101	0.9504	0.1475	0.5484	1.8031	0.69	0.67	0.67	0.65
VGG16	0.8228	0.4589	0.5605	1.3003	0.66	0.66	0.64	0.66
VGG19	0.7045	0.6873	0.5424	1.2559	0.64	0.59	0.64	0.58
InceptionResNetV2	0.6992	0.7659	0.5291	1.3470	0.63	0.64	0.63	0.60
DenseNet121	0.7186	0.6854	0.5751	1.1337	0.64	0.65	0.64	0.59

Table 3. Results of different classifiers on Dataset II containing 2 classes.

Model	**Training Accuracy**	**Training Loss**	**Validation Accuracy**	**Validation Loss**	**Testing Accuracy**	**Precision**	**Recall**	**F1 Score**
MobileNetV2	0.8594	0.3225	0.7869	0.5358	0.81	0.83	0.82	0.81
ResNet101	0.9825	0.2060	0.7966	0.7018	0.83	0.83	0.81	0.81
VGG16	0.9396	0.0334	0.7851	1.5386	0.82	0.83	0.83	0.83
VGG19	0.8515	0.3317	0.7809	0.5231	0.81	0.83	0.82	0.81
InceptionResNetV2	0.9043	0.2312	0.7833	0.6096	0.81	0.82	0.81	0.81
DenseNet121	0.8949	0.2419	0.8087	0.5264	0.82	0.83	0.82	0.82

V. DISCUSSIONS

In this section, we analyze the performance of the models on each dataset using confusion matrices. A confusion matrix illustrates the performance of a classifier by showing the number of correct and incorrect predictions for each class: the vertical axis depicts the actual class, and the horizontal axis represents the class predicted by the model. The confusion matrices of the best-performing model for each dataset are shown in Figure 4.

An ideal confusion matrix has the largest values on the diagonal, with values decreasing as we move away from it. Although not ideal, the confusion matrix for Dataset I shows fewer misclassifications as we move away from the diagonal, with some noticeable exceptions: class 0 has both the highest number of correct predictions and the most incorrect predictions, mostly misclassified as classes 1 and 2. Since classes 1 and 2 correspond to doubtful and minimal severity levels, these misclassifications are less concerning. Our derived datasets show a substantial improvement in the ratio of correct to incorrect predictions for each class. Dataset II, which predicts a positive or negative diagnosis, shows fewer false negatives than true negatives in Figure 4, meaning the accuracy for this class increased significantly; compared with Dataset I, where the best model achieved only 68% accuracy for class 0, Dataset II achieved 85% accuracy for a negative diagnosis. Dataset III, created to determine the severity of KOA once a positive diagnosis is made in Dataset II, also shows fewer misclassifications as we move away from the diagonal in Figure 4.

Based on these confusion matrices, we infer that our proposed multistep classification approach significantly reduces the ratio of misclassifications for each class. The classification accuracy values and confusion matrices for Datasets II and III show that our multistep method is more suitable than classifying all five original classes directly.

This study has a few limitations. The dataset was relatively small and highly unbalanced—for example, class 0 had 3857 images, whereas class 4 had only 295—and this imbalance could not be fully mitigated through augmentation, since augmenting 295 images to match 3857 would introduce a large amount of relatively unoriginal data. Another limitation was the lack of clean class boundaries: class 1, labeled ‘doubtful’, was frequently misclassified as class 0 in the confusion matrices, likely due to the lack of a fully deterministic decision by those who labeled the images. For future work, we plan to use a more comprehensive dataset and explore different computer vision techniques to enhance model performance. With a better understanding of KOA, structural features could be manually engineered to obtain a more accurate, automated solution for the diagnosis and severity classification of KOA.

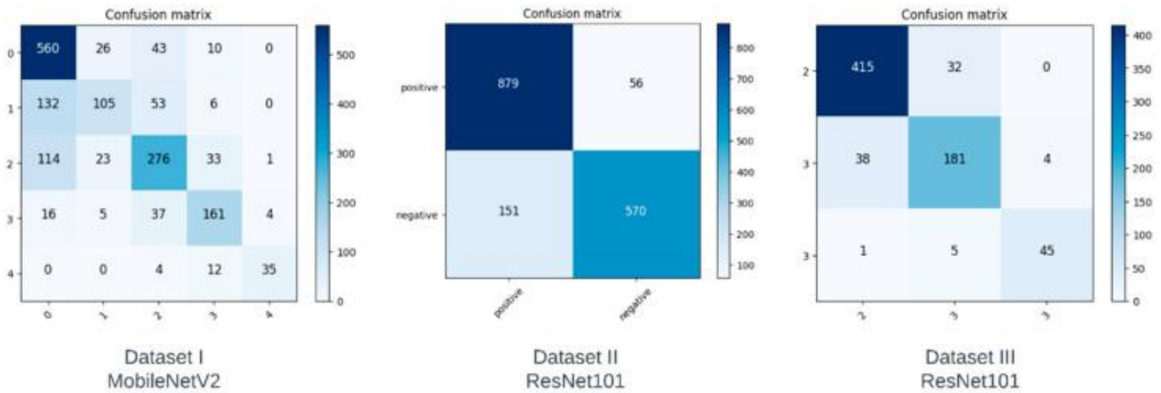

Figure 4. Confusion matrices obtained from our models.

VI. CONCLUSION

This research work addressed the identification and classification of knee osteoarthritis (KOA), one of the most challenging medical conditions in old-aged people. Our efforts were directed toward proposing, implementing, and testing an automated, fast, and accurate methodology that can help reduce the manual effort required of physicians and decrease the number of false diagnoses. For this purpose, we used the prediction capabilities of deep neural networks (DNNs), which automatically extract features from X-ray images. We trained six models (VGG16, VGG19, ResNet101, MobileNetV2, InceptionResNetV2, and DenseNet121) on the Osteoarthritis Initiative (OAI) dataset, consisting of 9786 images in total. Our experiments provided insights into how the number of classes affects classification accuracy: the ResNet101 model yielded maximum classification accuracies of 69%, 83%, and 89% on our self-formed Dataset I, Dataset II, and Dataset III, respectively. The results for Dataset II (binary classification of KOA diagnosis) and Dataset III (three-class classification of KOA severity) were better than those for Dataset I (five-class classification based on the Kellgren–Lawrence scale). The contribution of our work lies in the novel approach of combining classes to convert a five-level severity diagnosis into a multistep diagnosis that first determines whether a patient has KOA and then determines its severity.

Although our accuracy results exceeded the performance reported in similar work in the extant literature, we plan to further improve our approach by enhancing the dataset used, experimenting with additional DNN models, and exploring different computer vision tools for image segmentation and equalization.

VII. REFERENCES

[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/

[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] [PubMed] [Google Scholar]

[3] Kellgren J.H., Lawrence J.S. Radiological Assessment of Osteo-Arthrosis. Ann. Rheum. [DOI] [PMC free article] [PubMed] [Google Scholar]

[4] 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. [DOI] [PMC free article] [PubMed] [Google Scholar]

[5] Roy S., Meena T., Lim S.-J. Demystifying supervised learning in healthcare 4.0: A new reality of transforming diagnostic medicine. Diagnostics. 2022;12:2549. [DOI] [PMC free article] [PubMed] [Google Scholar]

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

[7] Latif G., Ben Brahim G., Iskandar D.A., Bashar A., Alghazo J. Glioma Tumors' classification using deep-neural-network-based features with SVM classifier. Diagnostics. 2022;12:1018. [DOI] [PMC free article] [PubMed] [Google Scholar].

[8] 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. [DOI] [PMC free article] [PubMed] [Google Scholar].

[9] Yong C.W., Teo K., Murphy B.P., Hum Y.C., Tee Y.K., Xia K., Lai K.W. Knee osteoarthritis severity classification with ordinal regression module. Multimed. Tools Appl. 2021;81:41497–41509. [DOI] [Google Scholar].

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

[11] Kwon S.B., Han H.S., Lee M.C., Kim H.C., Ku Y. Machine learning-based automatic classification of knee osteoarthritis severity using gait data and radiographic images. IEEE Access. 2020;8:120597–120603. doi: 10.1109/ACCESS.2020.3006335. [DOI] [Google Scholar]

[12] Gornale S.S., Patravali P.U., Hiremath P.S. Automatic Detection and Classification of Knee Osteoarthritis Using Hu's Invariant Moments. Front. Robot. AI. 2020;7:591827. doi: 10.3389/frobt.2020.591827 [DOI] [PMC free article] [PubMed] [Google Scholar].

[13] Kokkotis C., Moustakidis S., Papageorgiou E., Giakas G., Tsaopoulos D.E. Machine learning in knee osteoarthritis: A review. Osteoarthr. Cartil. Open. 2020;2:100069. doi: 10.1016/j.ocarto.2020.100069. [DOI] [PMC free article] [PubMed] [Google Scholar]

[14] 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. doi: 10.1007/s11760-020-01645-z. [DOI] [Google Scholar]

[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. doi: 10.1016/j.compmedimag.2019.01.007. [DOI] [PubMed] [Google Scholar]

[16] Nasser Y., Jennane R., Chetouani A., Lespessailles E., El Hassouni M. Discriminative Regularized Auto-Encoder for early detection of knee osteoarthritis: Data from the osteoarthritis initiative. IEEE Trans. Med. Imaging. 2020;39:2976–2984. doi: 10.1109/TMI.2020.2985861. [DOI] [PubMed] [Google Scholar]

[17] 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. doi: 10.1038/s41598-021-81786-4. [DOI] [PMC free article] [PubMed] [Google Scholar]

[18] 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. doi: 10.3390/ijerph16071281. [DOI] [PMC free article] [PubMed] [Google Scholar]

[19] Kundu S., Ashinsky B.G., Bouhrara M., Dam E.B., Demehri S., Shifat-E-Rabbi M., Rohde G.K. Enabling early detection of osteoarthritis from presymptomatic cartilage texture maps via transport-based learning. Proc. Natl. Acad. Sci. USA. 2020;117:24709–24719. doi: 10.1073/pnas.1917405117. [DOI] [PMC free article] [PubMed] [Google Scholar]

[20] Yunus U., Amin J., Sharif M., Yasmin M., Kadry S., Krishnamoorthy S. Recognition of Knee Osteoarthritis (KOA) Using YOLOv2 and Classification Based on Convolutional Neural Network. Life. 2022;12:1126. doi: 10.3390/life12081126. [DOI] [PMC free article] [PubMed] [Google Scholar]