IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 April 2023, accepted 11 May 2023, date of publication 16 May 2023, date of current version 23 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3276810

RESEARCH ARTICLE

Knee Osteoarthritis Detection and Classification Using X-Rays

TAYYABA TARIQ[1], ZOBIA SUHAIL[1], AND ZUBAIR NAWAZ[2]
[1]Department of Computer Science, University of the Punjab, Lahore 54590, Pakistan
[2]Department of Data Science, University of the Punjab, Lahore 54590, Pakistan

Corresponding author: Tayyaba Tariq (tayyaba.tariq@pucit.edu.pk)

ABSTRACT Knee osteoarthritis is a common form of arthritis, a chronic and progressive disease recognized by joint space narrowing, osteophyte formation, sclerosis, and bone deformity that can be observed using radiographs. Radiography is regarded as the gold standard and is the cheapest and most readily available modality. X-ray images are graded using Kellgren and Lawrence's (KL) grading scheme according to the order of severity of osteoarthritis from normal to severe. Early detection can help early treatment and hence slows down knee osteoarthritis degeneration. Unfortunately, most of the existing approaches either merge or exclude perplexing grades to improve the performance of their models. This study aims to automatically detect and classify knee osteoarthritis according to the KL grading system for radiographs. We have proposed an automated deep learning-based ordinal classification approach for early diagnosis and grading knee osteoarthritis using a single posteroanterior standing knee x-ray image. An Osteoarthritis Initiative(OAI) based dataset of knee joint X-ray images is chosen for this study. The dataset was split into the training, testing, and validation set with a 7: 2: 1 ratio. We took advantage of transfer learning and fine-tuned ResNet-34, VGG-19, DenseNet 121, and DenseNet 161 and joined them in an ensemble to improve the model's overall performance. Our method has shown promising results by obtaining 98% overall accuracy and 0.99 Quadratic Weighted Kappa with a 95% confidence interval. Also, accuracy per KL grade is significantly improved. Furthermore, our methods outperform state-of-the-art automated methods.

INDEX TERMS Detection and classification, knee osteoarthritis, ordinal classification, X-rays.

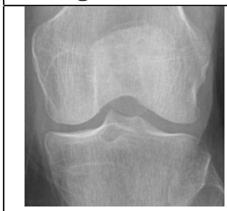
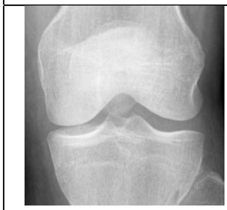
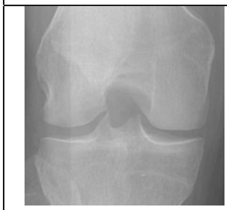
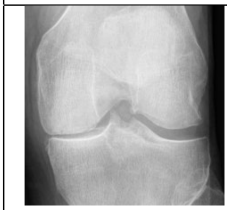
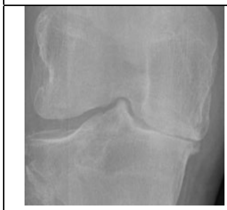
I. INTRODUCTION

Osteoarthritis (OA) is a disease with multiple factors, making it difficult to diagnose, detect and treat [1], [2]. It is a chronic degenerative disorder characterized by cartilage deterioration, eventually leading to bone deterioration. Knee osteoarthritis (KOA) is one type of osteoarthritis that affects the knee joint. Physical symptoms include pain, stiffness, swelling, and limited joint movements. Risk factors are age, gender, genetics, race, obesity, injury, vitamin D deficiency, and lifestyle [1], [2], [3], [4]. It is a progressive disease and has different stages of severity. According to a recent study [5], the global KOA prevalence is 16%. As reported by World Health Organization (WHO), this disease is more prevalent in women, i.e., 18.0%, than in men, i.e., 9.6%, and it affects people over 60 worldwide [1]. Knee osteoarthritis diagnosis is usually based on symptoms, arthroscopy, X-rays, and Magnetic Resonance Imaging (MRI). However, the early stages of OA are often hidden. In addition, there is a weak relationship between the degree of pain and dysfunction and the severity level of OA represented by the image. Thus, there is a need for a better diagnostic technique to detect OA in the initial stages. OA-related bio-markers can help in this situation [1].

Radiographs or X-rays to assess pain and restlessness are the foundation for detecting and diagnosing KOA [1], [3]. Key features that can be observed using X-rays are Joint Space Narrowing (JSN), osteophytes, cyst formation, and subchondral sclerosis. JSN refers to the loss of protective cartilage between knee joints. Osteophyte is a bony lump formed on bones or joints, while subchondral sclerosis is the abnormal thickening of the bone [3]. Kellgren and Lawrence's (KL)

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

TABLE 1. Kellgren and Lawrence's grading (KL) grading scheme.

Image	Grade Description
	Grade 0 (Normal) is assigned to normal bones and no symptoms on X-rays.
	Grade 1 (Doubtful) depicts doubtful JSN and the possibility of osteophytes.
	Grade 2 (Mild) specifies definite osteophytes and possible JSN.
	Grade 3 (Moderate) indicates multiple osteophytes with possible bone deformity.
	Grade 4 (Severe) shows large osteophytes, definite JSN, and severe sclerosis.

grading system is a semiquantitative method to assign grades to radiographs(x-rays) for KOA severity [6], [7]. According to this system, ordinal numbers are assigned according to the severity level for classification. KL grades are described in Table 1.

For computer-aided diagnosis and classification, images obtained from imaging modalities are processed using image processing and computer vision-based techniques. These techniques include image enhancement, segmentation, texture, and shape analysis [3], [8], [9]. Image segmentation approaches are applied to detect and localize knees in an image. Texture and shape features are then fed into the machine learning classifiers.

Recently, Deep Learning (DL) based Convolutional Neural Networks (CNN) have gained tremendous attention for computer vision and image analysis tasks. Several studies apply popular CNN architectures such as ResNet [10], VGG [11], and DenseNet [12] for various classification tasks. Transfer learning can pull off the benefits of existing architecture and its learned weights to save computation power and resources [13]. In this scenario, an ImageNet pre-trained deep learning network is only implemented as a feature extractor. The other method is fine-tuning this network to specialize for a particular dataset [11]. These networks are also applied for KOA classification.

Many automated methods and physician's grading systems are less reliable as they misclassify a KL grade to its nearby grade. In addition, since there are very few morphological and feature changes in successive KL grades, it becomes difficult to differentiate different grades [14].

The mainstream studies treat it as a multi-class classification task and ignore the inherent ordinal nature within KL grades [14]. Ordinal regression means that input values are continuous, and there is some order between the classes. Image ordinal classification uses handcrafted features passed to some regressor or classifier [15]. KOA severity level prediction is also an image classification task in which each KL grade is assigned a distinct category. KL grades maintain ordering information about the severity level of OA [14], [16], [17].

The KL grading is subjective. This depends on the expertise of the radiologist or radiograph reader. Inter-observer and Intra-observer reading reliability of quadratic Kappa can vary from 0.56 to 0.67 [18]. Therefore, it is challenging to build a consensus on the grade of a radiograph, specifically in the earlier stages of KOA.

In most studies, the initial stages of KOA have little accuracy [8], [19], [20], [21], [22] while, in some studies, the most difficult stages are merged for classification [23]. At the same time, some methods try combining X-ray features with other clinical data to improve performance [24]. As a result, the earlier it can be diagnosed, the earlier it can be treated, and knee degeneration leading to total knee replacement can be avoided. We need an assisting tool to prevent the development and worsen the disease.

In this study, we have attempted to resolve this gap by improving prediction accuracies for all KL grades.

The Osteoarthritis Initiative (OAI) is a multi-centric, ten-year, prospective observational study of KOA. They recruited 4796 men and women, sponsored by the National Institute of Health (part of the Department of Health and Human Services), with data from over 431,000 clinical and imaging visits and almost 26,626,000 images in this archive [25]. Knee X-ray images used in our study are based on this dataset [26].

1) Our method works on unilateral posteroanterior knee X-rays and does not require images from multiple angles or other clinical data.
2) Four ImageNet-based pre-trained models were fine-tuned. State-of-the-art results are achieved by each of these models individually.
3) An ensemble model is developed by combining predictions from the above-mentioned base models to improve overall performance.
4) Ordinal classification is considered using a customized ordinal loss function.
5) Finally, significant features identified by the model are visualized using class-specific heatmaps.

We focused on early diagnosis of KOA and improved the model's performance for all grades. Early diagnosis and the correct prediction of KOA grade will help physicians devise a better strategy for treating KOA at the early stages and will reduce the cost borne by the patients due to delayed detection [1]. Our model outperforms all existing methods to the best of our knowledge.

II. RELATED WORK

Anifah et al. [8] have used Contrast Limited Adaptive Histogram Equalization and Template matching for KOA grading. Their classification accuracy for KL grade 0 is 93.8%, for KL grade 1 is 70%, KL grade 2 is 4%, KL grade 3 is 10%, and KL grade 4 is 88.9%.

Chen et al. [19] have used YOLO2 Network for fully automated knee joint detection. They have tested multiple fine-tuned networks for classification, e.g., ResNet, VGG, and DenseNet. Their best-attained accuracy is 69.7%, and the Mean Absolute Error is 0.344.

Thomas et al. [27] developed an automated CNN-based model for knee osteoarthritis severity grading from radiographs. They had 32116 training images, 4074 for tuning, and 4090 for testing. Their reported accuracy is 0.71, and their obtained F1 score is 0.70 for the test set.

In another work, ResNet with Convolution Block attention Module (CBAM) has been implemented [20]. They used the Osteoarthritis Initiative (OAI) X-ray dataset for training and testing. Their obtained accuracy is 74.81%, Mean Squared Error (MSE) is 0.36, and Quadratic kappa score is 0.88. The accuracy for KL grade 0 is 83.81%, KL grade 1 is 48.66%, KL grade 2 is 65.68%, KL grade 3 is 85.67%, and KL grade 4 is 90.21%.

Another work is done [18] for classification to evaluate the effect of additional patient information on the prediction of the DL model for KOA severity. Two types of experiments were performed. First, only imaging information was used; in the second experiment, image data and clinical information were input. They used a private dataset of 3464 training images, 386 validation images, and 516 testing images. A CNN was developed with the six Squeeze and Excitation ResNet (SE-ResNet) modules. Their obtained AUCs with only image data for KL grades 0-4 are 0.91, 0.80, 0.69, 0.86, and 0.96. For DL with image data and patient information obtained, AUCs for KL grades 0-4 are 0.97, 0.85, 0.75, 0.86, and 0.95. It has been reported that KL grade 2 is the most complex and confusing to predict for the DL model. They observed that patient information improved the AUC for each stage.

In another study, [23], 25873 training images, 7779 validation images, and 5941 testing images are taken from the OAI dataset. Left and right Knee joints are localized using U-Net. Demographic information, i.e., age, gender, BMI, is fed to the DenseNet. Their method achieved sensitivity results for four levels of OA are normal 83.7%, mild 70.2%, moderate 68.9% Severe 86.0%, and Specificity normal 86.1%, mild 83.8% moderate 97.1%, and severe 99.1%. They have eliminated the KL1 doubtful grade. Moreover, their internal radiologists highlight the inter-observer reliability of KL classification varies from 0.51 to 0.89. It has also been observed that these wrong classifications are mainly made for adjacent KL grades.

A transfer learning-based approach was applied using pre-trained ResNet-34 architecture [24]. They used the OAI data of 728 participants. Knee radiographs and other clinical assessments, e.g., age, gender, ethnicity, and BMI, are fed to the model for possible KL grade and OA progression prediction. Using transfer learning, their achieved AUCs for KL grade prediction are 0.93, 0.80, 0.88, 0.96, and 0.99.

III. METHODOLOGY

In the following section, we have described the complete methodology to achieve the objectives mentioned above in this study. First, section III-A describes the dataset. Then,

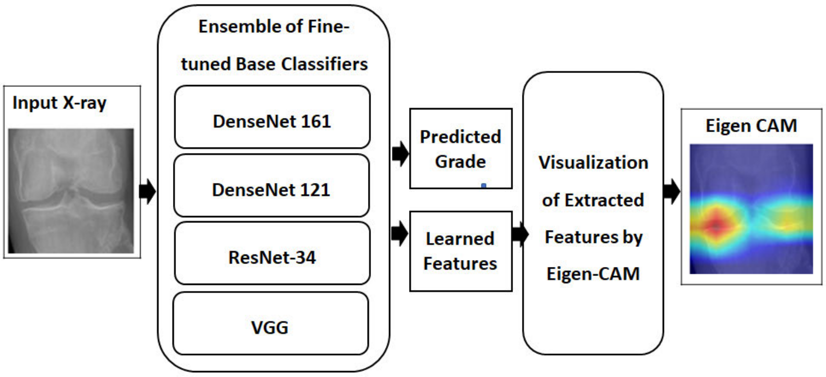

Input X-ray
Ensemble of Fine-tuned Base Classifiers
DenseNet 161
DenseNet 121
ResNet-34
VGG
Predicted Grade
Learned Features
Visualization of Extracted Features by Eigen-CAM
Eigen CAM

FIGURE 1. Methodology.

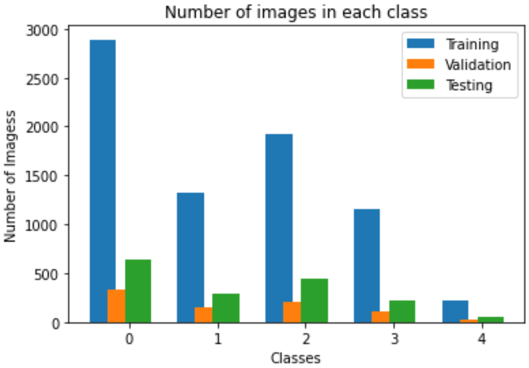


FIGURE 2. Number of images.

section III-B narrates the pre-trained deep learning networks used in our experiments. Finally, in section III-C, experimental settings and the training process are explained. Figure 1 depicts the overall methodology.

A. DATASET DESCRIPTION

The dataset used for this study is based on the Osteoarthritis Initiative dataset [9]. There are 9786 X-ray images graded according to the KL grading scheme. In grade 0, there are 3857 images, 1770 in grade 1, 2578 in grade 2, 1286 in grade 3, and 295 in grade 4. The size of each image is 224 × 224.

Data is highly unbalanced; hence, data has been split into the train, test, and validation classes considering the number of available samples for each category. Figure 2 reflects data distribution between training, testing, and validation. The same partitioning is used by [19] and [28]

B. NETWORKS

1) VISUAL GEOMETRIC GROUP NET (VGG)

VGG [11], winner of ImageNet Large Scale Visual Recognition Challenge (ILSVRSC) 2014, for localization and classification task, is based on multiple minute convolution filters and max pooling layer. On the other hand, VGG-19 is significantly deeper, with 144 million parameters, and is a variant of the VGG model that comprises 16 convolutional layers, three fully connected layers, five max pool layers, and one softmax layer. Max pool layers are crucial for spatial subsampling and generic feature extraction.

2) RESIDUAL NETWORK (ResNet)

ResNet, [10] that is comparatively shallow, has proven to perform better for image recognition tasks and won ILSVRC 2015. The ResNet-34 involves over 21 million trainable parameters and overcomes the problem of vanishing or exploding gradients by adding auxiliary connections. These connections help maintain a constant flow of information throughout the network and reduce computational costs.

3) DENSELY CONNECTED CONVOLUTIONAL NEURAL NETWORK (DenseNet)

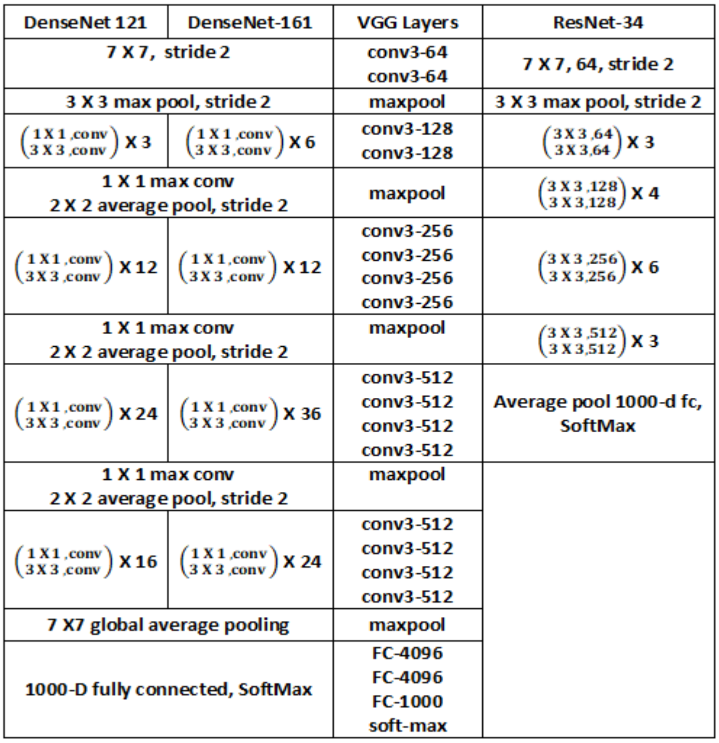
DenseNet [12] that connects each layer to each layer in a feed-forward way makes it dense, promotes feature reuse, reduces the number of parameters, and improves learning. DenseNets are based on auxiliary connections of ResNets and impose long-chained additional links to form dense blocks. The basic architectures of these four networks are compared in Figure 3.

The ensemble model joins predictions of divergent independently trained machine learning models, also called base classifiers, to reduce generalization error and improve performance. One of the different forms of the ensemble is the stacking ensemble, where outputs of base classifiers are combined and passed to another model for final prediction. We have separately fine-tuned VGG-19, ResNet-34, DenseNet 121, and DenseNet 161, inspired by the best models in the previous studies by Chen et al. [19], Tiulpin et al. [29], Mikhaylichenko and Demyanenko [28], Yong et al. [14] respectively. Finally, we developed an ensemble of these base models to improve the prediction precision and accuracy.

C. EXPERIMENTS

1) BASE CLASSIFIERS

For the training of base classifiers, the batch size is 28. Different image transformations are applied and chained together



DenseNet 121	DenseNet-161	VGG Layers	ResNet-34
7 X 7, stride 2		conv3-64 conv3-64	7 X 7, 64, stride 2
3 X 3 max pool, stride 2		maxpool	3 X 3 max pool, stride 2
$\begin{pmatrix}1\,X\,1\text{ ,conv}\\3\,X\,3\text{ ,conv}\end{pmatrix}$ X 3	$\begin{pmatrix}1\,X\,1\text{ ,conv}\\3\,X\,3\text{ ,conv}\end{pmatrix}$ X 6	conv3-128 conv3-128	$\begin{pmatrix}3\,X\,3\text{ ,64}\\3\,X\,3\text{,64}\end{pmatrix}$ X 3
1 X 1 max conv 2 X 2 average pool, stride 2		maxpool	$\begin{pmatrix}3\,X\,3\text{ ,128}\\3\,X\,3\text{,128}\end{pmatrix}$ X 4
$\begin{pmatrix}1\,X\,1\text{ ,conv}\\3\,X\,3\text{ ,conv}\end{pmatrix}$ X 12	$\begin{pmatrix}1\,X\,1\text{ ,conv}\\3\,X\,3\text{ ,conv}\end{pmatrix}$ X 12	conv3-256 conv3-256 conv3-256 conv3-256	$\begin{pmatrix}3\,X\,3\text{ ,256}\\3\,X\,3\text{,256}\end{pmatrix}$ X 6
1 X 1 max conv 2 X 2 average pool, stride 2		maxpool	$\begin{pmatrix}3\,X\,3\text{ ,512}\\3\,X\,3\text{,512}\end{pmatrix}$ X 3
$\begin{pmatrix}1\,X\,1\text{ ,conv}\\3\,X\,3\text{ ,conv}\end{pmatrix}$ X 24	$\begin{pmatrix}1\,X\,1\text{ ,conv}\\3\,X\,3\text{ ,conv}\end{pmatrix}$ X 36	conv3-512 conv3-512 conv3-512 conv3-512	Average pool 1000-d fc, SoftMax
1 X 1 max conv 2 X 2 average pool, stride 2		maxpool	
$\begin{pmatrix}1\,X\,1\text{ ,conv}\\3\,X\,3\text{ ,conv}\end{pmatrix}$ X 16	$\begin{pmatrix}1\,X\,1\text{ ,conv}\\3\,X\,3\text{ ,conv}\end{pmatrix}$ X 24	conv3-512 conv3-512 conv3-512 conv3-512	
7 X7 global average pooling		maxpool	
1000-D fully connected, SoftMax		FC-4096 FC-4096 FC-1000 soft-max	

FIGURE 3. **Basic architectures of imageNet pre-trained CNNs are compared. all four networks contain multiple convolutional and max pool layers and, finally, a fully connected layer to produce 1000 outputs.**

to improve model learning. These transformations include changing the brightness and saturation, flipping the image horizontally, random affine, and normalizing. This batch was then passed to the base network.

Considering the KL grade classification as an ordinal regression problem, a rank-consistent ordinal regression-based framework (CORN) [30] has been used for loss calculation and grade prediction. It used the chain rule of conditional probability distribution to obtain unconditional rank probabilities and was developed to be used with deep neural networks. The implementation of CORN is provided by the coral-PyTorch library. The idea is to solve the KL grading problem as an ordinal classification task.

Given a training set:

$$D = \{x^{[i]}, y^{[i]})\}_{i=1}^{N} \tag{1}$$

These target labels are extended into binary tasks to indicate if $y^{[i]}$ exceeds a certain grade g_k, such that $y_k^{[i]} \in 0, 1$. These are then passed to the base classifier model. We set the output layer for CORN to use g-1 classes associated with kl grade 1, grade 2, grade 3, and grade 4, i.e., g1, g2, g3, and g4, respectively, in the output layer of the base classifier model. Then CORN calculates conditional probability based on conditional training subsets. So, the output of the k-th binary task is

$$f_k(x^{[i]}) = P(y^{[i]} > g_k | y^{[i]} > g_{k-1}), \tag{2}$$

where these events are nested $\{y^{[i]} > g_k\} \subseteq \{y^{[i]} > g_{k-1}\}$

Four logits from the base classifier can be calculated using the following expression,

$$q[i] = 1 + \sum_{j=1}^{k-1} 1\ (p(y^{[i]} > g_j) > 0.5 \tag{3}$$

For an input image i, the index of KL predicted grade is, $g_{q[i]}$

For KL grade 0, the results of all sub-tasks are false, resulting in prediction 0. While for KL grade 4, the results of all four sub-tasks are true, which makes its sum equal to 4. Hence grade 4 is predicted.

For training CORN, the following loss function is minimized.

$$L(Z, y) = -\frac{1}{\sum_{j=1}^{k-1}|S_j|}\sum_{j=1}^{k-1}\sum_{i=1}^{|S_j|}\left[log\left(\sigma\left(z^{[i]}\right)\right).\right.$$
$$\left.\times\ 1y^{[i]} > g_j + log\left(\sigma\left(z^{[i]}\right) - z^{[i]}\right).1y^{[i]} \le g_k\right] \tag{4}$$

where $|S_j|$ is the size of the j-th conditional training set.

Z is the last layer's net inputs, and we call these logits.

All base classifiers calculate logits, predict labels, and calculate the loss. The optimizer is Adaptive Moment Estimation (ADAM) with an initial learning rate of 0.0001. The learning rate was reduced after every five epochs. The training was done for 100 epochs. After each training epoch, the model is validated on the validation set. Model results with the

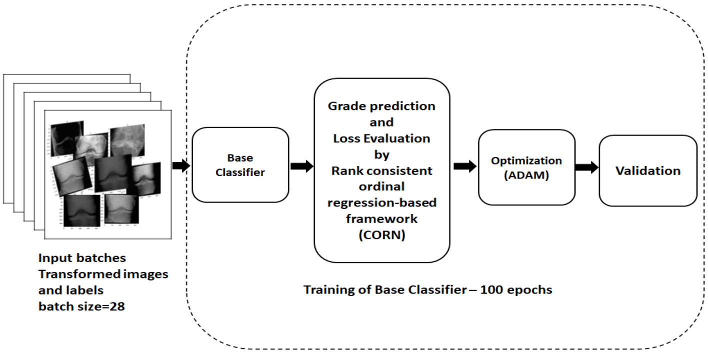

Input batches
Transformed images
and labels
batch size=28
Base Classifier
Grade prediction and Loss Evaluation by Rank consistent ordinal regression-based framework (CORN)
Optimization (ADAM)
Validation
Training of Base Classifier – 100 epochs

FIGURE 4. Training base classifiers: The batch size is 28. Input images are transformed using different transformations. Training is done for 100 Epochs. Rank Consistent Ordinal Regression based framework (CORN) is applied to calculate loss and predict grades. ADAM is used for the optimization of the network. Validation data is validated at the end of each epoch.

best performances are reported. The Eigen-CAM visualizes salient features [31], [32]. All four selected networks are fine-tuned according to the training process mentioned above, also reflected in Figure 4.

2) ENSEMBLE

The outputs from the base classifiers are joined using a fully connected layer in the Ensemble model, which uses the ADAM optimizer with an initial learning rate of 0.0001. The batch size was 28. The learning rate decayed every three epochs. After each epoch validation set was evaluated. Training is done for 25 epochs, and CrossEntropy is used to calculate the loss. The Ensembling process is demonstrated in Figure 5.

Implementation has been done with python 3.7, PyTorch-v1.12.1, and coral-PyTorch-1.4.0 in the GoogleColab framework.

D. EVALUATION METRICS

A machine learning model will partition the predictions into different classes for evaluation.

True Positive (TP) means an image is correctly identified as positive.

False Positive (FP) means an image is negative but identified as positive.

True Negative (TN) means an image is negative and correctly identified as negative.

False Negative (FN) means an image is positive but identified as negative.

Based on these outcomes, accuracy can be calculated.

Accuracy is the measure of correct predictions. In terms of FP, TP, TN, and FN, accuracy is

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{5}$$

Precision is how many images predicted as positive are positive.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

The recall is the ratio of images correctly identified as positive.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

The F1 score or F Measure is used to measure the accuracy of a model. The greater the F1 score, the better the performance of our model.

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

Another evaluation measure for Ordinal Regression is Mean Squared Error (MSE) is the mean or average of the square of the difference between actual and estimated values. For a dataset of n images, given the actual label and predicted label, the MSE is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(actual - predicted)^2 \tag{9}$$

Cohen's Quadratic Weighted Kappa (QWK) (κ) is useful when classification labels are ordered. It measures the agreement between classification accuracy (Proportion of observed agreement i.e., P_0) and the theoretical probability of chance agreement i.e.,P_e. Weights (w) can be assigned according to the ordering or severity information.

$$\kappa = \frac{P_0(w) - P_e(w)}{1 - P_e(w)} \tag{10}$$

The Receiver Operating Characteristics Curve (ROC) measures how accurately the model can differentiate between different classes, while Area Under the Curve (AUC) measures the entire 2-dimensional area underneath the ROC curve.

IV. RESULTS

Accuracy, precision, recall, F1-score, and AUC for all the models are compared in Table 2. The Ensemble model has secured the best overall results for almost all evaluation metrics except recall.

The Ensemble model achieved an overall accuracy of 0.98, an overall precision of 0.98, an overall F1-score of 0.97,

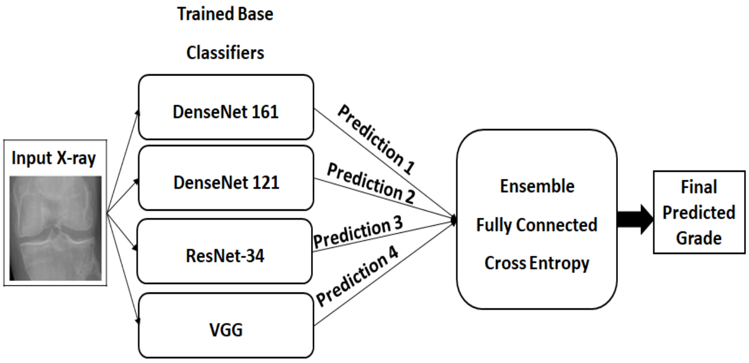


FIGURE 5. Ensembling: four imageNet pre-trained networks are fine-tuned, and then their predictions are joined as an ensemble to produce one final output.

TABLE 2. Performance comparison of five models overall and across different grades.

Metric	Model	Overall	Grade				
			0	1	2	3	4
Accuracy	DenseNet-121	0.96	0.99	0.88	0.96	0.97	0.96
	DenseNet-161	0.97	1.00	0.95	0.96	0.99	0.94
	ResNet-34	0.95	0.96	0.93	0.94	0.97	0.96
	Vgg-19	0.96	0.99	0.92	0.95	0.95	0.94
	Ensemble	0.98	1.00	0.94	0.97	0.99	0.92
Precision	DenseNet-121	0.97	0.96	0.95	0.94	0.98	1.00
	DenseNet-161	0.97	0.97	0.97	0.99	0.97	0.98
	ResNet-34	0.96	0.97	0.88	0.97	0.96	1.00
	Vgg-19	0.96	0.97	0.92	0.96	0.97	0.96
	Ensemble	0.98	0.98	0.97	0.98	0.96	1.00
Recall	DenseNet-121	0.95	0.99	0.88	0.96	0.97	0.96
	DenseNet-161	0.97	1.00	0.95	0.96	0.99	0.94
	ResNet-34	0.95	0.96	0.93	0.94	0.97	0.96
	Vgg-19	0.95	0.99	0.92	0.95	0.95	0.94
	Ensemble	0.96	1.00	0.94	0.97	0.98	0.92
F1-Score	DenseNet-121	0.96	0.98	0.91	0.95	0.97	0.98
	DenseNet-161	0.97	0.98	0.96	0.97	0.98	0.96
	ResNet-34	0.95	0.97	0.90	0.95	0.97	0.98
	Vgg-19	0.95	0.98	0.92	0.95	0.96	0.95
	Ensemble	0.97	0.99	0.96	0.97	0.97	0.96
AUC	DenseNet-121	0.97	0.98	0.94	0.97	0.98	0.98
	DenseNet-161	0.98	0.99	0.97	0.98	0.99	0.97
	ResNet-34	0.97	0.97	0.95	0.96	0.98	0.98
	Vgg-19	0.97	0.98	0.95	0.97	0.97	0.97
	Ensemble	0.98	0.99	0.97	0.98	0.99	0.96

and an overall AUC of 0.98. The highest QWK (κ), with a confidence interval of 0.95, is 0.99; minimum MAE and MSE are 0.027 and 0.032, respectively, which the Ensemble model also obtains. The confusion matrices of these models

TABLE 3. Cohen kappa, MAE, and MSE for all models.

Metric	Model				
	DenseNet-121	DenseNet-161	ResNet-34	VGG-19	Ensemble
Weighted Kappa	0.98	0.98	0.98	0.98	0.99
MAE	0.0471	0.0326	0.054	0.0465	0.0265
MSE	0.058	0.0459	0.065	0.055	0.0326

can observe the same results. For grade 0, 637 classes out of 639 were correctly predicted by the Ensemble model. ROC curves and AUC values also depict the same exciting fact that DenseNet-161 and Ensemble model performed best for grades 0, grade 2, and grade 3.

It can also be seen that the overall accuracy of VGG-19 and DenseNet-121 is equal. For overall accuracy, DenseNet-161 scored first, while ResNet-34 was at the lowest by acquiring a minimum accuracy of 0.95. The DenseNet-161 achieved better accuracies for most KL grades, i.e., grade 0, grade 1, and grade 3.

DenseNet-121 and ResNet-34 have obtained the best accuracy, precision, recall, and F1 score for grade 4. For the ResNet-34 confusion matrix, out of 639 x-rays for grade 0, 615 were correctly reported as grade 0, 21 were misclassified as grade 1, and only three were mislabeled as grade 2. For grade1, 14,6, and 1, X-rays were wrongly identified as grade 0, grade 2, and grade 3, respectively. Out of 447 x-rays, images of grade2 5, 18, and 5 were wrongly reported as grade 0, grade 1, and grade 3. For grade 3, only six images were improperly classified as grade 2. For grade 4, only two images needed clarification and were marked as grade 3.

It has been observed that the model wrongly labeled adjacent grades more often than distant ones. Overall final accuracy achieved as 95%. Accuracy, precision, recall, and F-Score results are shown in Table 2 and Table 3. Confusion Matrices and ROC curves can be seen in Table 4.

The Ensemble and DenseNet-161 obtained the best precision, recall, and F-scores. To abridge the research gaps discussed earlier in section I, we used the dataset set based on the OAI dataset. In addition, since the OAI study is conducted within defined protocols and images are annotated by multiple annotators' consensus, there are fewer chances of inter-rater disagreements. To improve the overall performance of the model following strategies were implemented. First, the input image size enables the model to extract useful information about the structures and features of bones. Secondly, the ratio of samples between each class is not equal. This problem of data unbalances resolved by applying different transformations to the training data. These variations helped the model learn variance, eliminating the need to treat left and right knee images separately, resulting in a robust model. It was also observed that training for more epochs could enhance the overall accuracy by compromising the accuracy of any KL grade. Since the number of samples in grade 0 is more abundant than in the rest of the grades, accuracy increases for grade 0. Hence more training leads to better accuracy of rates with more samples and starts overfitting for grades with fewer samples, thus reducing their accuracy.

The benefits of using CORN can also be observed. By considering the ordering information of the KL grades, the overall performances of the models are improved. Most of the wrong predictions are made to the nearby grade instead of making a wrong prediction to the distant grade, and most predictions now fall towards the diagonal. During training, the learning rate decays, which helps the model converge.

Accuracies, precision, recall, and F-score are reported. In addition, overall quadratic weighted kappa is also reported.

For computer-aided diagnosis, the focus is on the model's accuracy and precision hence we used the above-mentioned metrics for the model's evaluation. The computational complexity of deep learning models depends upon the number of parameters and layers, as shown in Figure 3. The number of Floating-Point Operations (FLOPs) is another metric that measures the performance of deep networks. ResNet-34 has 4 Billion FLOPs with 21 Million parameters. VGG-19, DenseNet-121, and DenseNet-161 have 20 Billion, 3 Billion, and 8 Billion FLOPs, respectively. Finally, the Ensemble model utilizes all these operations to predict better.

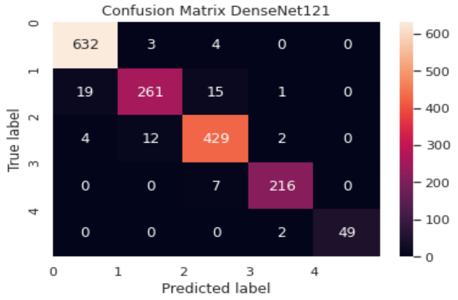
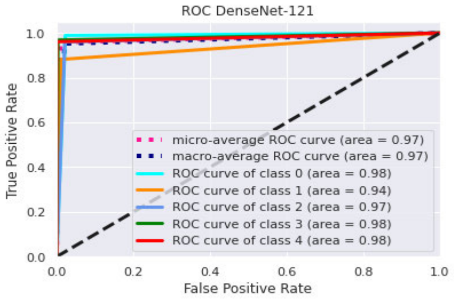
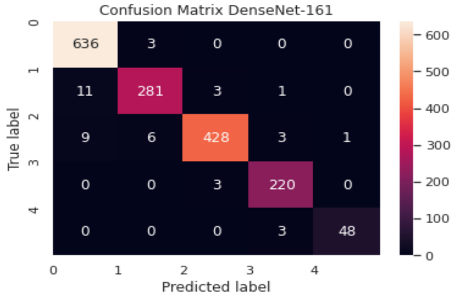
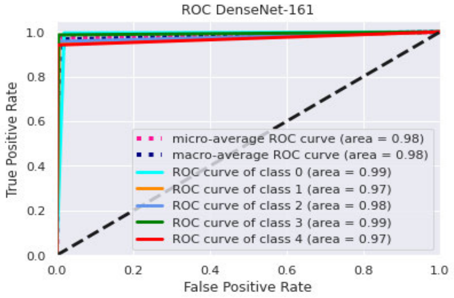
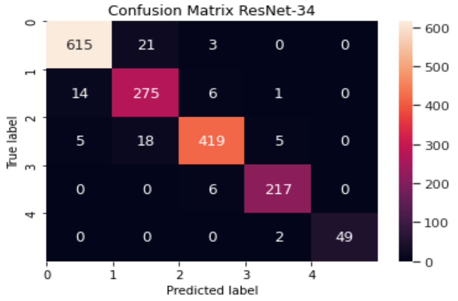
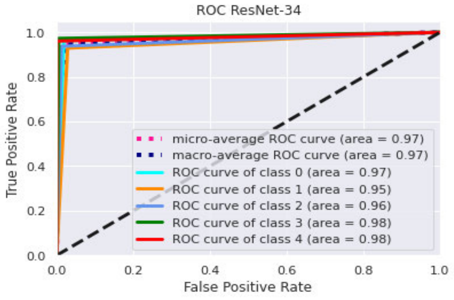
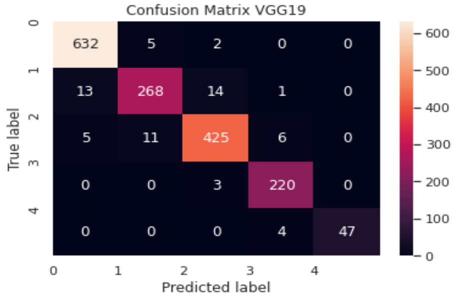
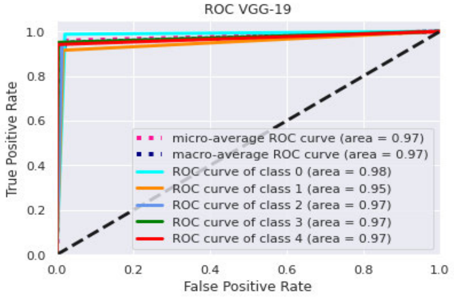
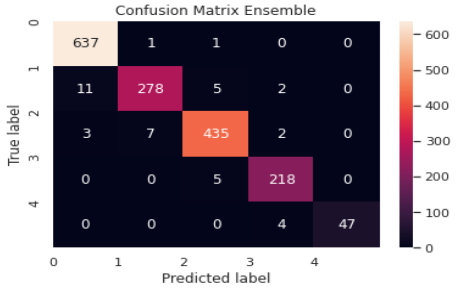
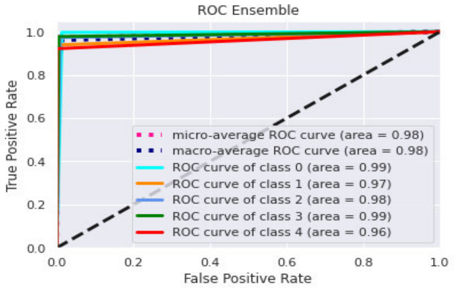
Table 5 presents the feature localization by Eigen-CAM. It can be concluded that our models could extract valuable features from the X-ray image. These models can be visualized to differentiate bone sclerosis, osteophytes, cartilage degeneration, and joint space narrowing, as the Eigen-CAM highlights their extracted features. For instance, for grade 0, the Ensemble model has almost 100% accuracy. Similarly, for grade 4, ResNet-34 and DenseNet-121 identified better features, as shown by Eigen-CAM.

Table 6 compares our results with other state-of-the-art techniques.

In almost all previous studies overall accuracy of the KL grade 1 was significantly less than the overall accuracy. Compared to our research, some studies have used a different dataset, and some have adopted another loss function.

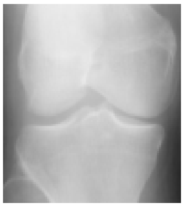
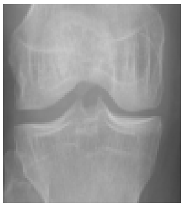
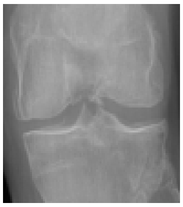
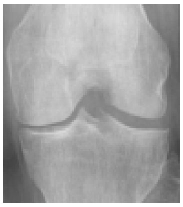
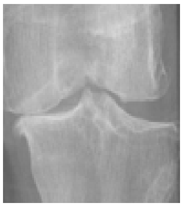
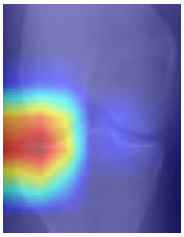
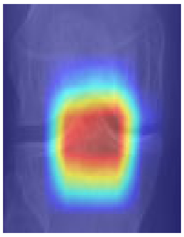
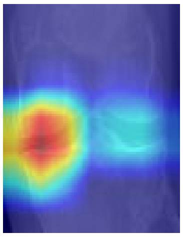
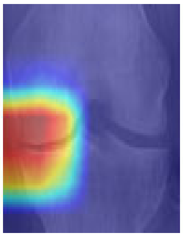
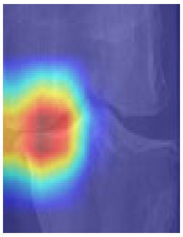
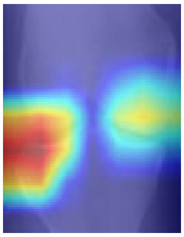
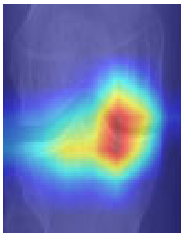
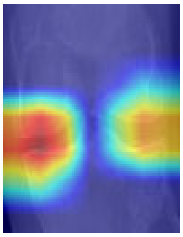
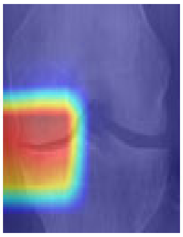
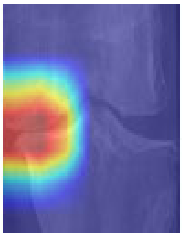
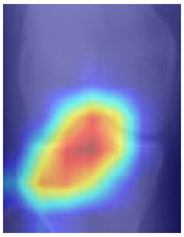
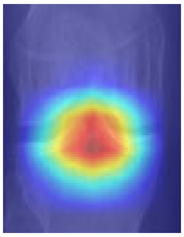
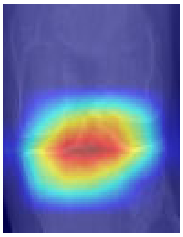
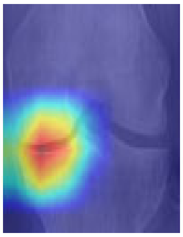
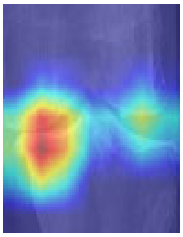
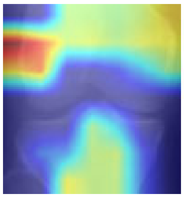
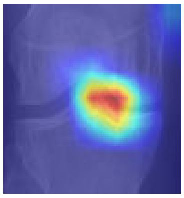
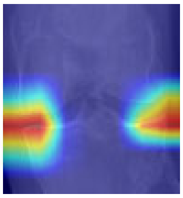
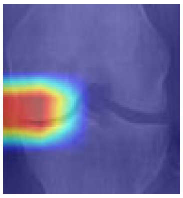
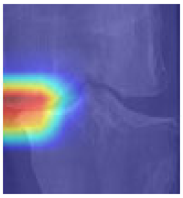
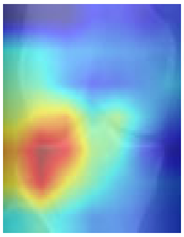
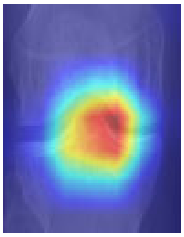
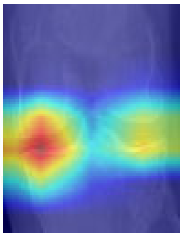
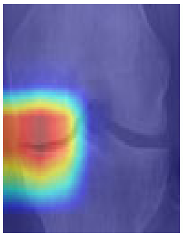
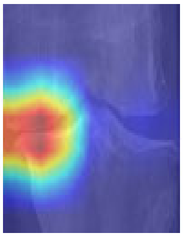
The most recent study by Liu et al. [35] has performed multiple experiments with Cross Entropy loss and Focal Ordinal Loss (FOL) using several classical pre-trained models such as VGG, ResNet, DenseNet, and GoogleNet. Moreover, they have performed experiments with augmented datasets and CBAM. Their best results are reported in Table 6. The overall accuracy is 66%. In addition, they have shown improvements in performance measures using FOL, but their results were unstable in accuracy.

TABLE 4. Results- confusion matrices and ROC curves.

Model	Confusion Matrix	ROC-AUC
DenseNet-121		
DenseNet-161		
ResNet-34		
VGG-19		
Ensemble		

If we refer to the work of Yong et al. [14], they have performed several experiments with state-of-the-art architectures. Overall accuracy was enhanced up to 88.09% with a QWK of 0.86 using an ordinal regression module with DenseNet-161. Their accuracies for initial grades can be improved.

TABLE 5. Eigen-CAM visualization all models.

	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4
Original					
DenseNet-121					
DenseNet-161					
ResNet-34					
VGG-19					
Ensemble					

Feng et al. [33] have implemented ResNet, without its residual part, with an attention module, where Mish is used as an activation function. However, with the same dataset and dataset distribution in a train, testing, and validation as ours, their overall performance and performance for each KL grade still need to be higher than our models.

TABLE 6. Comparison with other state-of-the-art techniques.

Reference	Overall Evaluation	Accuracy per KL Grade (%)

Antony et al. (2017) [17]	Accuracy 63.6%
	MSE= 0.70
		KL0=87 KL1=6 KL2=60
		KL3=72 KL4=78

Tiulpin et al. (2018) [29]	Accuracy=66.71%
	MSE=0.48
	Kappa coefficient = 0.83
	AUC=0.93
		KL0=78 KL1=45 KL2=52
		KL3=70 KL4=88

Chen et al. (2019) [19]	Accuracy=69.7%
	MAE=0.344
		KL0=87 KL1=18 KL2=75
		KL3=75 KL4=84

Mikhaylichenko et al. (2020) [28]	Accuracy = 71.08%	KL0=92 KL1=16 KL2=72
		KL3=83 KL4=63

Feng et al. (2021) [33]	Accuracy = 70.23%
	Recall = 68.23%
	Precision = 70.25%
	F1 = 67.55%
		KL0=92 KL1=15 KL2=70
		KL3=82 KL4= 84

Yong et al. (2021) [14]	Accuracy= 88.09%
	MAE=0.33
	QWK= 0.8609
		KL0=80 KL1=39 KL2=70
		KL3=80 KL4=86

Liu et al. (2022) [34]	Accuracy = 66%
	MSE = 0.48
		KL0=82 KL1=29 KL2=57
		KL3=85 KL4= 82

This work	Accuracy=98%
	MAE=0.027
	MSE=0.033
	QWK = 0.99
	AUC = 0.98
		KL0=100 KL1=94 KL2=97
		KL3=99 KL4=92

In comparison to the work by Mikhaylichenko and Demyanenko [28], which used variants of DenseNet, the accuracies of our, DenseNet-121 and DenseNet-161 are better. Our lowest-performing model, ResNet-34, performed better than their best ensemble model. They performed multiple experiments with DenseNet, training from scratch and with pre-trained models. They have also compared the results of CrossEntropy loss, and ordinal loss [19]. The main difference is that we used CORN to calculate losses and grade prediction.

Chen et al. [19] have applied knee joint localization. Then they made a comparison of fine-tuning variants of ResNet, VGG, DenseNet as well as Inception. They have also introduced an adjustable ordinal loss function considering KOA grading an ordinal regression problem. Our models outperform their results by using CORN loss.

Tiulpin et al. [29] developed and compared three types of models. One is the pre-trained ResNet-34 as a baseline network. Second is the re-implementation of CNN by [17]. Finally, their models are based on the siamese network trained using multiple hyper-parameters and different seeds. They use cropped images into two square patches and feed these images to the Siamese network, which makes their approach different from our research. Finally, they used ensembling to combine networks trained with different seeds. They have used separate datasets for training, testing, and validation. They have also included radiographs taken from multiple angles, i.e., 5^o, 10^o, and 15^o. To train our model, we applied different conversions, e.g., rotation and flipping, to the frontal images for better training.

Another impactful and inspiring study that is the basis of most of the studies mentioned above is by Antony et al. [17]. Unlike our work, they have performed experiments with two types of datasets. They have implemented and trained a CNN from scratch and another CNN for simultaneous regression and classification for KOA grading. Our models have shown a tremendous difference in performance.

A substantial limitation of our study is the need for more diversity in the dataset. For example, the same dataset is used for training, testing, and validation.

V. CONCLUSION

In this paper, we have applied a deep learning-based ordinal classification approach to grading knee osteoarthritis X-rays. We present new state-of-the-art results in automated KOA classification for all KL grades. In addition, we enhanced the performance of our models by making an ensemble of fine-tuned models. Our method provides a quick, early, and reliable evaluation of input knee X-rays, and medical practitioners can use it as an alternative option to save time. Ordinal classification improved the performance of our system significantly. Further Ensemble has also shown significant improvement for all evaluation metrics. In the future, we plan to incorporate multiple datasets from multiple settings.

REFERENCES

[1] Y. Badshah, M. Shabbir, H. Hayat, Z. Fatima, A. Burki, S. Khan, and S. U. Rehman, "Genetic markers of osteoarthritis: Early diagnosis in susceptible Pakistani population," *J. Orthopaedic Surgery Res.*, vol. 16, no. 1, pp. 1–8, Dec. 2021.

[2] S. K. Das and A. Farooqi, "Osteoarthritis," *Best Pract. Res. Clin. Rheumatol.*, vol. 22, no. 4, pp. 657–675, 2008.

[3] S. S. Gornale, P. U. Patravali, and R. R. Manza, "Detection of osteoarthritis using knee X-ray image analyses: A machine vision based approach," *Int. J. Comput. Appl.*, vol. 145, no. 1, pp. 20–26, 2016.

[4] M. N. Iqbal, F. R. Haidri, B. Motiani, and A. Mannan, "Frequency of factors associated with knee osteoarthritis," *J. Pakistan Med. Assoc.*, vol. 61, no. 8, p. 786, 2011.

[5] Y. X. Teoh, K. W. Lai, J. Usman, S. L. Goh, H. Mohafez, K. Hasikin, P. Qian, Y. Jiang, Y. Zhang, and S. Dhanalakshmi, "Discovering knee osteoarthritis imaging features for diagnosis and prognosis: Review of manual imaging grading and machine learning approaches," *J. Healthcare Eng.*, vol. 2022, pp. 1–19, Feb. 2022.

[6] J. H. Kellgren and J. S. Lawrence, "Radiological assessment of osteoarthrosis," *Ann. Rheumatic Diseases*, vol. 16, no. 4, pp. 494–502, Dec. 1957.

[7] D. Saini, T. Chand, D. K. Chouhan, and M. Prakash, "A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on X-ray images," *Biocybernetics Biomed. Eng.*, vol. 41, no. 2, pp. 419–444, Apr. 2021.

[8] L. Anifah, I. K. E. Purnama, M. Hariadi, and M. H. Purnomo, "Osteoarthritis classification using self organizing map based on Gabor kernel and contrast-limited adaptive histogram equalization," *Open Biomed. Eng. J.*, vol. 7, no. 1, pp. 18–28, Feb. 2013.

[9] A. Brahim, R. Jennane, R. Riad, T. Janvier, L. Khedher, H. Toumi, and E. Lespessailles, "A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis initiative," *Computerized Med. Imag. Graph.*, vol. 73, pp. 11–18, Apr. 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[13] E. Chaves, C. B. Gonçalves, M. K. Albertini, S. Lee, G. Jeon, and H. C. Fernandes, "Evaluation of transfer learning of pre-trained CNNs applied to breast cancer detection on infrared images," *Appl. Opt.*, vol. 59, no. 17, p. 23, 2020.

[14] C. W. Yong, K. Teo, B. P. Murphy, Y. C. Hum, Y. K. Tee, K. Xia, and K. W. Lai, "Knee osteoarthritis severity classification with ordinal regression module," *Multimedia Tools Appl.*, vol. 81, pp. 41497–41509, Jan. 2021.

[15] C. Zhang, C. Zhu, J. Xiao, X. Xu, and Y. Liu, "Image ordinal classification and understanding: Grid dropout with masking label," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[16] C. Beckham, "Techniques in ordinal classification and image-to-image translation," Ph.D. dissertation, Mémoire de maîtrise, Département de génie informatique et génie logiciel, Ecole Polytechnique, Montreal, QC, Canada, 2017. [Online]. Available: https://books.google.com.pk/books?id=GmXAwgEACAAJ

[17] J. Antony, K. McGuinness, K. Moran, and N. E. O'Connor, "Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. Data Mining Pattern Recognit.* Cham, Switzerland: Springer, 2017, pp. 376–390.

[18] 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.

[19] 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.

[20] B. Zhang, J. Tan, K. Cho, G. Chang, and C. M. Deniz, "Attention-based CNN for KL grade classification: Data from the osteoarthritis initiative," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 731–735.

[21] A. Tiulpin and S. Saarakkala, "Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks," *Diagnostics*, vol. 10, no. 11, p. 932, Nov. 2020.

[22] T. Tariq, Z. Suhail, and Z. Nawaz, "Machine learning approaches for the classification of knee osteoarthritis," in *Proc. 3rd Int. Conf. Electr., Comput., Commun. Mechatronics Eng. (ICECCME)*, 2023.

[23] B. Norman, V. Pedoia, A. Noworolski, T. M. Link, and S. Majumdar, "Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs," *J. Digit. Imag.*, vol. 32, no. 3, pp. 471–477, Jun. 2019.

[24] 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.

[25] (2006). *Oai*. [Online]. Available: https://nda.nih.gov/oai

[26] P. Chen. (2018) *Knee Osteoarthritis Severity Grading Dataset*. [Online]. Available: https://data.mendeley.com/datasets/56rmx5bjcr/1/

[27] K. A. Thomas, L. Kidzinski, E. Halilaj, S. L. Fleming, G. R. Venkataraman, E. H. G. 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.

[28] A. Mikhaylichenko and Y. Demyanenko, "Automatic grading of knee osteoarthritis from plain radiographs using densely connected convolutional networks," in *Proc. Int. Conf. Anal. Images, Social Netw. Texts*. Cham, Switzerland: Springer, 2020, pp. 149–161.

[29] 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, pp. 1–10, Jan. 2018.

[30] X. Shi, W. Cao, and S. Raschka, "Deep neural networks for rank-consistent ordinal regression based on conditional probabilities," 2021, *arXiv:2111.08851*.

[31] 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.

[32] Jacob Gildenblat and Contributors. (2021). *PyTorch Library for CAM Methods*. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

[33] Y. Feng, J. Liu, H. Zhang, and D. Qiu, "Automated grading of knee osteoarthritis X-ray images based on attention mechanism," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1927–1932.

[34] W. Liu, T. Ge, L. Luo, H. Peng, X. Xu, Y. Chen, and Z. Zhuang, "A novel focal ordinal loss for assessment of knee osteoarthritis severity," *Neural Process. Lett.*, vol. 54, no. 6, pp. 5199–5224, 2022.

[35] B. Liu, J. Luo, and H. Huang, "Toward automatic quantification of knee osteoarthritis severity using improved faster R-CNN," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 15, no. 3, pp. 457–466, Mar. 2020.

TAYYABA TARIQ is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science, University of the Punjab. She is an Assistant Professor with the Department of Computer Science, University of the Punjab. Her research interests include image processing and deep learning.

ZOBIA SUHAIL received the Ph.D. degree in computer science from Aberystwyth University, in 2019. She is currently an Assistant Professor with the Department of Computer Science, University of the Punjab. Her research interests include medical image processing and machine learning.

ZUBAIR NAWAZ received the Ph.D. degree in computer engineering from the Delft University of Technology, in 2011. He is currently an Assistant Professor with the Department of Data Science, University of the Punjab. His research interests include data science, machine learning, high-performance computing, compiler optimization, and scientific computing.