IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 April 2024, accepted 7 May 2024, date of publication 14 May 2024, date of current version 22 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3400987

RESEARCH ARTICLE

Knee Osteoarthritis Analysis Using Deep Learning and XAI on X-Rays

RAFIQUE AHMED AND ALI SHARIQ IMRAN, (Member, IEEE)
Intelligent Systems and Analytics (ISA) Research Group, Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway

Corresponding author: Ali Shariq Imran (ali.imran@ntnu.no)

ABSTRACT Knee osteoarthritis (OA) is a chronic disorder mainly arising from age-related factors affecting the knee joints. Its diagnosis is critically important and is usually done by medical practitioners using X-ray images. Although this process is accurate, it is time-consuming. X-ray images have facilitated the use of deep learning (DL) models for the automation of the diagnosis of knee OA, commonly employing convolutional neural network (CNN) based architectures. However, the lack of models' interpretability makes the results less trustworthy. This work builds on the existing state-of-the-art (SOTA) pre-trained DL models to understand the model's behavior in classifying highly complex knee OA cases utilizing a divide-and-conquer approach - from multi-class to a binary class for better results interpretability and explainability using explainable artificial intelligence (XAI). Five SOTA fine-tuned DL models are tested on Kellgren-Lawrence (KL) graded X-ray images. Both multi-class and binary-class (using the multiple subsets derived from the original dataset to examine how the models perform with different data combinations) classification approaches and their interpretability of findings using Gradient-weighted Class Activation Mapping (GradCAM) are undertaken in this study. The GradCAM visualization of EfficientNetb7 demonstrates that when the degree of variance between different classes increases, the model's efficiency in classifying knee OA also increases. Specifically, it becomes more effective at distinguishing normal and severe cases with 99.13% classification accuracy. However, the model's efficacy drops to 67% for other cases, indicating that it cannot classify knee OA as effectively as doctors.

INDEX TERMS Explainable artificial intelligence (XAI), deep learning, knee osteoarthritis, healthcare, diagnosis, classification.

I. INTRODUCTION

Knee osteoarthritis (OA) is a chronic degenerative disorder affecting the knee joint, resulting from conditions such as old age, pre-existing knee injuries, and being overweight. The typical evidence of this anomaly includes knee inflammation and pain [1]. Knee OA can be diagnosed using several approaches that involve physical examination by healthcare experts and imaging technology such as X-rays or magnetic resonance imaging (MRI). Imaging technology is superior to conventional diagnostic approaches in terms of providing a broader view of the disorder, as it enables visual confirmation by healthcare specialists [2]. Thanks to advancements in medical imaging technology and state-of-the-art techniques, it is now possible to classify and analyze medical images using various deep learning (DL)–based algorithms [3]. One of the most widely used DL methods is the convolutional neural network (CNN) based model, which is most often utilized in the domain of medical image classification [4]. Given the widespread utilization of CNN-based models in medical image classification, it is also possible to employ imaging-based diagnostics for knee OA [5]. To be more precise, images from X-rays can be utilized to develop a CNN-based model that can effectively assist in distinguishing between severe and normal cases of knee OA. This approach not only yields precise outcomes but also minimizes the time required for diagnosis [6].

The classification of X-rays for this objective is determined by the Kellgren-Lawrence (KL) classification method, often known as KL grading [7], [8]. According to the KL grading

The associate editor coordinating the review of this manuscript and approving it for publication was Genoveffa Tortora.

system, cases of knee OA can be classified into five grades, i.e., class/grade 0 represents a normal knee joint without any signs of OA, where the knee joint space is completely normal. Class/grade 1 indicates uncertainty regarding the presence of OA, class/grade 2 signifies a mild case of OA, class/grade 3 indicates a moderate case, and class/grade 4 corresponds to a severe condition characterized by narrowed joint space. The KL grading scheme is presented in Figure 1.

FIGURE 1. KL graded X-ray images from the dataset used for this study. Left to right: grade 0, grade 1, grade 2, grade 3, and grade 4.

In literature, several studies have been conducted on detecting, classifying, and analyzing knee OA using state-of-the-art algorithms. These studies are discussed in the section II, focusing on the knee OA classification using X-rays. This study primarily examines the effectiveness of advanced DL models in accurately classifying knee OA, comparing their performance to that of ratings by healthcare experts or doctors. This study aims to classify the knee OA using CNN-based pre-trained models. In addition, an explainable artificial intelligence (XAI) approach is employed to analyze the results produced from these models. To be more precise, the work is centered on:

- Use of multi-class approach for classification of knee OA using the fine-tuned CNN-based pre-trained models (VGG-16, VGG-19, ResNet-50, ResNet-101, and EfficientNetb7) taking into account the KL graded dataset.
- Assessment of the best-performing model for multi-class classification using the XAI approach Gradient-weighted Class Activation Mapping (GradCAM) to check if the model can accurately classify the X-ray images like the healthcare professionals by focusing on the region of interest (ROI).
- Use of inter-class/grade approach for binary classification to check if the limitations of multi-class approach can be explained better by the binary models.
- Assessment of the results obtained using binary classification to check whether the model considers the ROI for the classification of knee OA. Based on the ROI, the model can classify the X-ray images according to the KL grading.
- Exploring whether the size of test samples can impact the overall performance of the DL models while classifying knee OA.

The rest of the paper is structured as section II highlighting the previous work done for knee OA classification using X-ray images and the techniques utilized for explaining the model behavior, section III demonstrates the methodology adopted to achieve the objectives of this study, section IV presents the obtained results and section V is the conclusion.

II. RELATED WORK

DL algorithms have frequently been employed in the domain of medical image classification [9]. When it comes to the performance of these algorithms, the use of a balanced medical dataset results in good performance [10], [11]. Owing to the progress in this domain, a variety of XAI-based methodologies such as GradCAM [12], GradCAM++ [13], and EigenCAM [14] have been developed to comprehend the functioning of DL models during decision-making, enabling more comprehensive analysis and optimization of the DL model depending on the specific task [15], [16]. Various CNN-based models have been employed to classify knee OA using X-ray images [8], [17], [18]. Different XAI methodologies have been employed in existing research to comprehend and interpret the behavior of CNN-based models for knee OA classification tasks [19]. The following discussion focuses on the relevant research conducted on the classification of knee OA, specifically utilizing CNN-based models and X-ray images. Furthermore, the XAI method to comprehend the model's behavior during the decision-making process has also been provided for each study.

The authors in [20] proposed a method that utilizes a Deep Siamese CNN-based model to automatically assess the severity of knee OA by considering the KL grading system. Furthermore, GradCAM was employed to extract the factors that influenced the model's decisions. The study in [21], after detecting the knee joint using the you only look once (YOLO) v2 model, has fine-tuned variants of pre-trained CNN models such as VGG-19 and Inception v3. GradCAM has been utilized to comprehend the behavior of the model. Similarly, another study in [22] employed ResNet-34 with the transfer learning technique to forecast the likelihood of knee OA progression, specifically in terms of complete knee replacement and OA diagnosis. The utilization of GradCAM demonstrated the model's decision-making process during prediction.

Likewise, the study in [23] employed the "squeeze-and-excitation" (SE-ResNet) model to classify knee OA and utilized GradCAM to provide insight into the model's classification. The study investigated if incorporating extra patient information might enhance the precision of the knee OA classification model. The results demonstrated that including supplementary data enhances the model's accuracy in image classification. The research in [24] used a fine-tuned pre-trained DenseNet-169 model to perform a multiclass classification of knee OA. Additionally, the study utilized a saliency map approach to highlight the regions the model took into account for the classification.

The study in [25] also employed a DensNet and VGG-based framework to classify knee OA. Two approaches, GradCAM++ and layer-wise relevance propagation (LRP), were utilized to obtain. Another work in the study in [26] employed the YOLO model to identify the ROI, followed by the ResNet-50 model to classify knee OA. In this study, GradCAM was utilized to provide explanations for the model's results. A different study in [27] used ResNet-34 and

DenseNet-121 models to classify knee OA. This work used GradCAM to explain the model results.

Finally, the study in [28], employed four pre-trained models i.e. ResNet-34, VGG-19, DenseNet-121, and DenseNet-161 using transfer learning and fine-tuned. These models were combined in an ensemble to enhance the model's overall performance. The EigenCAM technique was employed to explain the results of model classification. Ultimately, the studies in [29] and [30] had gone one step ahead, i.e., instead of using a pre-trained CNN-based model, the researchers employed pre-trained vision transformers (ViT) to classify knee OA. Both studies utilized GradCAM to explain the behavior of models.

The literature analysis concludes that the use of pre-trained CNN models is widespread for the classification of knee OA. GradCAM has been frequently used to generate heatmaps that illustrate the specific regions the model focuses on when making decisions, hence facilitating the explanation of the model's results. It is worth mentioning that the use of an advanced CNN-based pre-trained model, such as EfficientNetb7 [31], for knee OA classification has to be further evaluated to understand its performance on the knee OA classification task. This pre-trained CNN-based model is a powerful neural network of the EfficientNet family model with the highest accuracy on the ImageNet dataset. Also, prior research on knee OA has primarily concentrated on multi-class classification. So far, no study has examined the binary classification approach, as far as we know, to evaluate the effectiveness of CNN-based models in this context, i.e., for which case (considering the normal grade and others) can the model provide satisfying results?

Also, explaining the model's performance dependency on the features and how closely the model performs like a medical doctor in classifying different stages of knee OA is worth mentioning. This analysis will determine whether we can trust and use complex DL models to predict the various stages of knee OA accurately or not.

III. METHODOLOGY

As mentioned in section II, EfficientNetb7 is the powerful network of the EfficientNet family; hence, this work focuses on using this model to classify knee OA. To compare the performance of EfficientNetb7 with other pre-trained CNN-based models, the variants of VGG and ResNet have been considered as these have been commonly used in literature. Transfer learning and the same fine-tuning method have been used for all models used in this study to ensure a fair comparison of the models on the same dataset. For fine-tuning the models, new layers have been added to all the pre-trained models i.e., VGG,[1] ResNet[2] and EfficientNet.[3] Upon evaluating and considering models' performance for multi-class classification, the binary classification approach has been used for further assessment of models employed in this study. The following subsections address this subject along with the description of the dataset used. GradCAM, a widely employed XAI technique in literature, has been utilized following its official implementation in Keras[4] to visualize the decision-making of best-performing models in all cases. This visualization has been done for both multi-class and binary-class classification. The details of models used for this work, along with GradCAM, are provided below:

- **VGG models:** VGG-16 and VGG-19 belong to the VGG family. The difference between the two is in network depth. VGG-16 has 16 layers, making it a simpler model than its counterpart, VGG-19. VGG-19 has 19 layers, which makes it more complex. This study used both VGG-16 and VGG-19. The pre-trained models on the ImageNet dataset were first fine-tuned. Later, fine-tuned models have been used in the study.
- **ResNet models:** ResNet101 and ResNet50 are part of the ResNet family. The difference between the two is in network depth. ResNet50 contains 50 layers, making it a more complicated model than VGG-16 and VGG-19 while remaining simpler than ResNet101. ResNet101 has 101 layers, making it more sophisticated than the architectures listed above. The pre-trained ResNet50 and ResNet101 models on the ImageNet dataset have been fine-tuned first and then employed in this work.
- **GradCAM:** GradCAM is an XAI approach used to visualize and analyze the region that the model considers when making a decision. The GradCAM's core principle uses the information about gradients that flows into the CNN's final convolutional layer to allocate significance scores to each neuron for a specific decision of importance. GradCAM has been used in this work to represent the regions targeted by the models in the present study, taking into account its intended application and task at hand. The goal is to achieve the purpose of this study, which is to determine whether the models can make decisions similar to those of physicians while focusing on the region of interest.

A. DATASET

The "Knee Osteoarthritis Severity Grading Dataset" [32] is a collection of X-ray images for knee OA graded using the KL grading system (0 to 4). The dataset has been developed by the University of Florida. The dataset contains a total number of 8260 images, out of which 70%, i.e., 5778 images, have been dedicated to the training set, 10%, i.e., 826, have been provided for the validation set, and 20%, i.e., 1656 images, have been used for the testing set. The dataset contains five classes.

B. MULTI-CLASS CLASSIFICATION

The dataset used for this work is graded based on the KL grading system, so a multi-class classification approach has

[1] https://keras.io/api/applications/vgg/
[2] https://keras.io/api/applications/resnet/
[3] https://keras.io/api/applications/efficientnet/
[4] https://keras.io/examples/vision/grad_cam/

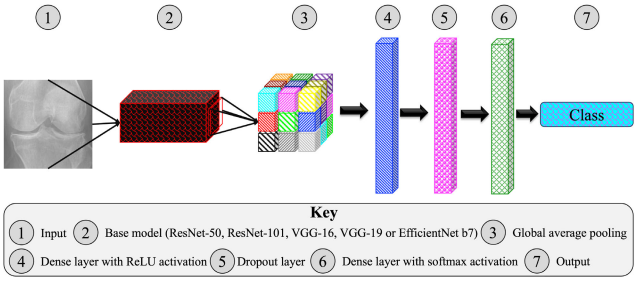

FIGURE 2. General architecture of models.

been utilized first. For this, two variants of VGG (VGG-16 and VGG-19), ResNet (ResNet-50 and ResNet-101), and EfficientNetb7 have been trained. For all models, a similar fine-tuning approach has been applied. The architecture of the fine-tuned model is presented in Figure 2.

In the fine-tuning stage, an output obtained from the pre-trained model is passed to a subsequent layer which averages the features to reduce the amount of features. These averaged features are then passed to the next layer, which is the ''Dense layer'', which learns from the previous features. A dropout layer has been added to regularize the output of the ''Dense layer'' before the prediction for output. Finally, the last layer takes the reduced features, and by taking advantage of the softmax function, it predicts the class for the current input image. With similar architecture, all the pre-trained models have been trained on the chosen dataset for multi-class classification.

C. BINARY CLASS CLASSIFICATION

Subsets from the original multi-class dataset have been created for binary class classification. The subsets are based on the pairs of class 0 (normal) and alternative class as the pair in the subset, that is, i). subset 1 contains class 0 (normal case) and class 1 (doubtful case), ii). subset 2 contains class 0 and class 2 (mild case), iii). subset 3 has class 0 and class 3 (moderate case), and iv). subset 4 contains class 0 and class 4 (severe case). As the samples present in the dataset for class 4 are very limited, certain pre-processing intensity-based image enhancement techniques like histogram equalization and contrast enhancement have been utilized to increase the training samples for class 4. Subset 1 consists of 4378 training and 935 testing samples, subset 2 contains 4347 training and 1086 testing images, subset 3 comprises 4557 training and 862 testing samples, and subset 4 possesses 2076 training and 690 testing samples.

On these subsets, all models with an architecture similar to that shown in Figure 2 have been trained for the classification of binary classes as presented in Table 1. The models trained for binary classification have the same architecture as the ones trained for multi-class classification with the only difference that, on the final dense layer instead of 5 classes, 2 classes have been defined considering the binary classification.

TABLE 1. Models trained for binary classification on subsets.

Subset	Classes	Total Models
1	0 and 1	5†
2	0 and 2	5†
3	0 and 3	5†
4	0 and 4	5†

† **Note:** Five models are VGG-16, VGG-19, ResNet-50, ResNet-101, and EfficientNetb7, respectively.

D. EXPERIMENTAL SETUP

For the completion of this work, Python version 3.8.18[5] has been used. All the models used for this work are based on Keras. The training of models has been carried out on the NVIDIA GeForce RTX 2080 GPU with the CUDA version 12.2.[6] To avoid over-fitting, an early stopping technique[7] has been used with patience value of 15, to monitor the validation loss of the models during training. This means that the model will stop training when the validation loss stops decreasing continuously until the defined patience value is reached. Table 2 presents the model training details.

TABLE 2. The parameters used for training all the models.

Parameter	Name/Value
Optimizer	Adam
Loss function	Categorical cross-entropy
Batch size	32
Number of epochs	1000
Learning rate	0.001
Patience	15

E. MODEL TRAINING

For the multi-class approach, five fine-tuned models have been trained for knee OA. The parameters presented in Table 2 have been used for training all the models. The training accuracy obtained for these five models is shown in Table 3.

TABLE 3. Training report of the models for multi-class classification.

Model	Accuracy [%]	Loss
VGG-16	54.60	1.05
VGG-19	57.13	1.00
ResNet-50	58.58	0.95
ResNet-101	57.97	0.98
EfficientNetb7	58.74	0.96

Similarly, twenty models have been trained for binary class classification using the subsets mentioned in section III-C. The training report for the models is provided in Table 4.

[5]https://www.python.org/downloads/release/python-3818/
[6]https://developer.nvidia.com/cuda-12-2-0-download-archive
[7]https://keras.io/api/callbacks/early_stopping/

IV. RESULTS

This section provides insights into the results obtained from the experiment. All the trained models for both multi-class and binary-class classification cases have been tested on test data. Model evaluation includes test accuracy, test loss, F1-score, recall, precision, and confusion matrix. Based on this evaluation of models' classification, for each case, GradCAM has been applied on the final convolutional layer [12] of the best-performing model to understand the model's decision-making behavior, i.e., can a DL-based model classify knee OA as classified/graded by the doctors? Furthermore, it has been demonstrated how the size of test samples can influence the accuracy of the models' test results.

A. MULTI-CLASS CLASSIFICATION

All five models trained for multi-class classification have been tested. The test report, as shown in Table 5, indicates that EfficientNetb7 achieves the highest accuracy compared to other models for multi-class classification. Further evaluation of these models shows that, except for the EfficientNetb7 model, none of the models could predict class 1, as shown in the "unique predictions" column in Table 5.

Although EfficientNetb7 predicts class 1, they are false predictions. Specifically, it predicts two samples from class 2 as class 1, as shown in Figure 3. Considering the behavior of the models, it is obvious that the DL-based models used in this work cannot classify knee OA as the healthcare professionals have classified. One reason behind this is that for the model, the data is highly correlated; hence, it cannot differentiate between class 0 and class 1.

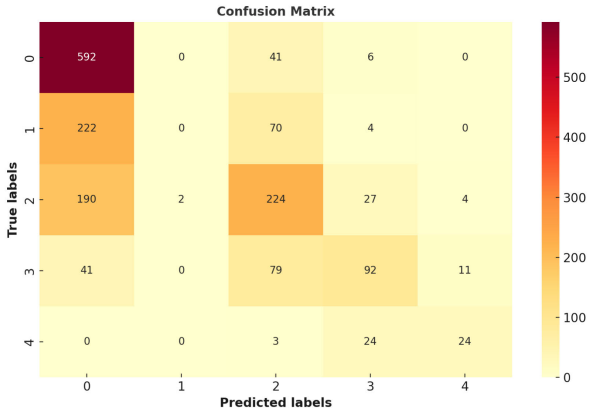

FIGURE 3. Confusion matrix of the best-performing multi-class classification model, i.e., EfficientNetb7.

Furthermore, GradCAM, as shown in Figure 4, has been applied to the predictions made by the best-performing model to visualize, while classifying, what is a region in the image where the model focuses on for classification. The GradCAM visualization shows that the model has focused on the knee-joint part for almost all the predictions with few exceptions. On average, considering its accuracy and confusion matrix, the model is unable to classify the test samples correctly. Even though the model focuses on the knee-joint part when making the decision, it cannot differentiate between classes because classes have been graded based on KL grading.

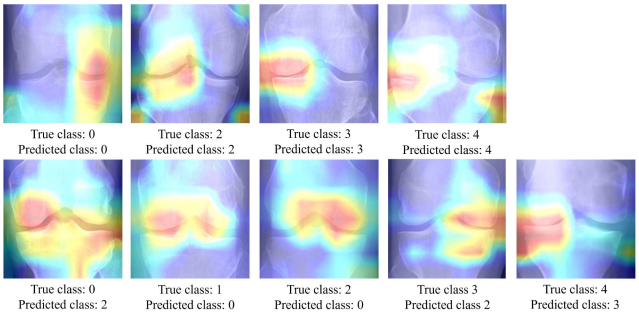

FIGURE 4. GradCAM† visualization of a few samples predicted by the model: row 1 is for true predictions, and row 2 is for false predictions. †Note: since none of the models make true predictions for class 1, hence, it is missing.

To check if the model can improve its performance, provided the subsets mentioned in the section for binary classification, all the trained models for binary class classification have also been tested and evaluated, as provided in the following section.

B. BINARY CLASS CLASSIFICATION

The trained models on the subsets have been tested on the test dataset. This result analysis of binary class classification can be divided as:

1) SUBSET 1: CLASS 0 AND 1

This was done to check how good the models were while differentiating between class 0 (normal) and class 1 (doubtful), as this is a peak case where all the models failed to make the predictions. The test report for all the models on this subset is provided in Table 7.

Though all models achieve almost the same test accuracy as shown in Table 7, EfficientNetb7 makes more correct predictions for class 1 compared to other models, which shows that it is more efficient than other models in this case, as presented in confusion matrix shown in Figure 5. The classification report presented in Table 8 shows that even EfficientNetb7 makes some correct predictions for class 1, but almost all the samples from class 1 are classified into class 0.

Similarly, GradCAM has been applied to the best-performing model in this case, as shown in Figure 6. In the case of the normal class (class 0), the model considers the features of the whole knee joint gap, whereas, in the case of the doubtful class (class 1), the model predicts by considering the features of the region in the knee joint where the gap is more than other sides.

2) SUBSET 2: CLASS 0 AND 2

Class 2, as per KL grading, is a mild case; hence, the X-rays for this subset differ from each other more compared

TABLE 4. Training report of the models for binary class classification.

Subset 1			Subset 2		
Model	Accuracy [%]	Loss	Model	Accuracy [%]	Loss
VGG-16	76.58	0.45	VGG-16	80.26	0.42
VGG-19	78.11	0.43	VGG-19	79.29	0.43
ResNet-50	76.06	0.46	ResNet-50	82.33	0.37
Resnet-101	76.03	0.45	Resnet-101	78.83	0.42
EfficientNetb7	75.46	0.49	EfficientNetb7	77.66	0.46
Subset 3			**Subset 4**		
Model	Accuracy [%]	Loss	Model	Accuracy [%]	Loss
VGG-16	94.55	0.15	VGG-16	99.74	0.01
VGG-19	91.92	0.18	VGG-19	99.25	0.08
ResNet-50	92.40	0.17	ResNet-50	96.72	0.07
Resnet-101	95.21	0.12	Resnet-101	99.95	0.002
EfficientNetb7	98.15	0.07	EfficientNetb7	99.88	0.005

TABLE 5. Test report of models for multi-class classification.

Model	Accuracy [%]	Loss	Unique predictions
VGG-16	51.87	1.12	{0,2,3,4}
VGG-19	50.42	1.14	{0,2,3,4}
ResNet-50	52.89	1.10	{0,2,3}
ResNet-101	51.99	1.11	{0,2,3,4}
EfficientNetb7	56.28	1.03	{0,1,2,3,4}

TABLE 6. Classification report of the best-performing models for multiclass classification.

	Precision	Recall	F1-score
0	0.93	0.57	0.70
1	0.00	0.00	0.00
2	0.50	0.54	0.52
3	0.41	0.60	0.49
4	0.47	0.62	0.53
accuracy			0.54
macro avg	0.46	0.46	0.45
weighted avg	0.76	0.56	0.63

TABLE 7. Test report of models for subset 1.

Model	Accuracy [%]	Loss
VGG-16	68.87	0.61
VGG-19	68.77	0.60
ResNet-50	68.34	0.61
ResNet-101	68.55	0.61
EfficientNetb7	67.16	0.61

TABLE 8. Classification report of the best-performing models for subset 1.

	Precision	Recall	F1-score
0	0.92	0.70	0.79
1	0.12	0.44	0.19
accuracy			0.67
macro avg	0.52	0.57	0.49
weighted avg	0.85	0.67	0.74

to subset 1. This means that the models should be more capable of differentiating between these classes, i.e., 0 and 2. The test report of all the models is given in Table 9, whereas the classification report and the confusion matrix

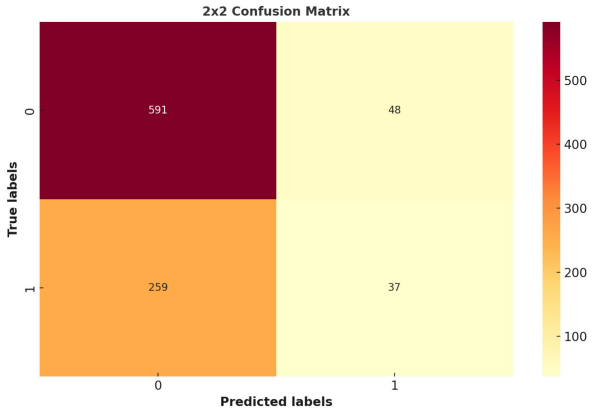

FIGURE 5. Confusion matrix of the best-performing model for subset 1, i.e., EfficientNetb7.

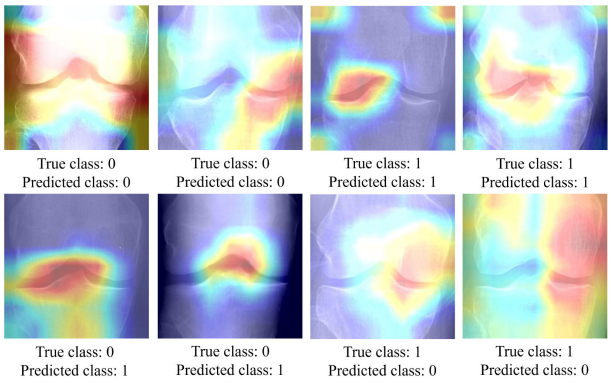

FIGURE 6. GradCAM visualization of a few samples predicted by the model: row 1 is for true predictions, and row 2 is for wrong predictions.

of the best-performing model are provided in Table 10 and Figure 7. The confusion matrix shows that though the results obtained are a bit improved compared to the previous case, the models still need variation between classes to classify knee OA correctly.

GradCAM has been applied to the best-performing model as illustrated in Figure 8 to understand the region in which the model focuses when making the classification decision. In the

TABLE 9. Test report of models for subset 2.

Model	Accuracy [%]	Loss
VGG-16	73.48	0.54
VGG-19	73.11	0.53
ResNet-50	71.82	0.53
ResNet-101	74.76	0.52
EfficientNetb7	76.51	0.50

TABLE 10. Classification report of the best-performing models for subset 2.

	Precision	Recall	F1-score
0	0.89	0.75	0.82
2	0.58	0.79	0.67
accuracy			0.77
macro avg	0.74	0.77	0.74
weighted avg	0.80	0.77	0.77

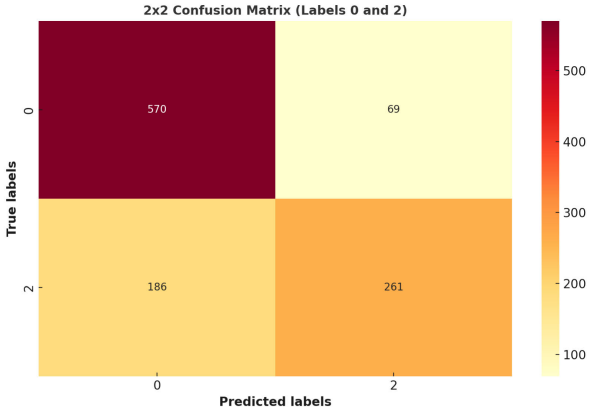

FIGURE 7. Confusion matrix of the best-performing model for subset 2, i.e., EfficientNetb7.

case of the normal class (class 0), the model considers the features of the whole knee joint gap or a significant portion of the gap, whereas, in the case of the mild class (class 2), the model predicts by considering the features of the region in the knee joint where the gap is more than other sides, for example in case of Figure 8 (row 1, image three and four), the left side of the joint has more gap as compared to right side.

3) SUBSET 3: CLASS 0 AND 3

For this subset, both classes differ significantly from the previous subgroup's classes. Class 3, as per KL grading, is a moderate case, i.e., the space between the knee joints is halfway closing. The test report of all the models for this case is given in Table 11. The classification report and the confusion matrix for the best-performing model are provided in Table 12 and Figure 9, respectively. The confusion matrix shows that the model has improved performance. This shows that the model's capability of differentiating between classes has increased, i.e., it can classify more correctly than the previous two subsets as the images differ from each other.

In the next step of model evaluation, GradCAM has been applied to the best-performing model. GradCAM

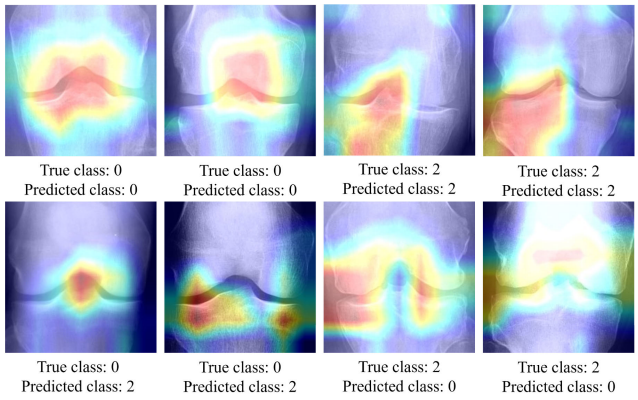

FIGURE 8. GradCAM visualization of a few samples predicted by the model: row 1 is for true predictions, and row 2 is for wrong predictions.

TABLE 11. Test report of models for subset 3.

Model	Accuracy [%]	Loss
VGG-16	89.09	0.28
VGG-19	86.31	0.32
ResNet-50	88.97	0.28
ResNet-101	88.16	0.29
EfficientNetb7	91.53	0.20

TABLE 12. Classification report of the best-performing models for subset 3.

	Precision	Recall	F1-score
0	0.95	0.93	0.94
3	0.81	0.86	0.83
accuracy			0.92
macro avg	0.88	0.90	0.89
weighted avg	0.92	0.92	0.92

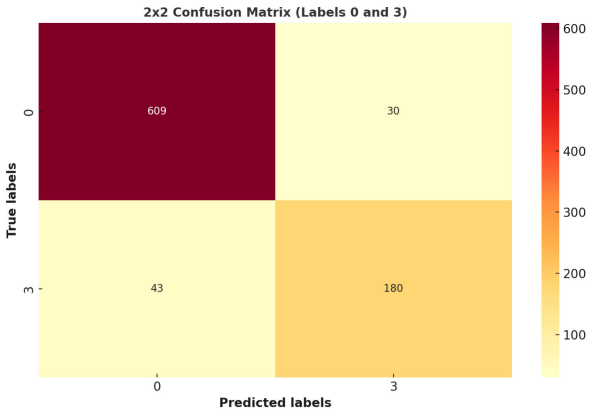

FIGURE 9. Confusion matrix of the best-performing model for subset 3, i.e., EfficientNetb7.

visualizations for some of the samples predicted by the model are presented in Figure 10. In the case of the normal class (class 0), the model considers the features of the whole knee joint gap or a significant portion of the gap, whereas, in the case of the moderate class (class 3), the model predicts by

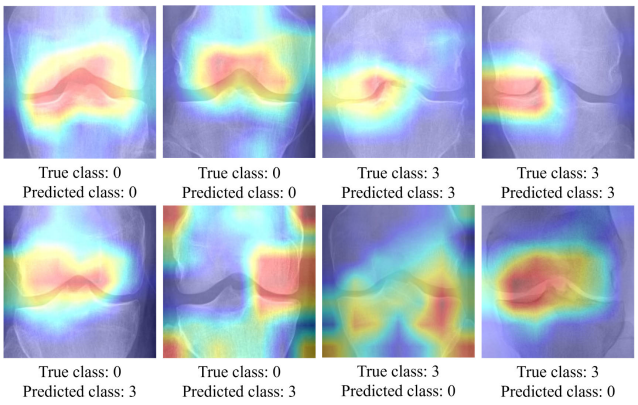

FIGURE 10. GradCAM visualization of a few samples predicted by the model: row 1 is for true predictions, and row 2 is for wrong predictions.

considering the features of the region in the knee joint where the gap is minimal as compared to the rest of the region in joint, for example in case of Figure 10 (row 1, image three and four), the left side of the joint has less gap as compared to right side. This shows that the variation between classes allows the model to focus on the region that should be considered while classifying, i.e., the gap between knee joints.

4) SUBSET 4: CLASS 0 AND 4

This subset contains the normal (class 0) and the severe case (class 4). Considering the nature of classes, all the models should be highly efficient in separating these two classes. For class 0, all the images have normal space between knee joints, whereas for class 4, all the images have narrow spaces between the joints. The test report of all models tested on this subset is provided in Table 13. For the best-performing model, the classification report and confusion matrix are provided in Table 14 and Figure 11, respectively.

TABLE 13. Test report of models for subset 4.

Model	Accuracy [%]	Loss
VGG-16	98.55	0.05
VGG-19	98.11	0.08
ResNet-50	98.40	0.04
ResNet-101	97.68	0.06
EfficientNetb7	99.13	0.03

TABLE 14. Classification report of the best-performing models for subset 4.

	Precision	Recall	F1-score
0	1.00	0.99	1.00
4	0.92	0.96	0.94
accuracy			0.99
macro avg	0.96	0.98	0.97
weighted avg	0.99	0.99	0.99

Finally, GradCAM has been applied to the best-performing model to visualize the region where the model focuses while

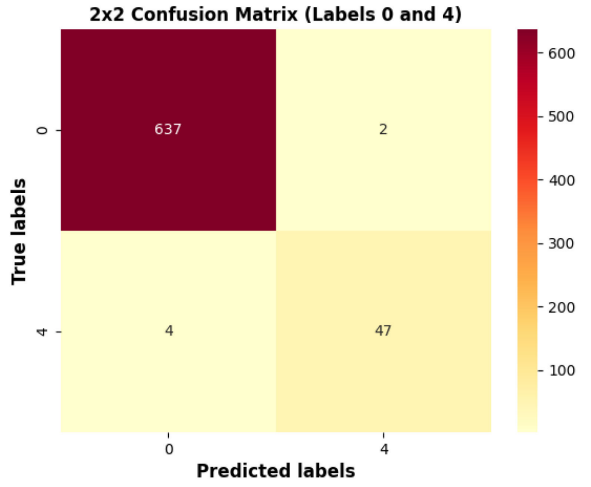

FIGURE 11. Confusion matrix of the best-performing model for subset 4, i.e., EfficientNetb7.

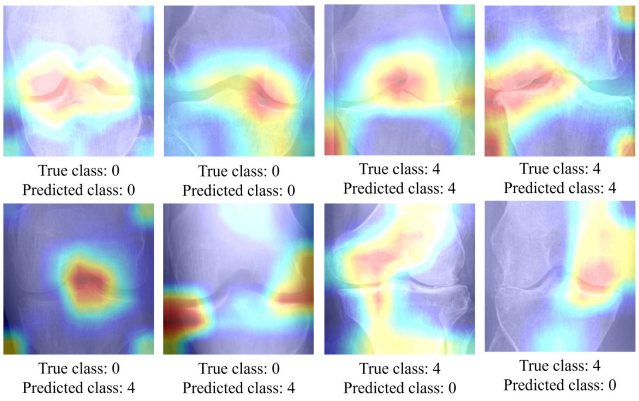

FIGURE 12. GradCAM visualization of a few samples predicted by the model: row 1 is for true predictions, and row 2 is for wrong predictions.

making the decision, as shown in Figure 12. Here, when the model considers the whole knee-joint region, then it makes the correct prediction, but when it considers a smaller portion of the joint or region other than the joint area, it makes false decisions and causes model confusion. Since there is no gap between joints, the true label is class 4, for which the model predicts class 4 as well, similar to class 0. Again, it can be seen that the decision-making capability of the model is heavily affected by the huge variation between classes, indicating that the model can perform efficiently and effectively for normal and severe cases only.

C. IMPACT OF TEST SAMPLES' SIZE ON MODELS' ACCURACY

Figure 11 illustrates that there is a significant disparity in the number of test samples between class 4 and class 1. Therefore, it has been investigated whether the quantity of test samples can affect the performance of the models. To tackle the problem of limited data, new additional samples have been generated for class 4 by employing histogram

TABLE 15. Test report of models after increased samples for class 4.

Model	Accuracy [%]	Loss
VGG-16	96.48	0.10
VGG-19	94.17	0.26
ResNet-50	96.68	0.09
ResNet-101	94.47	.14
EfficientNetb7	95.78	0.12

equalization and contrast enhancement techniques. With this, the number of samples for class 4 grows from 51 to 357. The models' evaluation of this new data leads to a decrease in test accuracies compared to the ones presented in Table 13, as shown in Table 15.

V. CONCLUSION

Knee OA is a medical disorder for elderly people. Diagnosing knee OA requires doctors to conduct a rigorous evaluation of the affected individual's knee joint X-rays. In this study, we used fine-tuned CNN-based pre-trained models to classify knee OA using multi-class and binary-class classification approaches. The objective was to determine if CNN-based models can accurately classify knee OA based on the KL grading system. Specifically, the focus was on determining if models' classifications align with those provided by doctors. For this study, two variants of VGG, ResNet, and one of EfficientNetb7 have been employed. From the results, it was observed that the models could not accurately predict any sample belonging to the doubtful class. GradCAM was utilized to analyze the decision-making process. The GradCAM results indicated that the model mostly focused on the knee-joint region when making decisions. Nevertheless, the model's results were highly incorrect due to its confusion in distinguishing between various classes. We also considered a binary class classification approach to determine the extent to which the models perform comparably to doctors in classifying X-rays based on KL grading. The results obtained from experiments demonstrated that EfficientNetb7 consistently outperforms competing models in all scenarios. Also, the GradCAM analysis indicated that the model considered the knee-joint region while making decisions for binary-class cases. The model had good results specifically for subsets, including normal and severe classes, compared to the other subsets. While the attained accuracy is satisfactory for this subset, the GradCAM findings indicated that the model continues to deliver false predictions even when considering the ROI. These incorrect results given by the model indicate that the models are less proficient at distinguishing between the classes compared to the doctor in the context of multi-class classification as well as binary class.

REFERENCES

[1] D. T. Felson, "Osteoarthritis of the knee," *New England J. Med.*, vol. 354, no. 8, pp. 841–848, Feb. 2006.
[2] M. J. Lespasio, N. S. Piuzzi, M. E. Husni, G. F. Muschler, A. Guarino, and M. A. Mont, "Knee osteoarthritis: A primer," *Permanente J.*, vol. 21, no. 4, pp. 75–79, Dec. 2017.
[3] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.
[4] Y. Yu, E. Favour, and P. Mazumder, "Convolutional neural network design for breast cancer medical image classification," in *Proc. IEEE 20th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2020, pp. 1325–1332.
[5] B. Liu, J. Luo, and H. Huang, "Toward automatic quantification of knee osteoarthritis severity using improved faster R-CNN," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 3, pp. 457–466, Jan. 2020.
[6] R. T. Wahyuningrum, L. Anifah, I. K. Eddy Purnama, and M. Hery Purnomo, "A new approach to classify knee osteoarthritis severity from radiographic images based on CNN-LSTM method," in *Proc. IEEE 10th Int. Conf. Awareness Sci. Technol. (iCAST)*, Oct. 2019, pp. 1–6.
[7] M. D. Kohn, A. A. Sassoon, and N. D. Fernando, "Classifications in brief: Kellgren–Lawrence classification of osteoarthritis," *Clin. Orthopaedics Rel. Res.*, vol. 474, no. 8, pp. 1886–1893, Aug. 2016.
[8] S. S. Abdullah and M. P. Rajasekaran, "Automatic detection and classification of knee osteoarthritis using deep learning approach," *La Radiologia Medica*, vol. 127, no. 4, pp. 398–406, Mar. 2022.
[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sanchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[10] A. Auriemma Citarella, L. Di Biasi, F. De Marco, and G. Tortora, "ENTAIL: Yet another amyloid fIbrils cLassifier," *BMC Bioinf.*, vol. 23, no. 1, p. 517, Dec. 2022.
[11] A. Ahmed, A. S. Imran, A. Manaf, Z. Kastrati, and S. M. Daudpota, "Enhancing wrist abnormality detection with YOLO: Analysis of state-of-the-art single-stage detection models," *Biomed. Signal Process. Control*, vol. 93, Jul. 2024, Art. no. 106144.
[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[13] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.
[14] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.
[15] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review," *Comput. Biol. Med.*, vol. 166, Oct. 2023, Art. no. 107555.
[16] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: A comprehensive review," *Artif. Intell. Rev.*, vol. 55, pp. 1–66, Jun. 2022.
[17] D. R. Sarvamangala and R. V. Kulkarni, "Grading of knee osteoarthritis using convolutional neural networks," *Neural Process. Lett.*, vol. 53, no. 4, pp. 2985–3009, May 2021.
[18] S. M. Ahmed and R. J. Mstafa, "Identifying severity grading of knee osteoarthritis from X-ray images using an efficient mixture of deep learning and machine learning models," *Diagnostics*, vol. 12, no. 12, p. 2939, Nov. 2022.
[19] Y. Xin Teoh, A. Othmani, S. Li Goh, J. Usman, and K. Wee Lai, "Deciphering knee osteoarthritis diagnostic features with explainable artificial intelligence: A systematic review," 2023, *arXiv:2308.09380*.
[20] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach," *Sci. Rep.*, vol. 8, no. 1, p. 1727, Jan. 2018.
[21] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," *Computerized Med. Imag. Graph.*, vol. 75, pp. 84–92, Jul. 2019.
[22] K. Leung, B. Zhang, J. Tan, Y. Shen, K. J. Geras, J. S. Babb, K. Cho, G. Chang, and C. M. Deniz, "Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: Data from the osteoarthritis initiative," *Radiology*, vol. 296, no. 3, pp. 584–593, Sep. 2020.
[23] D. H. Kim, K. J. Lee, D. Choi, J. I. Lee, H. G. Choi, and Y. S. Lee, "Can additional patient information improve the diagnostic performance of deep learning for the interpretation of knee osteoarthritis severity," *J. Clin. Med.*, vol. 9, no. 10, p. 3341, Oct. 2020.

[24] K. A. Thomas, Ł. Kidzinski, E. Halilaj, S. L. Fleming, G. R. Venkataraman, E. H. Oei, G. E. Gold, and S. L. Delp, "Automated classification of radiographic knee osteoarthritis severity using deep neural networks," *Radiol., Artif. Intell.*, vol. 2, no. 2, Mar. 2020, Art. no. e190065.

[25] Md. R. Karim, J. Jiao, T. Döhmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann, and S. Decker, "DeepKneeExplainer: Explainable knee osteoarthritis diagnosis from radiographs and magnetic resonance imaging," *IEEE Access*, vol. 9, pp. 39757–39780, 2021.

[26] Y. Wang, X. Wang, T. Gao, L. Du, and W. Liu, "An automatic knee osteoarthritis diagnosis method based on deep learning: Data from the osteoarthritis initiative," *J. Healthcare Eng.*, vol. 2021, pp. 1–10, Sep. 2021.

[27] Y. Wang, Z. Bi, Y. Xie, T. Wu, X. Zeng, S. Chen, and D. Zhou, "Learning from highly confident samples for automatic knee osteoarthritis severity assessment: Data from the osteoarthritis initiative," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 1239–1250, Mar. 2022.

[28] T. Tariq, Z. Suhail, and Z. Nawaz, "Knee osteoarthritis detection and classification using X-rays," *IEEE Access*, vol. 11, pp. 48292–48303, 2023.

[29] E. A. Alshareef, F. O. Ebrahim, Y. Lamami, M. B. Milad, M. S. Eswani, S. A. Bashir, S. A. Bshina, A. Jakdoum, A. Abourqeeqah, and M. O. Elbasir, "Knee osteoarthritis severity grading using vision transformer," *J. Intell. Fuzzy Syst.*, vol. 43, pp. 8303–8313, Nov. 2022.

[30] Z. Wang, A. Chetouani, M. Jarraya, D. Hans, and R. Jennane, "Transformer with selective shuffled position embedding and key-patch exchange strategy for early detection of knee osteoarthritis," 2023, *arXiv:2304.08364*.

[31] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[32] *Dataset: Knee Osteoarthritis Severity Grading Dataset*. Accessed: Oct. 2023. [Online]. Available: https://data.mendeley.com/datasets/56rmx5bjcr/1

RAFIQUE AHMED received the bachelor's degree in electronics engineering from the Mehran University of Engineering and Technology (MUET), Jamshoro, Pakistan, in 2021. He is currently pursuing the Erasmus Mundus Joint Master Degree (EMJMD) in computational color and spectral imaging (COSI) with the Norwegian University of Science and Technology, Gjøvik, Norway. His research interests include machine learning, deep learning, computer vision, and image processing.

ALI SHARIQ IMRAN (Member, IEEE) received the master's degree in software engineering and computing from the National University of Sciences and Technology (NUST), Pakistan, in 2008, and the Ph.D. degree in computer science from the University of Oslo (UiO), Norway, in 2013. He is currently an Associate Professor with the Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), Norway. He is also a member of the Intelligent Systems and Analytics Research Group, NTNU. With more than 15 years of teaching and research experience, he devised innovative ways to design effective multimedia learning objects and integrate the teaching-research nexus frameworks at the graduate level. He served as a Commission Member of the Ministry of Education of Macedonia in setting up Mother Theresa University in Skjope. He leads a capacity-building project called CONNECT (https://norpart-connect.com) funded by the Higher Education Commission of Norway, DIKU, under the NORPART scheme as a Coordinator and three Erasmus+ KA2 projects (PhDICTKES (https://phdictkes.eu), RAPID, and TKAEDiT) as the Project Manager at NTNU, along with an excited mini-project funded by NTNU. He is also leading a research group on Deep NLP (http://deep-nlp.net) and specializes in applied deep learning research to address various multi-modality media analysis application areas for audio-visual and text processing. He has coauthored over 100 peer-reviewed journals and conference publications. He is a member of ACM. He has served as an editor and a reviewer for many reputed journals.