Karaelmas Science and Engineering Journal
Journal home page: https://dergipark.org.tr/tr/pub/karaelmasfen
DOI: 10.7212/karaelmasfen.1480055

Research Article
Received / Geliş tarihi : 07.05.2024
Accepted / Kabul tarihi : 24.06.2024

Classification of Knee Osteoarthritis Severity by Transfer Learning from X-Ray Images

X-Ray Görüntülerinden Transfer Öğrenme ile Diz Osteoartriti Şiddetinin Sınıflandırılması

Fatma Zehra Solak*

Konya Technical University, Faculty of Engineering and Natural Sciences, Software Engineering Department, Konya, Türkiye

Abstract

Knee Osteoarthritis (KOA) is the most common type of arthritis and its severity is assessed with the Kellgren-Lawrence (KL) grading system based on evidence from both knee bones. Recent advancements point to an era where computer-assisted methods enhance KOA diagnostic efficiency. This study implemented binary and multiple classification processes based on X-ray images and deep learning algorithms for computer-aided KOA severity diagnosis. Pre-processing involved extracting the region of interest and contrast enhancement with CLAHE on the X-ray images from the included dataset. Using this dataset, 2, 3, 4, and 5 class classification processes were conducted with ResNet-50, Xception, VGG16, EfficientNetb0, and DenseNet201 transfer learning models. Each model was assessed with "rmsprop," "sgdm," and "adam" optimization algorithms. Study findings reveal that, the DenseNet201-rmsprop model achieved 87.7% accuracy, 87.2% F1-Score, and a 0.75 Cohen's kappa value for 2-class classification. For 3-class classification, it achieved 85.6% accuracy, 82.4% F1-Score, and a 0.71 Cohen's kappa value. For 4-class classification, the DenseNet201-rmsprop model provided 81.5% accuracy, 77.1% F1-Score, and a Cohen's kappa value of 0.67. In the 5-class classification, the highest success was with the Xception-rmsprop model, with 67.8% accuracy, 68.8% F1-Score, and a 0.55 Cohen's kappa value. The evaluation with varying class numbers and different transfer learning models highlights the proposed approach's effectiveness. Results of the study underscore the study's uniqueness and success in demonstrating how varying the number of classes, employing different transfer learning models and optimizers can provide clearer insights into KOA severity evaluation.

Keywords: Clahe, multiple classification, osteoarthritis, transfer learning, X-ray.

Öz

Knee Osteoartrit (KOA), her iki diz kemiğinden elde edilen kanıtlara dayanarak Kellgren-Lawrence (KL) derecelendirme sistemi ile değerlendirilen en yaygın artrit türüdür. Son gelişmeler, KOA tanı verimliliğini artırmak için bilgisayar destekli yöntemlerin kullanıldığı bir döneme işaret etmektedir. Bu çalışma, X-ışını görüntüleri ve derin öğrenme algoritmaları temelinde ikili ve çoklu sınıflandırma süreçleri uygulayarak KOA şiddeti tanısında bilgisayar destekli yöntemler geliştirmiştir. Önişleme işlemi, dahil edilen veri setindeki X-ışını görüntülerinden ilgi alanının çıkarılması ve kontrastın CLAHE ile arttırılmasını içermiştir. Bu veri seti kullanılarak, ResNet-50, Xception, VGG16, EfficientNetb0 ve DenseNet201 transfer öğrenme modelleri ile 2, 3, 4 ve 5 sınıf sınıflandırma süreçleri gerçekleştirilmiştir. Her model, "rmsprop," "sgdm," ve "adam" optimizasyon algoritmaları ile değerlendirilmiştir. Çalışmanın bulguları, DenseNet201-rmsprop modelinin 2-sınıf sınıflandırma için %87.7 doğruluk, %87.2 F1-Skoru ve 0.75 Cohen's kappa değeri elde ettiğini ortaya koymaktadır. 3-sınıf sınıflandırma için %85.6 doğruluk, %82.4 F1-Skoru ve 0.71 Cohen's kappa değeri elde edilmiştir. 4-sınıf sınıflandırmada, DenseNet201-rmsprop modeli %81.5 doğruluk, %77.1 F1-Skoru ve 0.67 Cohen's kappa değeri sağlamıştır. 5-sınıf sınıflandırmada, en yüksek başarı, %67.8 doğruluk, %68.8 F1-Skoru ve 0.55 Cohen's kappa değeri ile Xception-rmsprop modeli ile elde edilmiştir. Farklı sınıf sayıları ve farklı aktarım öğrenme modelleri ile yapılan değerlendirme, önerilen yaklaşımın etkinliğini vurgulamaktadır. Çalışmanın sonuçları, sınıf sayısının değiştirilmesinin, farklı transfer öğrenme modellerinin ve optimize edicilerin kullanılmasının KOA şiddeti değerlendirmesinde nasıl daha net bilgiler sağlayabileceğini gösterme konusunda çalışmanın benzersizliğini ve başarısını vurgulamaktadır.

Anahtar Kelimeler: Clahe, çoklu sınıflandırma, osteoartrit, transfer öğrenme, X-ray

*Corresponding author: fzgogus@ktun.edu.tr

Fatma Zehra Solak orcid.org/0000 0001 5035 7575

1. Introduction

Osteoarthritis (OA) is the most common form of musculoskeletal disorders worldwide and is a degenerative condition that is increasing with the ageing population (Martel-Pelletier 1999, Anderson and Loeser 2010). This disease, which is frequently seen in overweight and elderly individuals, usually causes friction of the bones, excessive joint pain and stiffness as a result of cartilage tear (Martel-Pelletier 1999). OA, which starts by affecting the knees, can also affect the hip and hand joints over time.

Knee Osteoarthritis (KOA) is a common condition affecting a wide age group from young to old. In the USA alone, there are 14 million people with symptomatic knees (Deshpande et al. 2016), affecting tens of millions of people worldwide (Vina and Kwoh 2018). KOA is usually diagnosed and assessed by methods such as X-ray imaging, ultrasound, computed tomography and magnetic resonance imaging (Wenham et al. 2014). Among these, radiographs (X-rays) have remained the gold standard for screening for KOA because of their cost-effectiveness, safety, wide accessibility and speed. According to radiologists, the most prominent pathological features of KOA that can be easily observed are joint space narrowing (JSN) and osteophyte formation, which can also be used in the Kellgren-Lawrence (KL) grading approach to determine KOA severity. With this approach, KOA severity is classified on the basis of a consensus, which is divided into five classes (class 0 to 4) (Kellgren and Bier 1956, Kohn et al. 2016). According to the Kellgren-Lawrence (KL) grading system, 0 indicates that there is no evidence of KOA, that is, healthy. 1 defines suspicious narrowing of the joint space, so, uncertainty regarding the presence of KOA (suspected KOA). 2 signify mild KOA as a result of possible joint space narrowing. 3 shows moderate KOA with definite evidence of narrowing of the joint space. And, 4 point the presence of severe KOA as a result of severe narrowing of the joint space.

In advanced degrees of KOA, available treatment options are limited, so early diagnosis and assessment of the disease is vital. The complex nature of KOA, the diversity of risk factors and the limitations of treatment options increase the importance of artificial intelligence-based diagnosis and evaluation methods in this field. Machine Learning (ML) and Deep Learning (DL) methods focusing on early diagnosis of KOA have the potential to improve the quality of life of patients (Wang et al. 2021). Therefore, studies on the early diagnosis and treatment of KOA patients using deep learning (DL) methods are currently making significant progress.

Tiulpin et al. (2018) developed knee osteoarthritis (KOA) diagnosis system using a deep Siamese convolutional neural network based on the Kellgren-Lawrence (KL) rating scale, achieving 66.71% accuracy and a squared Kappa value of 0.83 on the Osteoarthritis Initiative (OAI) dataset. Brahim et al. (2019) used Random Forest and Naïve Bayes algorithms, achieving 82.98% accuracy, 80.65% specificity, and 87.15% sensitivity on 1024 knee X-ray images from OAI for KOA detection. Nasser et al. (2020) proposed a discriminative regularized autoencoder, achieving 82.53% accuracy on 3900 knee radiographs from OAI. Jakaite et al. (2021) used a small dataset of 31 images and ML (ANN, SVM, and RF) techniques to achieve 85.0% accuracy in early KOA detection. Yong et al. (2022) created an ordinal regression module for neural networks, achieving 88.09% accuracy and a Quadratic Weighted Kappa score of 0.8609 on 8260 radiographs from OAI. It was reported that although the approaches gave good results, they produced misclassifications when classifying KOA images with KL degrees of 0 and 1. Wang et al. (2022) used Deep Neural Networks (DNNs) with preprocessing and a midpoint extraction technique, achieving 81.41% accuracy. Guan et al. (2022) combined traditional and deep learning models to improve the performance of early KOA detection. They achieved AUC of 0.807, 72.3% sensitivity, and 80.9% specificity on a dataset consisting of 1389 subjects with KOA and 3285 subjects without KOA obtained from the OAI. Kondal et al. (2022) presented a CNN-based approach to automatically grade knee radiographs on the Kellgren-Lawrence (KL) scale. It used an object detection model to isolate individual knees and a regression model to assign the KL scale to each knee. After fine-tuning, they achieved a mean absolute error reduction (from 1.09 to 0.28). Alshamrani et al. (2023) benefited from transfer learning models for the early detection of osteoarthritis from 3836 X-ray images labeled as KOA and non-KOA. Therefore, they performed the classification process to separate those with KOA diagnoses from those without KOA diagnoses with accuracy of 92%. Pi et al. (2023), proposed an ensemble network to detect severity of KOA using a dataset of 8260 images from the OAI open dataset. The best performance was obtained by them as accuracy of 76.93% and an F1-score of 0.7665. However, they specified that the proposed ensemble network frequently misclassifies KL grade 1 as KL grade 0 or KL grade 2. Sharma et al. (2023) proposed a model with Adam and Adamax optimizers for detection of osteoarthritis in the early stage. They trained the model with Knee Osteoarthritis Severity Grading dataset which contains various

X-Ray images of joints for healthy, moderate and severe categories. In their study, the best accuracy was achieved as 93.84% with Adam optimizer. Saini et al. (2023) studied on a dataset consisting of 3696 radiograph images to determine the severity of KOA. Three-stage pre-processing method has been proposed using a combination of different techniques. For the severity classification, 89.95% accuracy was obtained by the VGG16 model. Raza et al. (2024), presented a multi-faceted approach using feature extraction and machine learning (ML) to diagnose KOA stages from 3154 knee X-ray images and improve classification accuracy. Feature extraction methods such as Histogram of Oriented Gradients (HOG) with Linear Discriminant Analysis (LDA) and Min–Max scaling to prepare the data for classification were implemented. Using 6 ML classifiers, they achieved a best accuracy of 98.90%, distinguishing between healthy and unhealthy knees. Rehman and Gruhn (2024) introduced a novel hybrid model combining Convolutional Neural Networks (CNN) and VGG16 architectures. They created their own dataset consisting of 1650 digital X-ray images of knee joints. Their hybrid method (CNN-ResNet50) achieved an accuracy of more than 93% on the test dataset. Ahmed and Imran (2024) utilized VGG-16, VGG-19, ResNet-50, ResNet-101, and EfficientNetb7 models for the automation of the diagnosis of knee OA. Researchers performed both multi-class and binary-class classification approaches. They stated that they were particularly effective in distinguishing normal and severe cases with a classification accuracy of 99.13%, but in other cases, the effectiveness of the model decreased to 67%.

The literature review revealed that the difficulty of multiple classification of KOA severity reduces the success of studies. As a result, some researchers continue their work by considering only a few severities of KOA (Guan et al. 2022, Alshamrani et al. 2023, Sharma et al. 2023, Ahmed and Imran 2024, Raza et al. 2024).

In this study, to address the challenge of multi-class classification, multiple classification processes with different numbers of classes created based on various combinations of Kellgren-Lawrence (KL) grades were performed to determine the severity of KOA. While creating combinations, the treatment methods applied to the severity grades were taken as the basis. To implement this approach, firstly, X-ray images related to KOA were preprocessed and then subjected to 2-class, 3-class, 4-class and 5-class classification processes. For the classifications, ResNet-50, Xception, VGG16, EfficientNetb0 and DenseNet201 transfer learning models were utilized. Each of the models was run separately with each of the "rmsprop", "sgdm" and "adam" optimization algorithms.

As a result of the study, it can be concluded that utilizing multiple classifications with varying numbers of classes has proven effective in evaluating the behavior and performance of the models in classifying complex KOA cases. This study underscores the importance of utilizing diverse class combinations, different transfer learning models, and optimization strategies to improve classification accuracy and reliability.

2. Material and Methods

This section summarizes the recommended process for determining KOA severity. X-Ray Images related to KOA was first pre-processed. Then, they were subjected to multiple classification processes involving different numbers of classes using transfer learning models. Finally, the performances of the models were evaluated. Figure 1 shows the flow diagram of the study.

2.1. Dataset

In order to perform the experiments specified in the proposed flow diagram in Figure 1, a publicly available dataset consisting of 8,260 X-ray images of size 224x224x1 related to knee osteoarthritis was utilized. The dataset is accessible via "Hugging Face" (kneeosteoarthritis 2018). A larger version of the dataset is also available on "Kaggle".

The dataset provided by Chen (2018) contains x-ray images of the left and right knees from a total of 4796 participants, including both male and female patients aged 45 to 79 years. The dataset's X-ray images are categorized into five distinct groups, referred to as KL-grades (0 to 4), which represent the severity of knee osteoarthritis. In the grade 0 (healthy) category, there are 3,253 images, while there are 1,495 images classified as grade 1 (suspicious), 2,175 images as grade 2 (mild), 1,086 images as grade 3 (moderate), and 251 images as grade 4 (severe). Figure 2 indicates examples of knee joints from all degrees of CL.

2.2. Pre-Processing

Region of interest extraction and image enhancement with CLAHE were performed on the images in the dataset. Region of Interest Extraction is often used to remove unnecessary information and focus on prominent features before analyzing an image. This resulted in an image containing only the region to be analyzed. In this way, it helps to remove unnecessary information outside the area where the

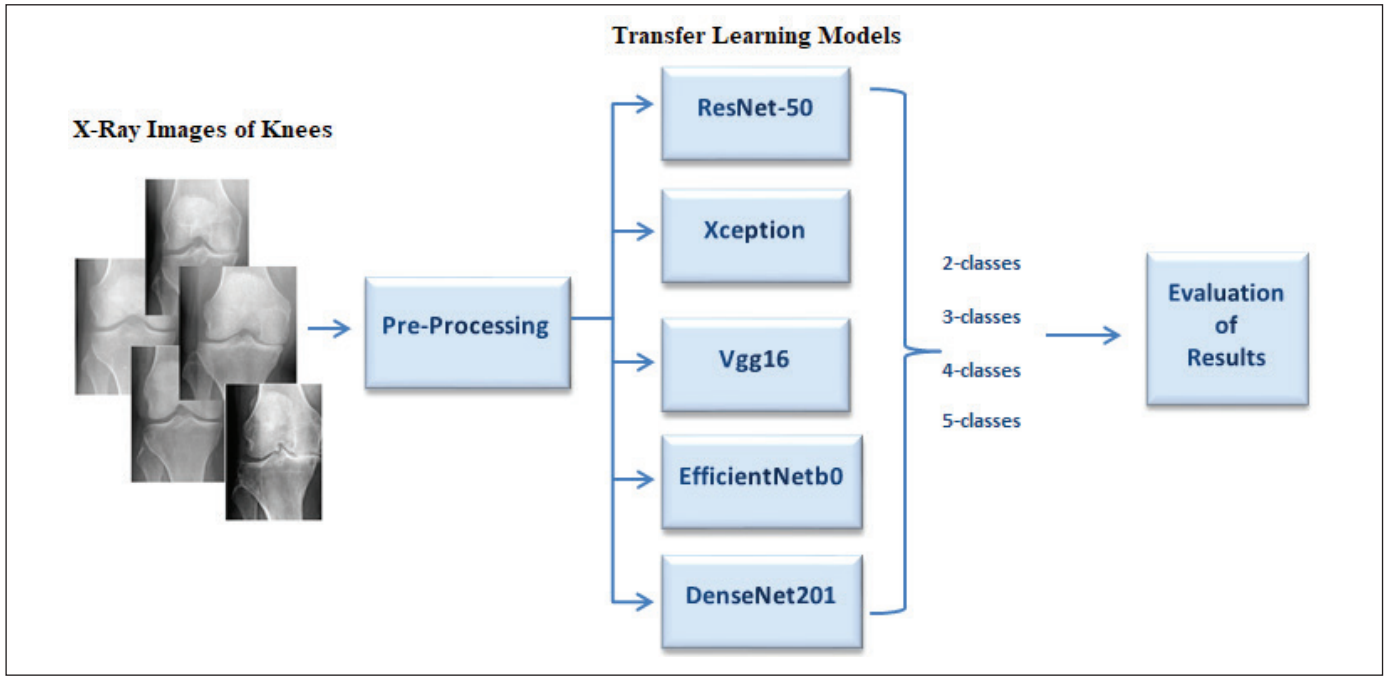

Figure 1: Flow diagram of the study.

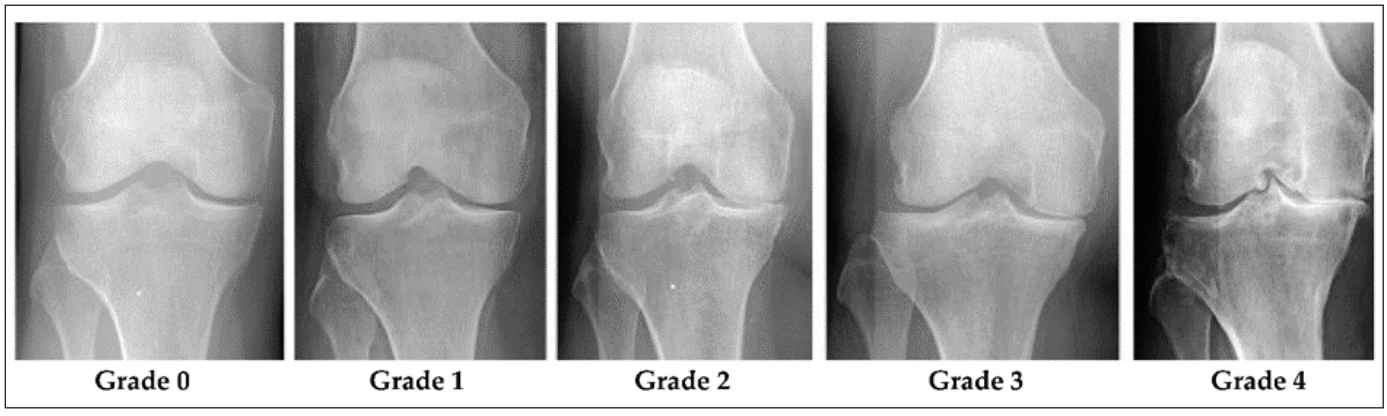

Figure 2: Knee joint specimens of all KL grades (Kim et al. 2020).

analysis focuses, making the process more effective. In this study, this process was accomplished by cropping to highlight the knee joint. Each of images was cropped by 20 pixels from all directions (top, bottom, right and left). Following the extraction the relevant area, Contrast Limited Adaptive Histogram Equazation (CLAHE) technique was benefitted to enhance the contrast of the images in the dataset. Parameters of this technique, namely, 'NumTiles' and 'ClipLimit' were chosen as (8 8) and 0.005, respectively. Figure 3 demonstrates the results of pre-processing on KOA images.

After all these processes, images in the dataset were randomly separated into training, validation and testing to be used in classification processes with transfer learning models. During the classification process, a randomly selected 15% of the data sets were used for testing, another randomly selected 15% of the remaining data were used for validation and the remaining data were used for training.

2.3. Classification

In this study, to determine the severity level of extremely complex KOA cases from X-Ray images, we performed several classification processes with different numbers of classes in addition to the 5-class classification according to the KL grading system. This approach aimed to better understand the models' behavior and address the challenges of multi-class classification. For the classifications, the num-

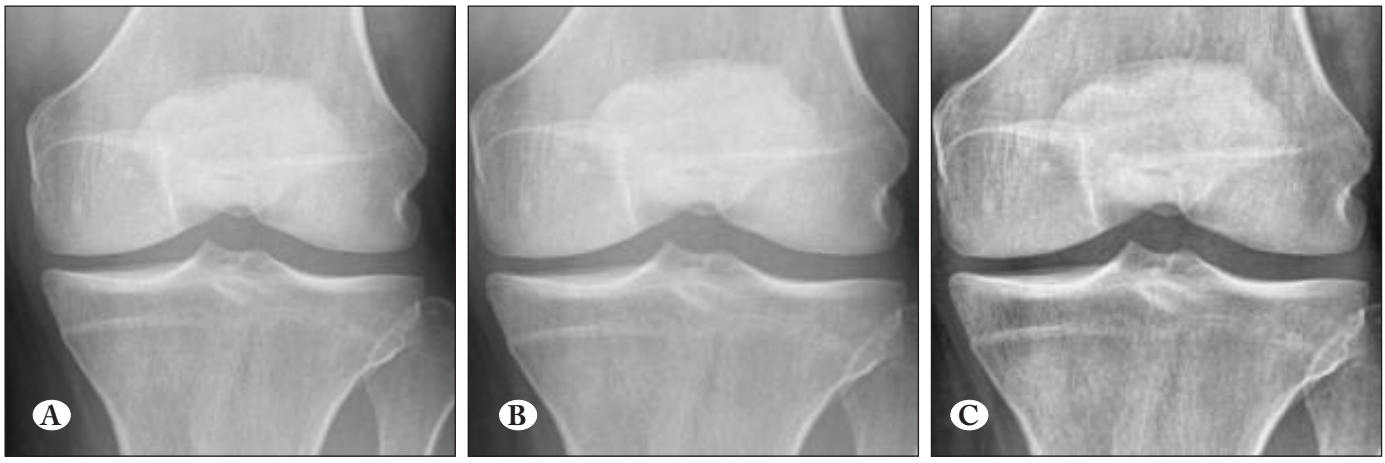

Figure 3: Applying pre-process on KOA images. **(A)** shows the original image, **(B)** illustrated the result region of interest extraction process from original image, and **(C)** CLAHE applied version of the image with region of interest extraction.

ber of classes was determined using various combinations of Kellgren-Lawrence (KL) grades. These combinations were formulated based on the treatment methods corresponding to different severity levels.

First of all, a 2-class classification approach was adopted in the study. According to KL grading system, cases that were considered healthy and suspicious were evaluated as healthy, that is, no KOA, and the others were considered as KOA cases. Then, a 3-class classification was carried out. In the 3-class classification, healthy and suspicious cases were merged into one category. Mild and moderate KOA cases were combined into another category. For these degrees of KOA cases, exercise, physical therapy, and medication are sufficient. And, severe KOAs are considered as a separate category since treatment of this degree usually requires surgery. Later on, 4-class classification was performed. In this classification, healthy and suspected cases were grouped together and regarded as a single category. Mild, moderate and severe KOAs were evaluated as separate categories due to differences in their treatment. Finally, as in other literature studies, a 5-class classification was made based on the severities in the KL grading system. Table 1 summarizes the class structures in classifications carried out in this study.

For the 2, 3, 4, and 5 class classification tasks aimed at determining the severity of KOA, as shown in Table 1, we utilized several transfer learning models including ResNet-50, Xception, VGG16, EfficientNetB0, and DenseNet201.

ResNet-50 introduced by He et al. (2016) represents a deep convolutional neural network (CNN) architecture notable for its utilization of residual connections. The incorporation of residual connections in deep neural networks alleviates the problem of vanishing gradients by enabling the learning of residual mappings (Sharma et al. 2022). It comprises 50 layers, consisting of convolutional, pooling, fully connected, and shortcut layers (He et al. 2016). As a consequence of the residual connections, gradients propagate more effectively through the network, facilitating accelerated learning and convergence (Shivadekar et al. 2023).

The Xception model, proposed by Chollet (2017), extends the inception architecture through the exclusive utilization of depth-wise separable convolutions (Chollet 2017). Xception, an extreme version of its predecessor model Inception (Szegedy et al. 2016), is a Convolutional Neural Network (CNN) architecture that consists of 71 layers in depth. The Xception model utilizes depth-wise separable convolutional neural networks with residual connections instead of standard convolution techniques. Depth-wise separable convolution, commonly known as 'Separable Convolutional,' represents an alternative to traditional convolutional layers, aiming to improve computational efficiency (Shaheed et al. 2022). In addition, these layers are lighter and more computationally efficient compared to traditional convolutional layers.

VGG16 proposed by Simonyan and Zisserman (2014) has a 16-layer network comprising 13 convolutional layers and 3 fully connected layers. It is a CNN architecture that creates a deep network using repeated convolutional layers, followed by max-pooling layers for spatial downsampling (Li et al. 2023). Increasing depth gradually while keeping the filter size modest at 3 × 3, the network aims to learn hierarchical image representations (Yang et al. 2023). The

Table 1. Structure of classifications**.**

Classification	Classes	Structure
1	2	Class 0: healthy and suspected cases **(do not require treatment)**
		Class 1: mild, moderate and severe cases **(a treatment is needed)**
2	3	Class 0: healthy and suspected cases **(*do not require treatment)***
		Class 1: mild and moderate KOA cases **(*exercise, physical therapy and medication are sufficient for treatment)***
		Class 2: severe KOA cases **(*treatment goes up to surgery)***
3	4	Class 0: healthy and suspected cases **(do not require treatment)**
		Class 1: mild KOA cases (***exercises can be performed at home*** **for treatment)**
		Class 2: moderate KOA cases **(for the treatment, medication and physiotherapy are sufficient)**
		Class 3: severe KOA cases **(*treatment is either heavy injections of medication directly into the affected area or surgery)***
4	5	Class 0: ***Healthy KOA cases***
		Class 1: ***Suspected KOA cases***
		Class 2: ***Mild KOA cases***
		Class 3: ***Moderate KOA cases***
		Class 4: ***Severe KOA cases***

utilization of smaller filters, rather than larger ones, enables a deeper network with fewer parameters.

EfficientNetB0, containing a total of 237 layers, was introduced in 2019, utilizing inverted residual blocks along with Squeeze and Excitation (SE) blocks, and employing swish activation (Tan and Le 2019). The model architecture comprises multiple convolutional layers employing a 3x3 receptive field and mobile inverted bottleneck convolutional layers (Mou and Razzak 2023). EfficientNetB0 was crafted through the AutoML Mnas neural architecture search, resulting in a network structured with mobile inverted bottleneck blocks, enhanced by squeeze-and-excitation optimization for superior performance and efficiency (Mou and Razzak 2023).

The DenseNet201 framework, initially proposed by Huang et al. (2017), is a recent advancement in dense-network architecture, known for its effectiveness in image recognition tasks. The architecture of DenseNet201 adopts a distinctive strategy wherein each layer is interconnected in a feed-forward manner with every other layer. Moreover, the model integrates both pooling layers and a compact structure. As a result of these design decisions, there is a notable reduction in parameter count and overall model complexity, leading to enhanced efficiency (Turkoglu 2021).

These models are pre-trained on the ImageNet database with 1000 classes. Since 2, 3, 4 and 5-class classification was made in this study, all models were fine-tuned according to these classifications. The ResNet-50 structure used is shown in Figure 4. Other used models were similarly fine-tuned according to the number of classes.

The Transfer Learning Models were configured with specific parameters as follows: 15 epochs were set for training, with a batch size of 8 samples per iteration. The cross entropy loss function was employed to evaluate model performance. Additionally, three different optimization algorithms were utilized for each network: Root Mean Square Propagation (rmsprop), Stochastic Gradient Descent with Momentum (sgdm), and Adaptive Moment Estimation (adam).

Classification Accuracy (CA), F1-Score (F1) and Cohen's Kappa Value (K) were used to evaluate the performance of the models. Classification Accuracy (CA) and F1-score values are given as a percentage (%).

3. Results and Discussion

Before loading the training set of X-ray images into the transfer learning models for determining the severity of KOA, data augmentation was performed to address class imbalance, enhance prediction accuracy, and reduce overfit-

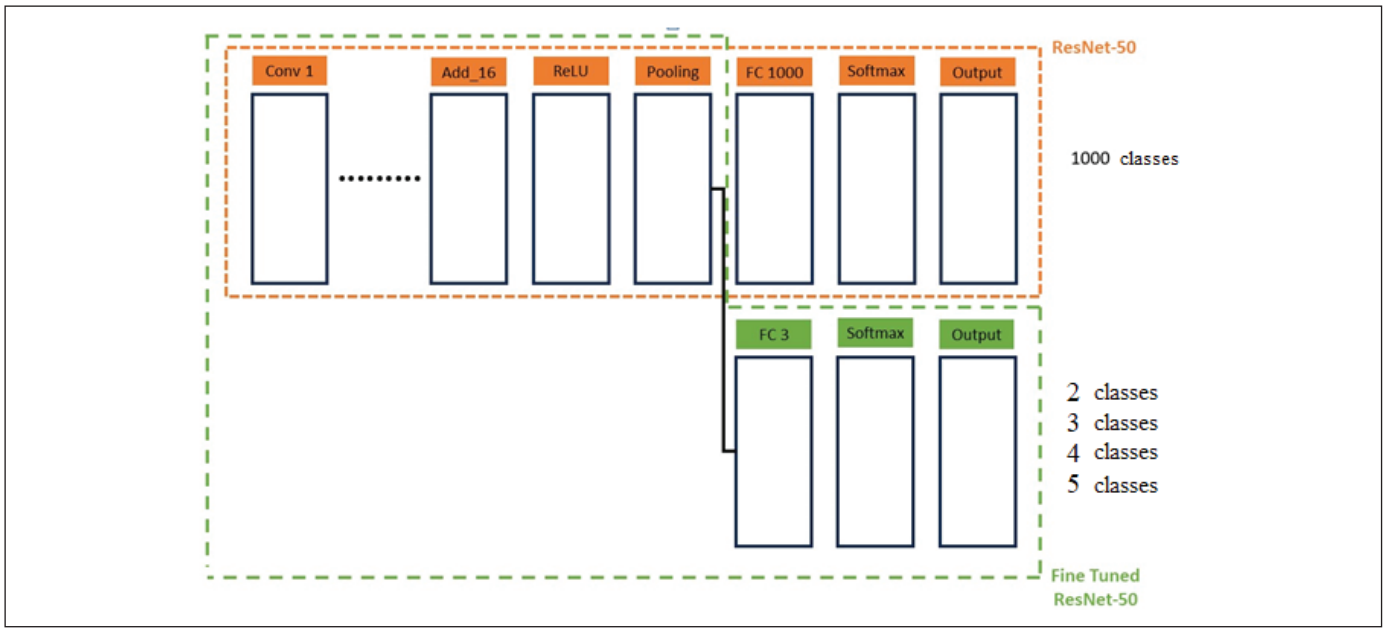

Figure 4: Used ResNet-50 structure.

Table 2. Classification results of ResNet-50 models for test set using different optimizers.

	2-class			3-class			4-class			5-class		
ResNet-50	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**
rmsprop	85.2	84.4	0.69	84.0	82.0	0.68	**79.9**	**77.7**	**0.64**	67.0	64.2	0.52
sgdm	85.1	84.7	0.69	82.6	80.7	0.66	77.5	76.8	0.62	64.7	63.6	0.50
adam	**86.4**	**85.6**	**0.72**	**84.3**	**82.0**	**0.68**	77.4	75.5	0.63	**67.0**	**66.4**	**0.54**

ting. For the augmentation, random rotation, random reflection in the left-right direction, random reflection in the top-bottom direction, range of horizontal shear and range of vertical shear were preferred. After this process, augmented X-ray images were given to fine-tuned transfer learning models and the models were trained with the determined parameters. Subsequently, the trained models were applied to the test set, consisting of 1239 X-ray images.

Table 2 evaluates the performances of ResNet-50 for the test set using rmsprop, sgdm and adam optimizers on 2-class, 3-class, 4-class and 5-class classification tasks.

According Table 2, across various classification tasks, different optimizers demonstrate varying performances. Notably, for 2-class classification, adam exhibits superior performance with an accuracy of 86.4%, leading in F1-Score (85.6%) and Kappa (0.72), indicating strong precision, recall, and reliability as the F1-score reflects a balance between precision and recall. Also high Kappa indicates agreement beyond chance. However, as the number of classes increases, the overall performance of all optimizers tends to decrease. Specifically, while adam and rmsprop perform relatively well in 3-class classification, achieving similar accuracy and reliability, sgdm lags slightly behind. In 4-class classification, rmsprop maintains a lead in accuracy (79.9%) and reliability, closely followed by adam. However, as the complexity increases to 5-class classification, the performance gap between adam and rmsprop narrows, yet adam demonstrates better overall performance and reliability, particularly reflected in its higher F1-score (66.4%) and Kappa (0.54). Nevertheless, across all scenarios, sgdm consistently trails behind in all metrics.

Test performances of Xception models using different optimizers are shown in Table 3.

Table 3 shows that, in 2-class classification, adam emerges as the top performer, boasting an accuracy of 87.1%, F1-Score of 86.8%, and Kappa of 0.74, demonstrating its robust precision and recall balance. rmsprop closely follows suit, showcasing comparable performance metrics with an

Table 3. Classification results of Xception models for test set using different optimizers.

	2-class			3-class			4-class			5-class		
Xception	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**
rmsprop	86.7	86.2	0.72	83.4	79.6	0.68	**80.3**	**77.8**	**0.66**	**67.8**	**68.8**	**0.55**
sgdm	81.4	80.7	0.62	78.4	73.6	0.58	72.7	69.0	0.53	59.8	53.6	0.42
adam	**87.1**	**86.8**	**0.74**	**84.3**	**83.2**	**0.69**	77.4	76.7	0.62	65.5	65.6	0.52

Table 4. Classification results of VGG16 models for test set using different optimizers.

	2-class			3-class			4-class			5-class		
VGG16	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**
rmsprop	76.0	76.0	0.53	79.7	53.5	0.60	63.4	32.6	0.24	51.7	42.8	0.36
sgdm	**86.7**	**86.3**	**0.73**	**83.5**	**82.5**	**0.68**	**78.8**	**78.7**	**0.64**	**67.3**	**67.0**	**0.55**
adam	84.6	83.6	0.68	82.6	77.6	0.66	78.8	74.6	0.63	65.9	62.0	0.52

accuracy of 86.7%, F1-Score of 86.2%, and Kappa of 0.72. These results highlight the importance of Kappa in assessing the model's agreement beyond what would be expected by random chance, emphasizing adam's superiority in reliability. However, as the classification complexity increases to 3 classes, adam maintains its lead with an accuracy of 84.3%, F1-Score of 83.2%, and Kappa of 0.69, albeit with slightly reduced metrics, while rmsprop remains a strong contender with an accuracy of 83.4%, F1-Score of 79.6%, and Kappa of 0.68. Conversely, sgdm exhibits behind significantly across all aspects, exhibiting lower performance metrics. As the task complexity escalates further to 4 classes, rmsprop takes the lead with an accuracy of 80.3%, F1-Score of 77.8%, and Kappa of 0.66, emphasizing a good balance between precision and recall, followed closely by Adam, indicating a consistent performance trend. However, sgdm once again falls short in performance metrics. Finally, in the 5-class classification, rmsprop exhibits notable performance in terms of F1-Score (68.8%) and Kappa (0.55), while adam demonstrates slightly lower accuracy (65.5%) and F1-Score (65.6%). However, both optimizers outshine sgdm, which consistently performs the lowest across all classification tasks.

Table 4 presents the test performances of VGG16 models using rmsprop, sgdm and adam optimizers.

In the 2-class classification scenario, sgdm emerges as the top performer for the VGG16 model, boasting the highest accuracy of 86.7% and an impressive F1-Score of 86.3%, indicating its ability to balance precision and recall effectively. Moreover, sgdm achieves a Cohen's Kappa coefficient of 0.73, signifying substantial agreement beyond random chance, which is crucial for binary classification tasks. Following closely, adam demonstrates strong performance with an accuracy of 84.6% and an F1-Score of 83.6%, although its Kappa coefficient of 0.68 suggests slightly lower reliability compared to sgdm. rmsprop trails behind both sgdm and adam in accuracy and F1-Score, indicating less robust performance in distinguishing between the two classes. As the classification tasks become progressively more complex (3-class, 4-class, and 5-class), sgdm consistently maintains its superiority, showcasing its versatility and effectiveness across various classification challenges. However, it's noteworthy that as the number of classes increases, the performance metrics for all optimizers, including sgdm, witness a decline. This decline is particularly evident in the diminishing F1-Score and Kappa values of adam and rmsprop, highlighting sgdm's resilience and reliability using with VGG16 in tackling increasingly intricate classification scenarios.

For the test set, performances of EfficientNetb0 models using rmsporp, sgdm and adam optimizers are seen in Table 5.

In the 2-class classification task, both rmsprop and adam achieve an accuracy of 85.3%, with sgdm slightly behind at 82.1%. adam exhibits the highest F1-Score (84.9%) and Kappa coefficient (0.70), indicating its effectiveness in achieving a balance between precision and recall and showing strong agreement beyond random chance. As the classification tasks become more complex, rmsprop and adam consistently perform similarly, maintaining competitive ac-

Table 5. Classification results of EfficientNetb0 models for test set using different optimizers.

	2-class			3-class			4-class			5-class		
EfficientNetb0	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**
rmsprop	85.3	84.7	0.70	**83.4**	82.5	0.67	**78.3**	**76.6**	**0.62**	64.1	63.3	0.49
sgdm	82.1	81.3	0.63	77.8	70.0	0.56	70.7	64.3	0.47	58.2	52.8	0.39
adam	**85.3**	**84.9**	**0.70**	82.6	**83.6**	**0.67**	77.1	76.3	0.61	**64.6**	**66.2**	**0.51**

Table 6. Classification results of DenseNet201 models for test set using different optimizers.

	2-class			3-class			4-class			5-class		
DenseNet201	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**	**CA**	**F1**	**K**
rmsprop	**87.7**	**87.2**	**0.75**	**85.6**	**82.4**	**0.71**	**81.5**	**77.1**	**0.67**	**67.8**	**67.2**	**0.55**
sgdm	84.7	84.2	0.69	83.5	81.1	0.67	78.0	75.2	0.62	64.9	63.0	0.49
adam	86.9	86.7	0.73	84.0	81.6	0.68	78.9	76.8	0.63	67.9	65.2	0.54

curacy and F1-Score values across all tasks. However, sgdm trails behind in performance metrics in all classification scenarios, suggesting lower effectiveness in handling diverse classification challenges.

The Table 6 displays the classification results of DenseNet201 models for test set utilizing different optimizers across various classification tasks (2-class, 3-class, 4-class, and 5-class)

Across all classification tasks, including 2-class, 3-class, 4-class, and 5-class scenarios, the DenseNet201 models exhibit varying performances contingent on the optimizer utilized. In the 2-class classification, rmsprop notably achieves the highest accuracy at 87.7%, closely followed by adam (86.9%) and sgdm (84.7%). However, rmsprop maintains a significant lead in both F1-Score (87.2%) and Cohen's Kappa coefficient (0.75), underscoring its robustness in balancing precision and recall and achieving substantial agreement beyond random chance. As the complexity of classification tasks increases, there is a slight decline in performance metrics across all optimizers. In the 3-class scenario, rmsprop leads with an accuracy of 85.6%, adam follows closely with 84.0%, and sgdm trails with 83.5%. Again, rmsprop excels in F1-Score (82.4%) and Kappa (0.71), demonstrating its effectiveness in accurately classifying instances across multiple categories. In the 4-class classification, rmsprop maintains its lead with an accuracy of 81.5%, followed by sgdm (78.0%) and adam (78.9%). Despite the decrease in performance compared to the 2-class scenario, rmsprop demonstrates the highest F1-Score (77.1%) and Kappa (0.67), showcasing its capability to handle classification challenges involving multiple categories. In the 5-class classification, rmsprop continues to dominate with an accuracy of 67.8%, followed by adam (67.9%) and sgdm (64.9%). Although there is a further decrease in performance metrics as the number of classes increases, rmsprop maintains its superiority in F1-Score (67.2%) and Kappa (0.55), indicating its reliability and effectiveness even in more complex classification scenarios with increased class diversity.

The tables (Table 2-6) show that as the number of classes increases, Classification Accuracy (CA), F1 Score, and Cohen's Kappa (K) metrics generally decrease for all models and optimizers. This decrease can be attributed to the inherent difficulties associated with distinguishing between a larger number of classes, resulting in decreased model accuracy and predictive power. Additionally, as the number of classes increases, the F1 Score decreases due to the increasing difficulty of striking a balance between precision and recall, while Cohen's Kappa coefficient also decreases as the agreement between predicted and actual labels becomes more difficult to achieve beyond random chance. DenseNet201 consistently achieves the highest performance across different class configurations, particularly with the rmsprop optimizer. Xception and ResNet-50 also perform well, especially with the adam optimizer, but their performance drops more significantly with increasing class complexity. VGG16 shows a substantial decline in performance with more classes, indicating it is less effective for complex tasks. EfficientNetb0 has moderate performance but does not surpass DenseNet201 or Xception. Overall, it is said that rmsprop

consistently maintains relatively higher performance across various class settings compared to sgdm and adam. This suggests that rmsprop may offer more stable and reliable performance, making it a preferable choice for optimizing classification models, particularly in scenarios with higher class diversity. In addition, DenseNet201 with rmsprop stands out as the most robust combination for multi-class classification, but all models face challenges as the classification task becomes more complex.

When comparing the models based on classification results with the same number of classes, several conclusions can be drawn from examining all the tables.

For the 2-class classification, the DenseNet201 model achieved the highest performance in terms of all metrics among all models when using "rmsprop" optimizer. Figure 5 shows the confusion matrix of this classification with DenseNet201-rmsprop. From this figure, it can be seen that Class 0, which contains images of healthy and suspicious cases, is better classified.

Figure 5: Confusion matrix of 2-class classification with DenseNet201-rmsprop.

For the 3-class classification, the highest results across all metrics were achieved when using DenseNet201 with the rmsprop optimizer. The confusion matrix for this classification is depicted in Figure 6. As observed from the matrix, images categorized as Class 0, representing healthy and suspicious cases, were classified with the highest accuracy. Conversely, Class 2, consisting of X-Ray images indicating severe KOA, exhibited the lowest classification success. This discrepancy can be primarily attributed to the significant imbalance in data distribution among the classes.

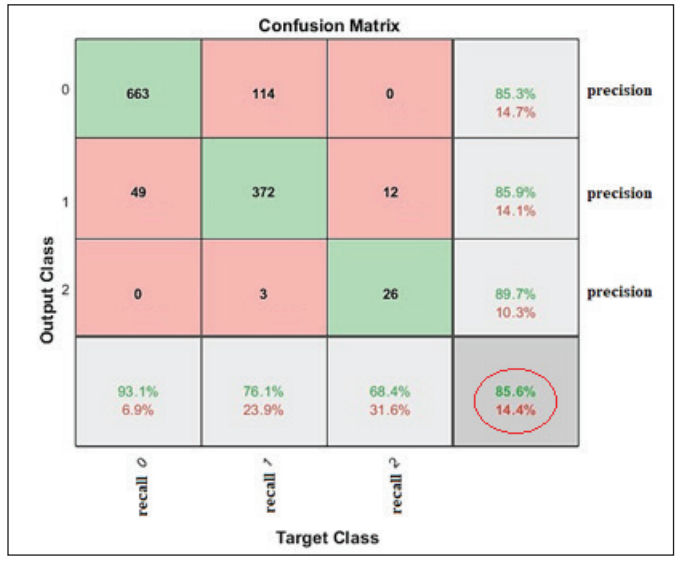

Figure 6: Confusion matrix of 3-class classification with DenseNet201-rmsprop.

In the 4-class classification, the DenseNet201 model with the rmsprop optimizer demonstrated the highest success across all metrics. Figure 7 presents the confusion matrix for this classification. It is evident from the matrix that the highest accuracy was achieved in Class 0, which comprises images of healthy and suspicious cases. Conversely, the

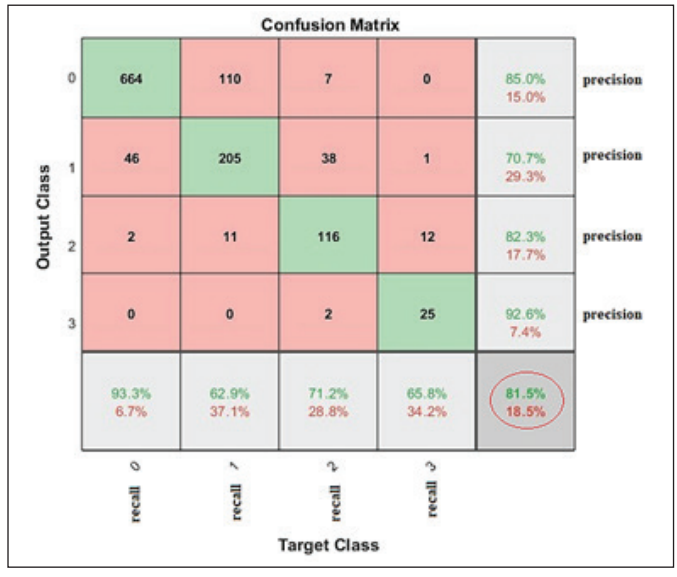

Figure 7: Confusion matrix of 4-class classification with DenseNet201-rmsprop.

lowest accuracy was observed in Class 1, representing mild KOA images.

Finally, in the 5-class classification process, although the Xception-rmsprop model achieved the same accuracy and K value as DenseNet201-rmsprop, it achieved a higher success in terms of F1-score. For this reason, the Xception-rmsprop model provided the best performance in this classification. In this classification, classes represent the degrees of KL grading system. Figure 8 shows the confusion matrix of this classification.

According to Figure 8, the highest success with the Xception-rmsprop model was achieved in the classification of Class 4 i.e. images of severe KOA cases. The next highest success was in the classification of healthy images (Class 0). The lowest classification success was in the Class 1 classification of images in which the presence of KOA was suspected. These images could not be differentiated very well from images of healthy and mild KOA cases.

Since the DenseNet201-rmsporp model achieved the same accuracy and kappa value in this classification, if we examine the confusion matrix in Figure 9, the highest rate was achieved in the classification of images of severe KOA cases, while the second successful result was obtained in the classification of those of moderate KOA. Similar to the Xception model, with DenseNet201-rmsprop, images of suspected KOA cases were classified at the lowest rate.

Training times in hours (hr) and minutes (min) when using the optimizers where all models give the highest performance for 2, 3, 4 and 5-class classifications are given in Table 7.

As can be seen from Table 7, the longest training took while training the Densenet201 model. The shortest training took place while training the VGG16 model. While there are big differences between them in terms of duration, we cannot talk about big differences in terms of performance. The Xception model also provided the highest performance for 5 classes and completed its training in a not too long time. In VGG16, it performed very closely with it for 5-class classification, even its kappa values were equal, and it completed the training period in a very short time.

A brief comparison with other studies for KOA severity detection using deep learning models is summarized in Table 8.

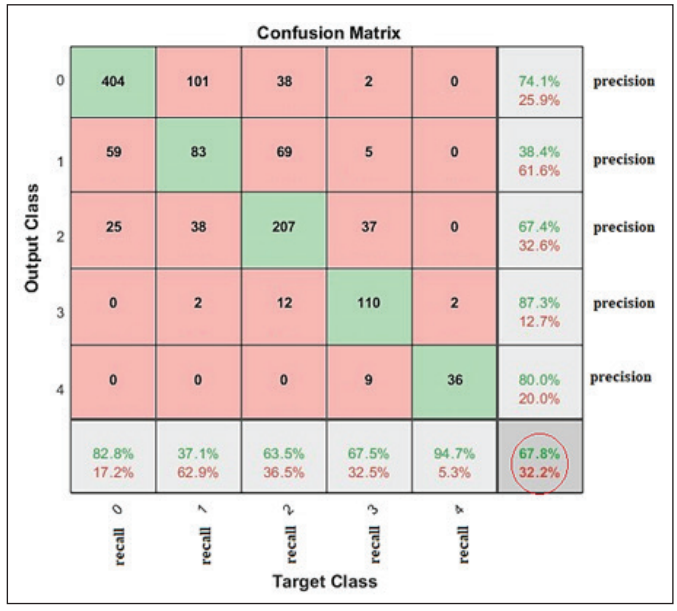

Figure 8: Confusion matrix of 5-class classification with Xception-rmsprop.

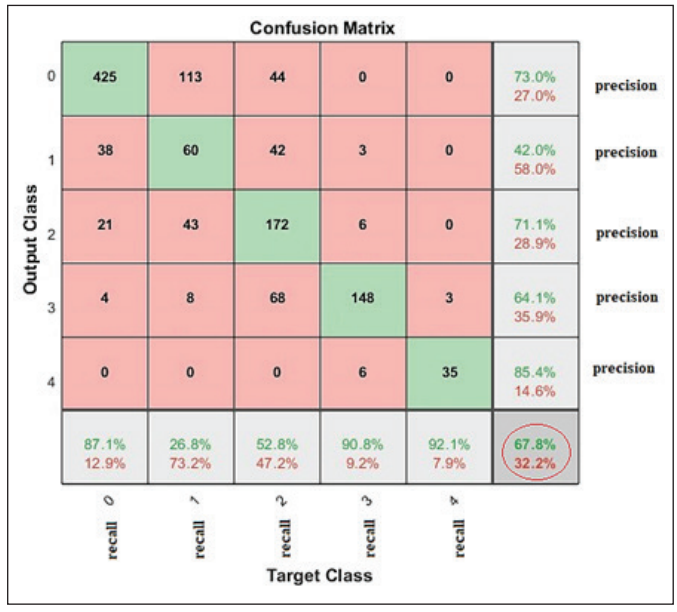

Figure 9: Confusion matrix of 5-class classification with DenseNet201-rmsprop.

This study explored multiple classifications providing a more comprehensive analysis of KOA severity compared to most other studies in Table 8 which focused only on 5-class or a single classification scheme. The study achieved high classification accuracy in the 2-class (87.7% CA) and 3-class (85.6% CA) classifications, showing strong performance in less complex classification tasks. The accuracy for the 5-class

Table 7. Training times of models.

Models	Optimizer	Classification	Training Times
DenseNet201	rmsprop	2-class	8hr 13min
	rmsprop	3-class	8hr 9min
	rmsprop	4-class	8hr 5min
	rmsprop	5-class	8hr 9min
Xception	adam	2-class	1hr 27min
	adam	3-class	1hr 28min
	rmsprop	4-class	1hr 21min
	rmsprop	5-class	1hr 21min
VGG16	sgdm	2-class	38min
	sgdm	3-class	38min
	sgdm	4-class	39min
	sgdm	5-class	39min
ResNet-50	adam	2-class	1hr 16min
	adam	3-class	1hr 16min
	rmsprop	4-class	1hr 7min
	adam	5-class	1hr 16min
EfficientNet	adam	2-class	3hr 12min
	adam	3-class	3hr 12min
	rmsprop	4-class	2hr 56min
	adam	5-class	3hr 10min

classification (67.8% CA) was higher than studies carried out by (Tiulpin et al. 2018) and (Ahmed and Imran 2024). However, it lower compared to several other studies such as Rehman and Gruhn (2024) with 93.27% CA, Saini et al. (2023) with 89.95% CA, Pi et al. (2023) and Yong et al. (2022) with 88.09% CA. This indicates room for improvement in handling the complexity of KOA severity levels. Nevertheless, there is a point to consider here that Rehman and Gruhn (2024) and Saini et al. (2023) used a different number of data in their study than our study. The number of data, especially the distribution of data into classes, significantly affects the performance of the models. Therefore, it would not be appropriate to directly compare our study with these two studies.

As shown in Table 8, Alshamrani et al. (2023) performed a 2-class classification on a dataset of 3,836 samples labeled KOA and Non-KOA. In contrast, our study performed a 2-class classification that included all KL grades, which were evaluated into 2 classes according to treatment status. Additionally, the number of data points analyzed in our study was different from theirs. A similar situation exists in the 3-class classification carried out by Sharma et al. (2023). Therefore direct performance comparison is not possible.

This study's primary advantages include its comprehensive evaluation of KOA severity through multiple classification levels (2, 3, 4, and 5 classes), flexibility in analysis, and high accuracy in less complex classifications. It also offers a detailed comparison of different models (ResNet-50, Xception, VGG16, EfficientNetb0, DenseNet201) and optimization algorithms (rmsprop, sgdm, adam). Moreover, through multiple classifications, the study reveals the behavior and impact of models and optimizers across classifications with varying numbers of classes. However, the study faces limitations due to data imbalance, which affects classification success, especially in more complex classifications. This situation highlights the need for strategies to address this

Table 8. Comparison of the literature.

Reference	Number of Data	Classification	Performance
Tiulpin et al. (2018)	27293	5-class (0, 1, 2, 3, 4 of KL grades)	CA: 66.71%
Yong et al. (2022)	8260	5-class (0, 1, 2, 3, 4 of KL grades)	CA_{macro}: 88.09%
Wang et al. (2022)	6380	5-class (0, 1, 2, 3, 4 of KL grades)	CA: 81.41%
Guan et al. (2022)	6567	5-class (0, 1, 2, 3, 4 of KL grades)	0.807 AUC (area under ROC curve)
Alshamrani et al. (2023)	3836	2-class (KOA and Non-KOA)	CA: 92%
Pi et al. (2023)	8260	5-class (0, 1, 2, 3, 4 of KL grades)	CA: 76.93%
Sharma et al. (2023)	4590	3-class (0, 3, 4 of KL grades)	CA: 93.84%
Saini et al. (2023)	3696	5-class (0, 1, 2, 3, 4 of KL grades)	CA: 89.95%
Rehman and Gruhn (2024)	1650	5-class (0, 1, 2, 3, 4 of KL grades)	CA: 93.27%
Ahmed and Imran (2024)	8260	5-class (0, 1, 2, 3, 4 of KL grades)	CA: 67%
This Study	**8260**	**5-class (0, 1, 2, 3, 4 of KL grades)** **4-class** **3-class** **2-class** } Different combination of KL grades	**CA: 67.8%** **CA: 81.5%** **CA: 85.6%** **CA: 87.7%**

imbalance. Additionally, while the study uses a significant dataset (8260 images), it is still smaller, potentially limiting generalizability and robustness.

4. Conclusion and Suggestions

This study exhaustively evaluated the severity of knee osteoarthritis (KOA) using various deep learning models (ResNet-50, Xception, VGG16, EfficientNetb0, DenseNet201) and optimization algorithms (rmsprop, sgdm, adam) across multiple classification levels (2, 3, 4, and 5 classes). The DenseNet201 model with the rmsprop optimizer consistently achieved the highest performance, particularly excelling in less complex classification tasks (2-class and 3-class). However, as the number of classes increased, all models experienced a decrease in performance metrics, highlighting the challenges associated with more complex classifications. Compared to existing literature, this study achieved competitive results, though there is room for improvement in handling the complexity of KOA severity levels. Notably, data imbalance significantly impacted classification success, particularly in higher class classifications.

To enhance future research, addressing data imbalance through advanced augmentation techniques or acquiring more balanced datasets is crucial. Additionally, exploring other optimization algorithms and hybrid models may improve classification accuracy. Expanding the dataset size and diversity could further enhance the generalizability and robustness of the models. Overall, this study's multi-level classification approach offers valuable insights into the performance dynamics of various models and optimizers, providing a foundation for further advancements in KOA severity detection.

Acknowledgment: This article is an original work and all results have not published previously. Also, this study did not receive any funding or research grants during the study, research or assembly of the manuscript.

Author contribution: Authors contributed to the all stages of the work, namely, collection of data, preprocessing, classification using ensemble deep learning models, the analysis of the results and also the writing of the manuscript, equally.

5. References

Ahmed, R., Imran, AS. 2024. Knee osteoarthritis analysis using deep learning and XAI on X-rays. IEEE Access, 12: 68870-68879. Doi: 10.1109/ACCESS.2024.3400987

Alshamrani, HA., Rashid, M., Alshamrani SS., Alshehri AH. 2023. Osteo-net: An automated system for predicting knee osteoarthritis from x-ray images using transfer-learning-based neural networks approach. Healthcare, 11(9): 1206. Doi: 10.3390/healthcare11091206

Anderson, AS., Loeser, RF. 2010. Why is osteoarthritis an age-related disease?. Best Pract Res Clin Rheumatol., 24(1): 15-26. Doi: 10.1016/j.berh.2009.08.006

Brahim, A., Jennane, R., Riad, R., Janvier, T., Khedher, L., Toumi H., Lespessailles, E. 2019. A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis Initiative. Computerized Medical Imaging and Graphics, 73: 11-18. Doi: 10.1016/j.compmedimag.2019.01.007

Chen, P. 2018. Knee osteoarthritis severity grading dataset. Mendeley Data, 1. Doi: 10.17632/56rmx5bjcr.1

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 1800-1807, USA

Deshpande, BR., Katz, JN., Solomon DH., Yelin, EH., Hunter, DJ., Messier, SP., Suter, LG., Losina, E. 2016. Number of persons with symptomatic knee osteoarthritis in the US: impact of race and ethnicity, age, sex, and obesity. Arthritis care & research, 68(12): 1743-1750. Doi: 10.1002/acr.22897

Guan, B., Liu, F., Mizaian, AH., Demehri, S., Samsonov, A., Guermazi, A., Kijowski, R. 2022. Deep learning approach to predict pain progression in knee osteoarthritis. Skeletal Radiol., 51(2):363-373. Doi: 10.1007/s00256-021-03773-0

He, K., Zhang, X., Ren, S., Sun, J. 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, p. 770-778. Doi: 10.48550/arXiv.1512.03385

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, KQ. 2017. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. p. 4700-4708. Doi: 10.48550/arXiv.1608.06993

Jakaite, L., Schetinin, V., Hladůvka, J., Minaev, S., Ambia, A., Krzanowski, W. 2021. Deep learning for early detection of pathological changes in x-ray bone microstructures: case of osteoarthritis. Sci. Rep., 11(1): 2294. Doi: 10.1038/s41598-021-81786-4

Kellgren, J., Bier, F. 1956. Radiological signs of rheumatoid arthritis: a study of observer differences in the reading of hand films. Ann. Rheum. Dis., 15(1): 55–60. Doi: 10.1136/ard.15.1.55.

Kim, DH., Kim, SC., Yoon JS., Lee YS. 2020. Are there harmful effects of preoperative mild lateral or patellofemoral degeneration on the outcomes of open wedge high tibial osteotomy for medial compartmental osteoarthritis? Orthop. J. Sports. Med., 8(6). Doi: 10.1177/2325967120927481

kneeosteoarthritis (2018). Knee Osteoarthritis Dataset with severity grading, https://huggingface.co/datasets/SilpaCS/kneeosteoarthritis/blob/main/data.zip, Access Data: 02.02.2024

Kohn, MD., Sassoon, AA., Fernando, ND. 2016. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin. Orthop. Relat. Res., 474(8): 1886–1893. Doi: 10.1007/s11999-016-4732-4

Kondal, S., Kulkarni, V., Gaikwad, A., Kharat, A., Pant, A. 2022. Automatic grading of knee osteoarthritis on the Kellgren-Lawrence scale from radiographs using convolutional neural networks. Advances in Deep Learning, Artificial Intelligence and Robotics: Proceedings of the 2nd International Conference on Deep Learning, Artificial Intelligence and Robotics, (ICDLAIR), p. 163-173, Springer.

Li, W., Yu, S., Yang, R., Tian, Y., Zhu, T., Liu, H., Jiao, D., Zhang, F., Liu, X., Tao, L. 2023. Machine learning model of resnet50-ensemble voting for malignant–benign small pulmonary nodule classification on computed tomography images. Cancers, 15(22): 5417. Doi: 10.3390/cancers15225417

Martel-Pelletier, J. (1999). Pathophysiology of osteoarthritis. Osteoarthritis and Cartilage., **7**(4): 371-373. Doi: 10.1053/joca.1998.0214

Mou, SF., Razzak, SA. 2023. Brain disease classification from MRI scans using EfficientNetB0 feature extraction. 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), IEEE. p. 336-340.

Nasser, Y., Jennane, R., Chetouani, A., Lespessailles, E., El Hassouni, M. 2020. Discriminative Regularized Auto-Encoder for early detection of knee osteoarthritis: data from the osteoarthritis initiative. IEEE transactions on medical imaging, 39(9): 2976-2984. Doi: 10.1109/TMI.2020.2985861

Pi, SW., Lee, BD., Lee, MS., Lee, HJ. 2023. Ensemble deep-learning networks for automated osteoarthritis grading in knee X-ray images. Scientific Reports**,** 13(1): 22887. Doi: 10.1038/s41598-023-50210-4

Raza, A., Phan, TL., Li, HC., Hieu, NV., Nghia, TT., Ching, CTS. 2024. A Comparative study of machine learning classifiers for enhancing knee Osteoarthritis Diagnosis. Information, 15(4): 183. Doi: 10.3390/info15040183

Rehman, SU., Gruhn, V. 2024. A Sequential VGG16+ CNN based Automated Approach with adaptive input for efficient detection of knee Osteoarthritis stages. IEEE Access,12: 62407 - 62415. Doi: 10.1109/ACCESS.2024.3395062

Saini, D., Khosla, A., Chand, T., Chouhan, DK., Prakash, M. 2023. Automated knee osteoarthritis severity classification using three-stage preprocessing method and VGG16 architecture. International Journal of Imaging Systems and Technology, 33(3): 1028-1047. Doi: 10.1002/ima.22845

Shaheed, K., Mao, A., Qureshi, I., Kumar, M., Hussain, S., Ullah, I., Zhang, X. 2022. DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition. Expert Systems with Applications, 191: 116288. Doi: 10.1016/j.eswa.2021.116288

Sharma, AK., Nandal, A., Dhaka, A., Koundal, D., Bogatinoska, DC., Alyami, H. 2022. Enhanced watershed segmentation algorithm-based modified ResNet50 model for brain tumor detection. BioMed Research International, 2022.Doi: 10.1155/2022/7348344

Sharma, G., Anand, V., Kumar, V. 2023. Classification of Osteo-Arthritis with the help of deep learning and transfer learning. 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), India. Doi: 10.1109/ICIRCA57980.2023.10220816

Shivadekar, S., Kataria, B., Hundekari, S., Wanjale, K., Balpande VP., Suryawanshi, R. 2023. Deep learning based image classification of lungs radiography for detecting COVID-19 using a Deep CNN and ResNet 50. International Journal of Intelligent Systems and Applications in Engineering, 11(1s): 241-250.

Simonyan, K., Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. Doi: 10.48550/arXiv.1409.1556

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. 2016. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, p. 2818-2826. Doi:10.1109/CVPR.2016.308

Tan, M., Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning, p. 6105-6114, California.

Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., Saarakkala, S. 2018. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci. Rep., 8(1): 1727. Doi: 10.1038/s41598-018-20132-7

Turkoglu, M. 2021. COVID-19 detection system using chest CT images and multiple kernels-extreme learning machine based on deep neural network. Irbm, 42(4): 207-214. Doi: 10.1016/j.irbm.2021.01.004

Vina, ER., Kwoh, CK. 2018. Epidemiology of osteoarthritis: literature update. Current opinion in rheumatology, 30(2):160-167. Doi: 10.1097/BOR.0000000000000479

Wang, Y., Wang, X., Gao, T., Du, L., Liu, W. 2021. An automatic knee osteoarthritis diagnosis method based on deep learning: data from the osteoarthritis initiative. J. Healthc. Eng., 2021: 1-10. Doi: 10.1155/2021/5586529

Wang, Y., Li, S., Zhao, B., Zhang, J., Yang, Y., Li, B. 2022. A ResNet-based approach for accurate radiographic diagnosis of knee osteoarthritis. CAAI Transactions on Intelligence Technology, 7(3): 512-521. Doi: 10.1049/cit2.12079

Wenham, C., Grainger, A., Conaghan, P. 2014. The role of imaging modalities in the diagnosis, differential diagnosis and clinical assessment of peripheral joint osteoarthritis. Osteoarthritis Cartilage., 22(10): 1692-1702. Doi: 10.1016/j.joca.2014.06.005

Yang, L., Xu, S., Yu, X., Long, H., Zhang, H., Zhu, Y. 2023. A new model based on improved VGG16 for corn weed identification. Front. Plant Sci., 14. Doi: 10.3389/fpls.2023.1205151

Yong, CW., Teo, K., Murphy, BP., Hum, YC., Tee, YK., Xia, K., Lai, KW. 2022. Knee osteoarthritis severity classification with ordinal regression module. Multimedia Tools and Applications, 81 (2): 1-13. Doi: 10.1007/s11042-021-10557-0