



Knee osteoarthritis severity prediction using an attentive multi-scale deep convolutional neural network

Rohit Kumar Jain¹ · Prasen Kumar Sharma¹ · Sibaji Gaj² · Arijit Sur¹ · Palash Ghosh^{3,4,5}

Received: 14 August 2021 / Revised: 5 February 2022 / Accepted: 18 April 2023 /

Published online: 20 June 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Knee Osteoarthritis (OA) is a destructive joint disease identified by joint stiffness, pain, and functional disability concerning millions of lives across the globe. It is generally assessed by evaluating physical symptoms, medical history, and other joint screening tests like radiographs, Magnetic Resonance Imaging (MRI), and Computed Tomography (CT) scans. Unfortunately, the conventional methods are very subjective, which forms a barrier in detecting the disease progression at an early stage. This paper presents a deep learning-based framework, namely OsteoHRNet, that automatically assesses the Knee OA severity in terms of Kellgren and Lawrence (KL) grade classification from X-rays. As a primary novelty, the proposed approach is built upon one of the most recent deep models, called the High-Resolution Network (HRNet), to capture the multi-scale features of knee X-rays. In addition, an attention mechanism has been incorporated to filter out the counterproductive features and boost the performance further. Our proposed model has achieved the best multi-class accuracy of 71.74% and MAE of 0.311 on the baseline cohort of the OAI dataset, which is a remarkable gain over the existing best-published works. Additionally, Gradient-based Class Activation Maps (Grad-CAMs) have been employed to justify the proposed network learning.

Keywords Classification · Deep learning · Hrnet · Kellgren lawrence grade · Knee osteoarthritis · Knee x-ray · Osteo hrnet

✉ Rohit Kumar Jain
jkrohit03@gmail.com

¹ Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Guwahati, India

² Cleveland Clinic, Cleveland, ON, USA

³ Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati, India

⁴ Jyoti and Bhupat Mehta School of Health Sciences and Technology, Indian Institute of Technology Guwahati, Guwahati, India

⁵ Centre for Quantitative Medicine, Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

1 Introduction

Knee OA is a common joint disorder caused by the eroding of the articular cartilage between the joints, which leaves the knee bones touching and rubbing against each other. In general, it occurs in the synovial joints and results from a combination of genetic factors, injury, and overuse [21, 25]. The pain, swelling, and stiffness in the joints are gradual in the onset and begin to worsen by the rigorous activity and stress compared to other inflammatory arthritis where activity and exercising improve symptoms. It may eventually lead to instability, joint deformity, and reduction in joint functionality [25]. The progression of knee OA is characterized by the flattening of space between the knee joints due to loss of cartilage [21]. The following key changes [8, 33], described by the word LOSS, determines the presence and progression of knee OA:

- L- “loss of joint space”, caused by the cartilage loss,
- O- “osteophytes formations”, projections that form along the margins of the joint,
- S- “subarticular sclerosis”, increase in bone density along the joint line, and
- S- “subchondral cysts”, caused due to holes in the bone filled with fluid along the joints [1].

Radiographic screening (X-Rays), MRI, and CT scans are the traditionally adopted ways to detect the knee joint’s structural changes and diagnose knee OA’s biological condition. However, the traditional treatment for knee OA are not effective enough to fix the disease completely, as there is no known cure for OA [22]. Therefore, it is of utmost importance to detect the deformation of the joint at such a stage before which it becomes impossible to reverse the loss [31]. In today’s time, OA treatment requires a personalized medicine approach [7]. Generally, the knee OA severity is measured in terms of the World Health Organization (WHO) approved KL grading scale [13]. KL grading is a 5-point semi-quantitative progressive ordinal scale ranging from grade 0 (*low severity*) to 4 (*high severity*). Figure 1 shows the disease progression along with its corresponding KL grade (Source: OAI) [3].

1.1 Challenges

A complete cure for this disease remains quite challenging to find, and OA management is mainly palliative [21, 33]. MRI screenings and CT scans are effective for the diagnosis as they highlight the three-dimensional structure of the knee joints [11, 24]. However, they have certain drawbacks, including limited availability, extreme device expenses, the time required in diagnosing, and the inclination to image ancient rarities [14, 36]. At the same time, X-Rays are the most effective and economically feasible way of diagnosing the

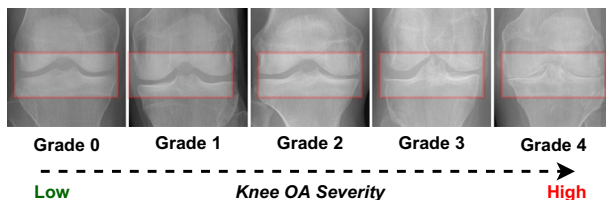


Fig. 1 Knee OA disease progression: A qualitative demonstration of sample X-rays and their corresponding KL grades. Source: OAI [3]

disease, given the routine knee OA diagnosis. However, the currently adopted methods for assessing the disease progression from X-Ray images are not effective enough. The diagnosis requires a very skilled practitioner to analyze the radiographic scans accurately and are thus absolutely subjective and time-consuming [6]. The analysis may differ based on their expertise and sometimes may be inaccurate and doubtful [6]. Further, multiple tests may be costly for some of the patients.

A better and in-depth understanding of knee OA may result in timely prevention and treatment. It is believed that early treatment and preventive measures are the most effective way of managing knee OA. Unfortunately, there is no reliable method which exists for detecting knee OA at a reversible stage [16]. Recently, the use of Machine Learning (ML) and Deep Convolutional Neural Networks (CNNs) for knee OA analysis have shown remarkable supremacy in detecting even the slightest differences in biological joint structural variations in the X-Rays [15].

Deep CNNs have been widely adopted in many medical imaging tasks, including classifications of COVID-19, pneumonia, tumor, bone fracture, polyps detection, etc. For *e.g.*, CheXNet [26], a 121-layers deep CNN, performed astonishingly better than the average performance of four specialists in assessing pneumonia using plain radiographs [44]. However, it is difficult to collect the medical images, as the collection and annotation of such data are challenged by the expert availability and the data privacy concerns [44].

1.2 The Osteoarthritis Initiative (OAI) dataset

OAI is a distributed, observational study of patients, which is publicly available¹. It facilitates the scientific and research community worldwide to work on knee OA progression and develop new treatments and techniques beneficial for detection and treatment. This work has utilized the data acquired from the OAI repository and made available by Chen et al. [3, 4]. The dataset comprises knee bilateral posterior-anterior fixed flexion radiographs of 4796 participants, including male and female subjects from the baseline cohort. Figure 1 shows sample X-ray images pertaining to each KL grade.

2 Related developments

Several schemes have been developed for the Knee OA severity prediction in the past few years. Shamir et al. [32] utilized a weighted nearest neighbors algorithm that incorporated the hand-crafted features like Gabor filters, Chebyshev statistics, multi-scale histograms, etc. Antony et al. [2] proposed to utilize the transfer learning of the existing pre-trained deep CNNs. Later, Antony et al. [20] customized a deep CNN from scratch and optimized the network using a weighted combination of the traditional cross-entropy and the mean squared error, which served as dual-objective learning. Tuilpin et al. [39] developed a method inspired from the deep Siamese network [5], for learning the similarity metric between the pair of radiographs. Gorriz et al. [9] developed an end-to-end attention-based network, bypassing the need to localize knee joint, to quantify the knee OA severity automatically. Chen et al. [4] proposed to utilize pre-trained VGG-19 [34] along with an adjustable ordinal loss for the proportionate penalty to the misclassification. Yong et al. [45] utilized the pre-trained DenseNet-161 [10], along with an ordinal regression module (ORM), in order

¹Dataset source: <https://nda.nih.gov/oai/>

to treat the ordinality of the KL grading. They further optimized the network using the cumulative link (CL) loss function.

2.1 Motivation

Deep CNNs are renowned for learning the highly correlated features in an image [41]. In addition, it is a widely known fact that the first few layers of a deep CNN contribute to the learning of low-level features in an image. Whereas the last few layers contribute to the learning of the high-level features, enabling the final classification by adaptively learning spatial hierarchies of features [19]. While the low-level features are the minute details of an image, including points, lines, edges, etc., the high-level features comprise several low-level features, which make up the more prominent and robust structures for classification. Furthermore, existing works [12, 17, 37, 38, 42] have demonstrated how fusion of these features yields superior classification performance.

However, in general, the knee X-Rays lacks depth and due to which it may be difficult for a deep CNN to learn an efficient classification particularly, in the case of knee OA, where one KL grade is not very distinctive from the other unless carefully inspected (see Fig. 1). A few of the most recent state-of-the-art methods [4, 45] have directly utilized the existing popular image classification models in a plug-and-play fashion without supervising the network engineering relevant to the given problem. It should be mentioned that while a majority of those methods were built for a generic image classification problem, a few of them were explicitly designed using architectural search, e.g., MobileNetV2 [28].

Moreover, for the knee OA severity classification, the presented best-performing deep CNNs were enormous in size, exceeding 500 MB [4], to be precise. As a result, such models may require substantially high computational resources, making it challenging to deploy in real-time environments. Therefore, it may be said that the direct usage of popular classification models may not be appropriate. Although some recent methods [9, 39, 45, 46], have started to design the models specific to knee OA given the amount of information present in the knee X-rays. However, they still lack in terms of accuracy and computational overhead. For e.g., Zhang et al. [46] utilized the Convolutional Block Attention Module, namely CBAM [43], after every residual layer in their proposed architecture, which may not be computationally pleasant. The attention module has performed undoubtedly well in many high-level vision tasks. However, one must not overlook its computational overhead considering the presence of fully connected layers.

The applicability of deep CNNs in medical imaging heavily depends on the amount of data available for efficient learning. As an alternative, many deep learning-based methods have utilized the data augmentation [18] techniques to further boost the performance, which has not been much considered in the existing works.

2.2 Our contributions

Based on the aforementioned drawbacks of the existing best-published works, our contributions are five-fold, as follows:

1. An efficient deep CNN has been proposed for the knee OA severity prediction in terms of KL grades using X-ray images. Unlike existing methods, the proposed scheme is not a direct plug-and-play of popular deep models. The proposed scheme has been built upon a high-resolution network; namely, HRNet [42], that takes the spatial scale of the X-Ray image into account for efficient classification.

2. This work also proposes to utilize the attention mechanism only once in the entire network to reduce the computational overhead and adaptive filtering of the counterproductive features just before classification.
3. Also, instead of relying on traditional entropy-based minimization, the ordinal loss [4] have been adopted to optimize the proposed scheme.
4. To further boost the performance of the proposed scheme, data augmentation techniques have been incorporated, which have not been much considered in any recent work so far. The images are flipped horizontally and vertically at random to add variability in the dataset.
5. Lastly, an extensive set of experiments and Grad-CAM [27] visualization have been presented to justify the importance of each module of the proposed framework.

The rest of the paper is organized as follows: Section 3 presents the proposed method and the adopted cost function. Section 4 briefly describes the incorporated dataset, training details, competing methods, and evaluation metrics. Section 5 presents the quantitative and qualitative comparison against the best-published works, along with the ablation study followed by a brief discussion on the learning of proposed scheme in terms of Grad-CAM visualization and finally, the paper is concluded in Section 6.

3 Proposed method

This section presents the details of the proposed model, followed by a brief description of the incorporated cost function. The proposed framework is built upon the HRNet and Convolution Block Attention Module (CBAM) in a serially cascaded manner. A descriptive representation of the proposed model is shown in Fig. 2.

3.1 High resolution network

High-Resolution Network (HRNet) [42] is a novel and revolutionary multi-resolution deep CNN, which tends to maintain high-resolution feature representations throughout the network. It starts as a stream of 2D convolutions and subsequently adds up the high-to-low resolution streams to form the following stages. It then merges the multi-resolution streams in parallel for information exchange [42] as shown in Fig. 2 (marked as *High-Resolution Network*). HRNet tends to generate reliable multi-resolution representations with strong spatial sensitivity. It has been achieved by utilizing parallel connections instead of serial (see Fig. 3(a)) and recurrent fusion of the intermediate representations from multi-resolution streams (see Fig. 3(b)), as shown in Fig. 3. As a result, it enables the network to learn more

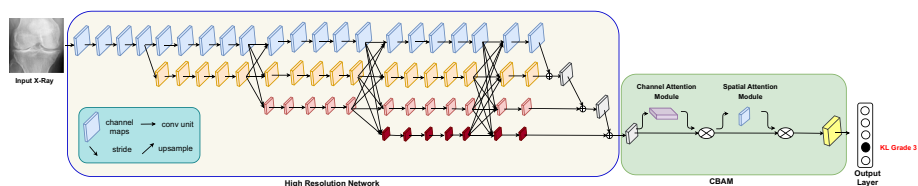


Fig. 2 An overview of the architecture of the proposed OsteoHRNet for the knee OA severity prediction. Blocks with different colors denote convolution features at different spatial scales. The proposed model takes knee X-Ray image as input and estimates the OA severity in terms of KL grade

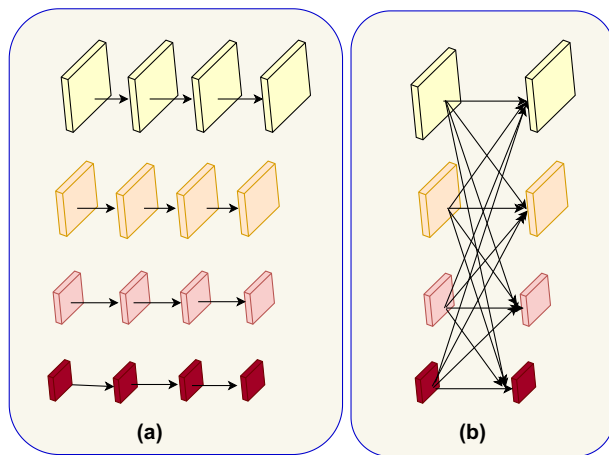


Fig. 3 Connections in HRNet: (a) Multi Resolution convolution in parallel, (b) Fusion of Multi-Resolution convolution. Different colors denote feature resolution at various scales

highly correlated and semantically robust spatial features. This motivates us to incorporate HRNet for processing the knee X-Ray images, which lack such rich spatial features.

To formally define, let \mathcal{D}_{ij} denotes the sub-network in the i^{th} stage of j^{th} resolution index. The spatial resolution in this branch is $1/2^j - 1$ of that of the high-resolution (HR) branch. For *e.g.*, HRNet, which consists of four different resolution scales, can be illustrated as follows:

$$\begin{array}{ccccccc}
 \mathcal{D}_{11} & \rightarrow & \mathcal{D}_{21} & \rightarrow & \mathcal{D}_{31} & \rightarrow & \mathcal{D}_{41} \\
 & \searrow & \mathcal{D}_{22} & \rightarrow & \mathcal{D}_{32} & \rightarrow & \mathcal{D}_{42} \\
 & & & \searrow & \mathcal{D}_{33} & \rightarrow & \mathcal{D}_{43} \\
 & & & & & \searrow & \mathcal{D}_{44},
 \end{array} \quad (1)$$

Later, the obtained multi-resolution feature maps are fused to exchange the learned variscaled information, as shown in Fig. 4. For this, HRNet utilizes bilinear upsampling followed by the 1×1 convolution to adjust the number of channels when transforming the lower resolution feature map to a higher resolution scale, or a strided 3×3 convolution otherwise.

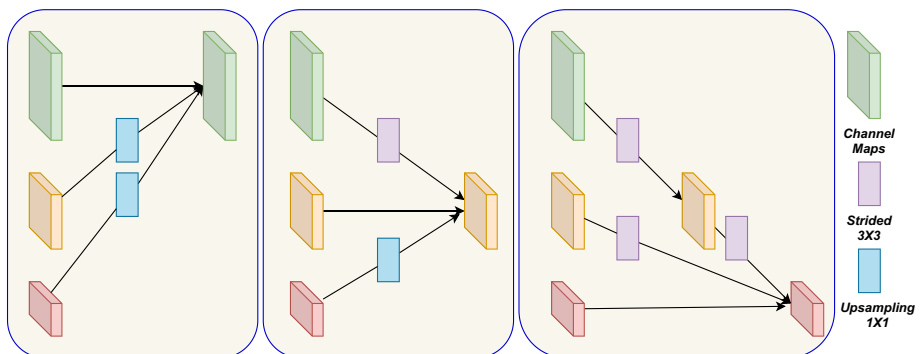


Fig. 4 Graphical demonstration of how HRNet fuses information from different resolutions

3.2 Convolutional block attention module

Convolutional Block Attention Module (CBAM) consists of two sequential sub-modules : (a) channel attention module, and (b) spatial attention module [43]. Given an input feature map, $\mathbf{P} \in \mathbb{R}^{C \times H \times W}$, CBAM sequentially infers a one-dimensional channel attention map $Map_c \in \mathbb{R}^{C \times 1 \times 1}$ and a two-dimensional spatial attention map $Map_s \in \mathbb{R}^{1 \times H \times W}$. Thus a final refined attention map is obtained, here denoted as \mathbf{T} , and the comprehensive attention mechanism can be summarized as:

$$\begin{aligned}\mathbf{P}^c &= Map_c(\mathbf{P}) \otimes \mathbf{P}, \\ \mathbf{T} &= Map_s(\mathbf{P}^c) \otimes \mathbf{P}^c,\end{aligned}\quad (2)$$

where \otimes signifies element-wise multiplication. Map_c is first generated by making use of the cross-channel relationship of the features, as,

$$Map_c(\mathbf{P}) = g(MLP(\mathcal{A}(\mathbf{P}))) + MLP(\mathcal{M}(\mathbf{P})), \quad (3)$$

where g , MLP , \mathcal{A} , and \mathcal{M} denote sigmoid function, multi-layer perceptron, average pool and max pool, respectively.

Whereas, the Map_s is generated efficiently by performing \mathcal{M} and \mathcal{A} along the channel axis. Next, the pooled descriptors are concatenated together to generate a reliable and efficient feature descriptor by utilizing the inter-spatial correlation of the features. It can be written as,

$$Map_s(\mathbf{P}) = g(k^{7 \times 7}([\mathcal{A}(\mathbf{P}); \mathcal{M}(\mathbf{P})])), \quad (4)$$

where $k^{7 \times 7}$ denotes the convolution operation with kernel of size 7×7 .

3.3 Network architecture

In this work a deep CNN, called OsteoHRNet, has been proposed, that utilizes the HRNet as the backbone and is further empowered with an attention mechanism for the knee KL grade classification. CBAM is integrated at the end of the HRNet, followed by a fully connected (FC) output layer, as depicted in Fig. 2. It may be said that the integration of the CBAM module after HRNet has been beneficial in learning adaptive enriched features for an efficient KL grade classification. It can also be observed that the proposed one-time integration of CBAM is computationally pleasant, compared to the multiple additions in the existing work [46]. The resultant output from the CBAM is then fed into the final fully connected layer, which outputs the probabilities of the KL grade for the given input X-Ray image. HRNet has been considered for reliable feature extraction, whereas the capabilities of CBAM are leveraged to help the model better focus on relevant features. The resolution for input x-ray image has been set to 224×224 .

3.4 Cost functions

A majority of the existing works [40, 46] on knee OA have adopted traditional cross entropy for classification which treats the KL grade as nominal data and penalises every misclassification equally. However, inspired by the idea of Chen et al. [4], this task has been approached as an ordinal regression problem and therefore ordinal loss function has been utilised instead of the traditional cross-entropy. The ordinal loss function used in this paper is a weighted ratio of the traditional cross-entropy. Given the ordinality in the KL grading, it must be acknowledged that extra information is provided by progressive grading [30]. This approach penalizes the distant grade misclassification more than the nearby grade according

to the penalty weights. For *e.g.*, a grade 1 classified as grade 3 is penalized more severely than it is classified as grade 2 and even more for being classified as grade 4. An ordinal matrix $C_{n \times n}$ is considered as the penalty weights between the outcome and the true grade, i.e., c_{uv} denotes the penalty weight for predicting a grade v as u with $n = 5$. In this study, with five KL grades to classify and $c_{uu} = 1$, the adopted ordinal loss can be written as

$$\mathcal{L}_o = \sum_{u=0}^{n-1} c_{uv} * q_u, \quad (5)$$

where u, v are the predicted and true KL grades of the input image, respectively, p_u is the output probability by the final output layer of the architecture with $q_u = p_u$ if $u \neq v$ and $q_u = 1 - p_u$, otherwise. The following penalty matrix has been utilised for our experimentation.

$$\begin{bmatrix} 1 & 3 & 6 & 7 & 9 \\ 4 & 1 & 4 & 5 & 7 \\ 6 & 4 & 1 & 3 & 5 \\ 9 & 7 & 4 & 1 & 4 \\ 11 & 9 & 7 & 5 & 1 \end{bmatrix}$$

If the matrix is closely observed, it could be noted that nearby grade misclassifications are assigned less penalty while far away grade misclassifications are assigned high penalty. The diagonal weights are 1 as in that case the predicted grade is equal to the true grade. To explain further, if first row (Grade 0) is considered, predicting it incorrectly as Grade 1 has been assigned a penalty weight of 3, similarly predicting it incorrectly as Grade 4 has been assigned the highest penalty weight of 9. The penalty weights employed are experimental and have been found to give the best results among several permutations of the penalty weight matrix.

4 Experimental details

4.1 Dataset

The X-ray radiographs acquired from the OAI repository, made available by Chen et al. [3], have been utilized in this study. The images obtained are of 4796 participants, including men and women. Given that, the focus is primarily on the KL grades, radiographs with annotated KL grades from the baseline cohort are acquired to assess our method. The dataset of a total of 8260 radiographs, including the left and right knee, was split into train, test, and validation sets in the ratio of 7:2:1 with balanced distribution across all KL grades [3]. Table 1 shows the train, test, and validation distribution of the dataset.

Table 1 Distribution of the dataset

Dataset	Grade0	Grade1	Grade2	Grade3	Grade4	Total
Training	2286	1046	1516	757	173	5778
Testing	639	296	447	223	51	1656
Validation	328	153	212	106	27	826
Total	3253	1495	2175	1086	251	8260

4.2 Training details

The entire code is developed using Pytorch [23] framework, and all the training and testing experiments have been conducted on a 12GB Tesla K40c GPU. Furthermore, the training of all the experimental models was optimized using stochastic gradient descent (SGD) for 30 epochs with an initial learning rate of $5e-4$. Additionally, owing to the GPU capacity, the batch size was set to 24.

4.3 Competing methods

In [4], the authors proposed to utilize the pre-trained VGG-19 [35] network with a novel ordinal loss function. Yong et al. [45] proposed to utilize the DenseNet-161 [10] with the ordinal regression module (ORM). The results obtained from OsteoHRnet have been compared against the results obtained by the above mentioned studies mentioned for a robust comparison. The results are presented in details in Section 5.

4.4 Evaluation metrics

In this study, the following evaluation metrics have been utilized to analyze and compare the performance of our proposed model: (a) Multi-class accuracy, (b) Mean Absolute Error (MAE), (c) Quadratic Weighted Cohen's Kappa coefficient (QWK), (d) Precision, (e) Recall, and (f) F1-score.

Traditionally, multi-class accuracy is defined as the average number of outcomes matching the ground truth across all the classes. Accuracy for five classes with N instances is formulated as below

$$Accuracy = \frac{1}{N} \sum_{i=1}^5 \sum_{x:g(x)=i} F(g(x) = \hat{g}(x)), \quad (6)$$

where, x is a test instance, $g(x)$ & $\hat{g}(x)$ are the true and predicted outcome respectively for the test instance, and F is a function which returns 1 if the prediction is correct and 0 otherwise.

MAE is the mean of the absolute error of the individual prediction over all the input instances. The error in the prediction value is determined by the difference between the predicted and the true value for that given instance. MAE for five classes with N instances can be expressed as below

$$MAE = \frac{\sum_{i=1}^N abs(y_i - \hat{y}_i)}{N}, \quad (7)$$

where, y_i & \hat{y}_i are the true and the predicted grade, respectively.

A weighted Cohen Kappa is a metric that accounts for the similarity between predictions and the actual values. The Kappa coefficient is a chance-adjusted index of agreement measuring the reliability of inter annotator for qualitative prediction. The Quadratic Weighted Kappa (QWK) is evaluated using a predefined table of weights which measures the extent of non-alignment between the two raters. The greater the disagreement, the greater the weight.

$$\kappa = 1 - \frac{\sum_{p,\hat{p}} w_{p,\hat{p}} O_{p,\hat{p}}}{\sum_{p,\hat{p}} w_{p,\hat{p}} E_{p,\hat{p}}}, \quad (8)$$

O is the contingency matrix for K classes such that $O_{p,\hat{p}}$ denotes the count of \hat{p} grade images predicted as p . The weight, is defined as

$$w_{p\hat{p}} = \frac{(p - \hat{p})^2}{(1 - K)^2}. \quad (9)$$

Next, E is calculated as the normalized product between the predicted grade's and original grade's histogram vector.

Precision measures the proportion of a class that was predicted to be in given class and are actually in that class. Precision value helps in identifying the quality of correct predictions by the model.

$$Precision_i = \frac{TP_i}{P_i}, \quad (10)$$

where i corresponds to a KL grade, TP_i corresponds to number of correctly predicted positives (True Positives) for a grade i , and P_i corresponds to the total positive predictions (True Positives and False Positives) made by the model for a grade i .

Recall measures the proportion of the true class predictions that were correctly predicted to be in a class over the number of true predictions possible.

$$Recall_i = \frac{TP_i}{N_i}, \quad (11)$$

where i corresponds to a KL grade, TP_i corresponds to number of correctly predicted positives for a grade i , and N_i corresponds to the total number of positive predictions possible for a grade i .

F1-score is defined as the measure of harmonic mean of precision and recall. F1-score provides more comprehensive insights of the performance.

$$F1 - Score_i = \frac{2 * (Precision_i * Recall_i)}{Precision_i + Recall_i}, \quad (12)$$

where i denotes the corresponding KL grade. The values for Precision, Recall and F1-score lies between 0 and 1.

5 Results & discussion

This section presents the quantitative and qualitative results, ablation study to justify the efficiency of each module in the network and finally followed by a discussion.

5.1 Comparison against state-of-the-art methods

It can be observed from Table 2, that the proposed method has outperformed the existing best-published works [4, 45] in terms of classification accuracy, MAE, and QWK. It should be mentioned that Yong et al. [45] reported the macro accuracy² and contingency matrix of their best model. For a fair comparison, equivalent to the above, their result has been reported as multi-class accuracy of 70.23%. Whereas Chen et al. [4] has reported the best multi-class accuracy of 69.69%. OsteoHRNet has reported a maximum multi-class accuracy of 71.74%, multi-class average accuracy of 70.52%, MAE of 0.311, and QWK of 0.869 which is a significant improvement over [4, 45]. Figure 5 represents the confusion matrix

²Macro accuracy: 88.09%

Table 2 Quantitative comparison against the existing methods [4, 45] in terms of multi-class accuracy, MAE, and QWK

Method	Accuracy	MAE	QWK
VGG 19 - Ordinal [4]	69.69 %	0.344	0.8460
DenseNet 161 - ORM [45]	70.23 %	0.330	0.8609
OsteoHRNET	71.74 %	0.311	0.8690

The bold entries signify the best results

obtained by using the proposed and existing methods [4, 45] which when fed with 1656 test images. Table 3 represents the comparison results in terms of precision, recall and F1-score. OsteoHRNet has obtained a mean F1-score of **0.722** while the mean F1-score obtained using Chen *et al*’s approach [4] and Yong *et al*’s approach [45] is 0.696 and 0.706 respectively. It is evident that OsteoHRNet outperforms the existing works in terms of all the above mentioned evaluation metrics.

Furthermore, Gradient-weighted Class Activation Maps (Grad CAM) [27] visualization technique has been employed to demonstrate the superiority and decision-making of the proposed OsteoHRNet. It helps in showcasing the most relevant regions which the network has learned to focus on in the X-ray images. Figures 6, 7, 8, 9, and 10 shows the qualitative comparison of the proposed model against the existing methods [4] in terms of Grad-CAM visualization. It can be observed that the proposed OsteoHRNet considers both features and the area between the knee joints for an efficient severity classification (*darker colors up the scales signifies more focus*). On close observation of the visualizations, it can clearly be said that OsteoHRNet is consistent in its decision making and gives importance to only the relevant areas in the knee X-rays while VGG-19 (used in [4]) is slightly confused and also focuses on irrelevant areas (can be identified by following the color shades of the VGG-19 visualization). The proposed model has efficiently learned the prominent features such as joint-space narrowing, osteophytes formations, and bone deformity, thus predicting the most relevant radiological KL grading. Moreover, it can be said that the decision-making of OsteoHRNet aligns in accordance with the actual real-world medical criterion of KL grade classification. This validates the enriched and superior results obtained by the proposed OsteoHRNet model.

5.2 Ablation study

This section presents an ablation study to demonstrate the contributions made by each sub-module of the proposed OsteoHRNet. For this, the following baselines have been performed:

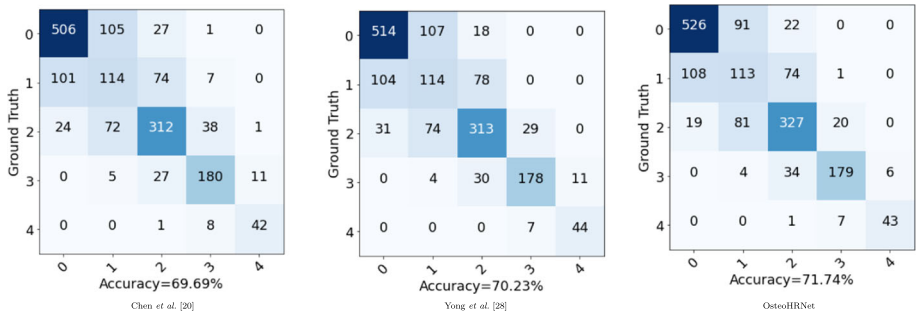


Fig. 5 Confusion matrices for KL grade prediction using different competing approaches [4, 45] and OsteoHRNet

Table 3 Quantitative comparison against the existing methods [4, 45] in terms of Precision, Recall and F1-score

KL Grade	VGG-19 [4]			Densenet-161 [45]			OsteoHRNet		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Grade 0	0.80	0.79	0.80	0.79	0.80	0.80	0.81	0.82	0.81
Grade 1	0.39	0.39	0.39	0.38	0.39	0.38	0.39	0.38	0.39
Grade 2	0.71	0.70	0.70	0.71	0.70	0.71	0.71	0.73	0.72
Grade 3	0.77	0.81	0.79	0.83	0.80	0.81	0.86	0.80	0.83
Grade 4	0.78	0.82	0.80	0.80	0.86	0.83	0.88	0.84	0.86
Average	0.69	0.702	0.696	0.702	0.71	0.706	0.73	0.714	0.722

The bold entries signify the best results

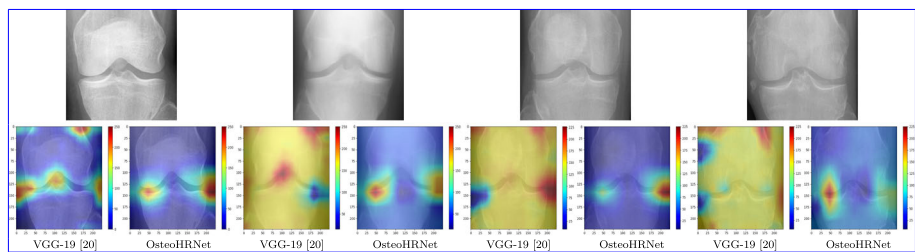


Fig. 6 Grad-CAM visualizations generated against KL grade 0 test images using VGG-19 of Chen et al. [4] and OsteoHRNet

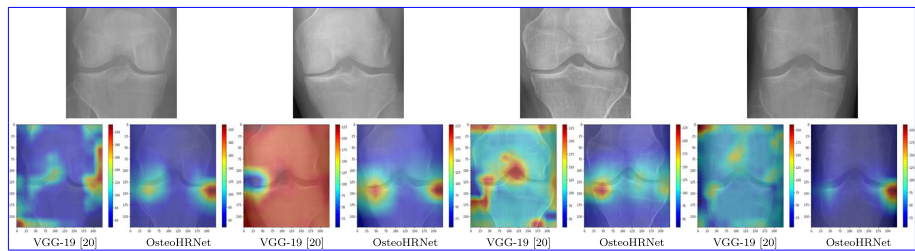


Fig. 7 Grad-CAM visualizations generated against KL grade 1 test images using VGG-19 of Chen et al. [4] and OsteoHRNet

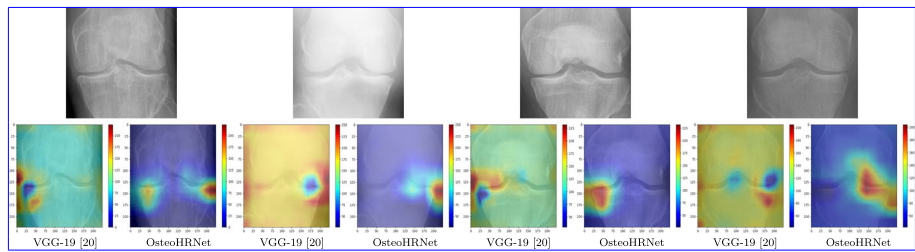


Fig. 8 Grad-CAM visualizations generated against KL grade 2 test images using VGG-19 of Chen et al. [4] and OsteoHRNet

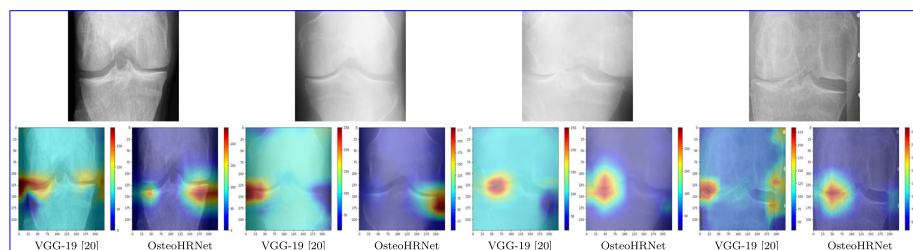


Fig. 9 Grad-CAM visualizations generated against KL grade 3 test images using VGG-19 of Chen et al. [4] and OsteoHRNet

1. **HRNet**: Original HRNet trained by utilizing the adopted dataset.
2. **HRNet + CBAM**: Original HRNet followed by the CBAM module trained using the adopted dataset.
3. **OsteoHRNet**: Original HRNet followed by the CBAM module trained using the adopted dataset. Further, during training, data augmentation techniques have been adopted to enhance the performance of the proposed model.

It can be observed from Table 4, how the addition of CBAM module has improved the performance compared to the raw HRNet module. Similarly, on incorporating data augmentation with CBAM, the performance has improved immensely compared to its curated baseline. The CBAM module has helped to adaptively learn the relevant features from the HRNet. Such features may have contributed more towards an efficient classification compared to the features learned by the original HRNet [42], VGG-19 [35] in Chen *et al.*'s study [4], or DenseNet161 [10] in Yong *et al.*'s study [45].

Figure 13 demonstrates the Grad-CAM visualizations for the ablation study. It can be observed that the proposed OsteoHRNet has learned the robust features progressively on addition of each component of our proposed network. It must be observed both qualitatively (from Fig. 13) and quantitatively (from Table 4), how the decision making of the model gets refined on usage of ordinal loss against the traditional cross entropy. To explain it further, let us consider the second column of Fig. 13 for Grade 1, if followed along down the column, each cell represents the visualizations of different baseline models. The decision making of the model keeps improving as the sub-modules are added to the pipeline, and reaches its best in the last visualization, which is obtained with OsteoHRNet. Thus, it is verified that each component of our network contributes to the final knee OA KL grade prediction.

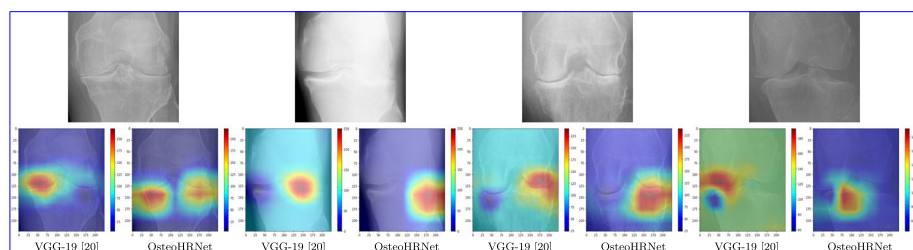


Fig. 10 Grad-CAM visualizations generated against KL grade 4 test images using VGG-19 of Chen et al. [4] and OsteoHRNet

Table 4 Effects of different network modules & cost function

Architecture	Cross Entropy		Ordinal Loss	
	Accuracy	MAE	Accuracy	MAE
HRNet	64.10 %	0.460	65.00 %	0.440
HRNet + CBAM	65.30 %	0.423	66.70 %	0.392
OsteoHRNet	69.90 %	0.373	71.74 %	0.311

The bold entries signify the best results

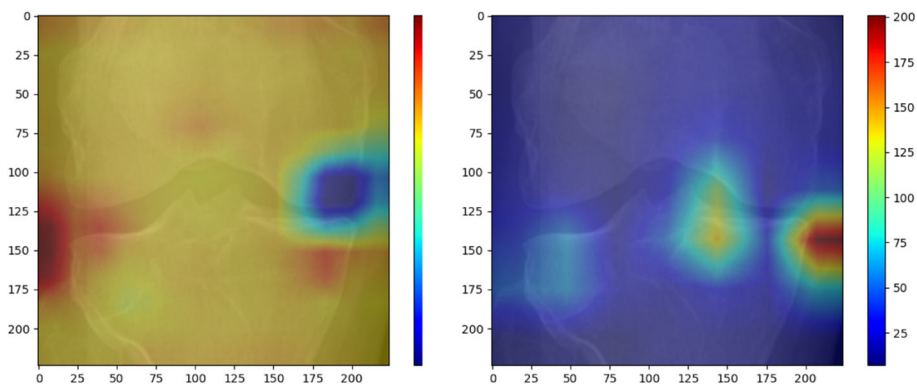


Fig. 11 Grad-CAM visualization for the incorrect classification by Chen et al. [4] (VGG-19; *left*) and proposed OsteoHRNet (*right*) for grade 2 radiograph

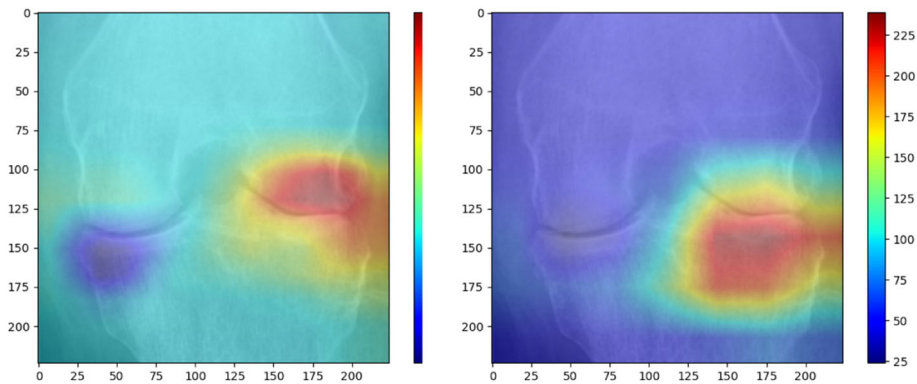


Fig. 12 Grad-CAM visualization for the incorrect classification by Chen et al. [4] (VGG-19; *left*) and proposed OsteoHRNet (*right*) for grade 4 radiograph

5.3 Discussion

It is evident from Fig. 5 that the OsteoHRNet has outperformed the previous works [4, 45], significantly. It should be mentioned that the OsteoHRNet classifies the higher grade X-rays very accurately while reducing the misclassification between far away grades. In comparison to existing methods, there has been a significant increase in correct classifications for grade 2. Furthermore, the nearby misclassifications between higher grades (grade 2-grade 3, grade 3-grade 4) are minimum for the proposed method, which needs to

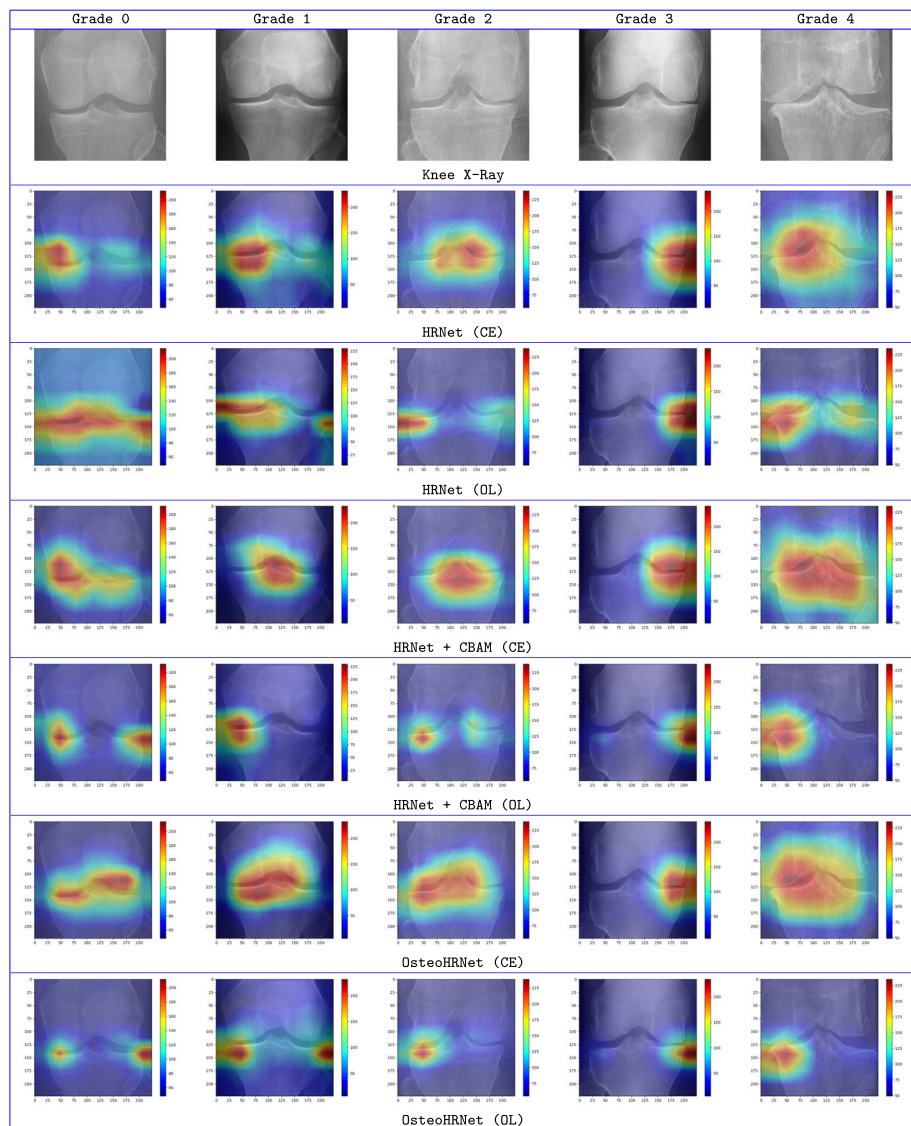


Fig. 13 Grad-CAM visualizations for the ablation study. CE stands for Cross-entropy and OL stands for Ordinal Loss

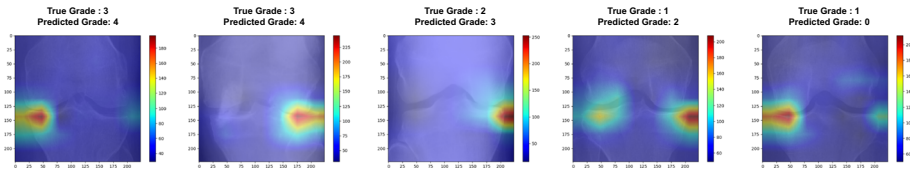


Fig. 14 Grad-CAM visualizations for the incorrectly classified radiographs obtained by using OsteoHRNet

be acknowledged. Also, by way of analysis using obtained Grad-CAM visualization of such incorrect classifications, it can be observed that OsteoHRNet is trying to locate joint space narrowing and osteophytes in accordance with the medical characteristics. At the same time, VGG-19 [4] is confused and focuses on the entire knee, giving importance to irrelevant features for KL grade classification, as seen in Figs. 11 and 12.

Owing to its superior network learning, our model is extremely relevant to the medical setting of KL grade classification. Furthermore, the Grad-CAM visualization of our model can be extended for the use of the medical practitioner to provide confidence in the findings. However, our study has some limitations, and certain radiographs could not be correctly classified due to the lack of rich features in the radiographs (Fig. 13). Figure 14 shows nearby grade misclassifications, which to a great extent is unavoidable. But, there is high inter and intra-observer variability (correlation coefficient = 0.83) for manual knee KL grading [29]. Thus, our proposed fully automated KL grading method can be extended in clinical settings for getting reliable and reproducible OA grading. However, there exists some limitations in our experiments. The choice of penalty matrix is empirical. Although exhaustive sets of experiments have been performed to choose the best set of penalty matrix, there may exist even better set of weights which can yield better performance.

6 Conclusion

This paper proposes a novel OsteoHRNet by adopting the HRNet as the backbone and integrating the CBAM module for an improved knee OA severity prediction results from plain radiographs. The proposed network was able to perform exceptionally well and attain significant improvements over the previously proposed methods owing to the HRNet's capability to maintain high-resolution features throughout the network and its ability to capture reliable spatial features. The intermediate extracted features were significantly refined with the help of the attention mechanism; therefore, the radiographs with a similarity between classes and variations within classes could be distinguished better. Moreover, the Grad-CAM visualizations have been employed to validate that the model has learned the most relevant spatial features in the radiographs. In the future, we will work on the entire OAI multi-modal data and consider all the cohorts in our study.

Declarations

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Abd Razak HR, Andrew TH, Audrey HX (2014) The truth behind subchondral cysts in osteoarthritis of the knee. *The open orthopaedics journal*, p 8
2. Antony J, McGuinness K, O'Connor NE, Moran K (2016) Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks, pp 1195–1200
3. Chen P (2018) Knee osteoarthritis severity grading dataset
4. Chen P, Gao L, Shi X, Allen K, Lin Y (2019) Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Comput Med Imaging Graph* 75:84–92
5. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. 1:539–546 vol 1
6. Deepikaraj A, Ballal R, Shashikala R, Shetty DM (2017) Analysis of osteoarthritis in knee x-ray image. *International Journal of Scientific Development and Research - IJSDR* 2(6):416–422
7. Dubois R, Herent P, Schiratti J-B (2021) A deep learning method for predicting knee osteoarthritis radiographic progression from mri. *Arthritis Res Ther*, p 23
8. Gold GE, Braun HJ (2012) Diagnosis of osteoarthritis: imaging. *Bone* 51(2):278–288
9. Górriz M, Antony J, McGuinness K, Giró-i-Nieto X, O'Connor NE (2019) Assessing knee oa severity with cnn attention-based end-to-end architectures. 102:197–214, 08–10
10. Huang G, Liu Z, Van Der Maaten L, Kilian Q (2018) Weinberger Densely connected convolutional networks
11. Kashyap S, Zhang H, Rao K, Sonka M (2018) Learning-based cost functions for 3-d and 4-d multi-surface multi-object segmentation of knee mri: Data from the osteoarthritis initiative. *IEEE Trans Med Imaging* 37(5):1103–1113
12. Kaur S, Goel N (2020) A dilated convolutional approach for inflammatory lesion detection using multi-scale input feature fusion (workshop paper). In: 2020 IEEE 6th International Conference on Multimedia Big Data (BigMM), pp 386–393
13. Kellgren JH, Lawrence JS (1957) Radiological assessment of osteo-arthritis. *Ann Rheum Dis* 16(4):494–502
14. Khalid H, Hussain M, Al Ghamdi MA, Khalid T, Khalid K, Khan MA, Fatima K, Masood K, Almotiri SH, Farooq MS, Ahmed A (2020) A comparative systematic literature review on knee bone reports from mri, x-rays and ct scans using deep learning and machine learning methodologies. *Diagnostics*, 10(8)
15. Kokkoti C, Moustakidis S, Papageorgiou E, Giakas G, Tsaopoulos DE (2020) Machine learning in knee osteoarthritis: A review. *Osteoarthritis Cartil* Open 2:100069, 05
16. Kundu S, Ashinsky BG, Bouhrara M, Dam EB, Demehri S, Shifat-E-Rabbi M, Spencer RG, Urish KL, Rohde GK (2020) Enabling early detection of osteoarthritis from presymptomatic cartilage texture maps via transport-based learning. *Proc Natl Acad Sci* 117(40):24709–24719
17. Liu X, Zheng X, Li W, Xiong J, Wang L, Zomaya AY, Longo A, Tang C (2020) Defusionnet: Defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features. *IEEE Trans Pattern Anal Mach Intell* 44(2):955–968
18. Mathur M, Sahoo S, Jain D, Goel N, Vasudev D (2020) Crosspooled fishnet: transfer learning based fish species classification model. *Multimed Tools Appl* 79(41):31625–31643
19. Nishio M, Do RKG, Yamashita R et al (2018) Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 9:611–629
20. O'Connor NE, Antony J, McGuinness K, Moran K (2017) Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks
21. Oka H, Muraki S, Akune T, Mabuchi A, Suzuki T, Yoshida H, Yamamoto S, Nakamura K, Yoshimura N, Kawaguchi H (2008) Fully automatic quantification of knee osteoarthritis severity on plain radiographs. *Osteoarthritis Cartil* 16(11):1300–1306
22. Pariyo GB, Agarwal AK, Vijay V, Vaishya R (2016) Non-operative management of osteoarthritis of the knee joint. *J Clin Traumatol-Sur* 7(3):170–176
23. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Lu F, Bai J, Pytorch SC (2019) An imperative style, high-performance deep learning library. Curran Associates Inc. pp 8024–8035
24. Peterfy CG, Schneider E, Nevitt M (2008) The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis Cartil* 16(12):1433–1441
25. Piuze NS, Husni ME, Muschler GF, Guarino A, Mont MA, Lepsasio MJ (2017) Knee osteoarthritis: A primer. *The Permanente journal*, p 21

26. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP, Ng AY (2017) Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning
27. Ramprasaath Rs, Das A, Vedantam R, Cogswell M, Parikh D, Batra D (2016) Grad-cam: Why did you say that? 11
28. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2019) Mobilenetv2: Inverted residuals and linear bottlenecks
29. Sassoon AA, Fernando ND, Kohn MD (2016) Classifications in brief: Kellgren-lawrence classification of osteoarthritis. *Clinical orthopaedics and related research*, 474(8)
30. Sassoon AA, Fernando ND, Kohn MD (2016) Classifications in brief: Kellgren-lawrence classification of osteoarthritis. *Clin Orthop Relat Res* 474(8):1886–1893
31. Schlüter-Brust KU, Eysel P, Michael JW (2010) The epidemiology, etiology, diagnosis, and treatment of osteoarthritis of the knee. *Deutsches Arzteblatt international* 107(9)
32. Shamir L, Ling SM, Scott WW, Bos A, Orlov N, Macura TJ, Mark Eckley D, Ferrucci L, Goldberg IG (2009) Knee x-ray image analysis method for automated detection of osteoarthritis. *IEEE Trans Biomed Eng* 56(2):407–415
33. Shamir L, Ling SM, Scott W, Hochberg M, Ferrucci L, Goldberg IG (2009) Early detection of radiographic knee osteoarthritis using computer-aided analysis. *Osteoarthr Cartil* 17(10):1307–1312
34. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition
35. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *International conference on learning representations*
36. Swart-NM, Bloem JL, Bierma-Zeinsträ S, Algra PR, Bindels P, Koes BW, Nelissen R, Verhaar J, Luijsterburg P, Reijnen M, van den Hout WB, van Oudenaarde K (2018) General practitioners referring adults to mr imaging for knee pain: A randomized controlled trial to assess cost-effectiveness. *Radiology* 288(1)
37. Swiecicki A, Li N, O'Donnell J, Said N, Yang J, Mather RC, Jiranek WA, Mazurowski MA (2021) Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. *Comput Biol Med* 133:104334
38. Tang C, Liu X, An S, Pichao W (2021) Br²net: Defocus blur detection via a bidirectional channel attention residual refining network. *IEEE Trans Multimed* 23:624–635
39. Thevenot J, Rahtu E, Lehenkari P, Saarakkala S, Tiulpin A (2018) Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific Reports*
40. Tiulpin A, Thevenot J, Rahtu E, Saarakkala S (2017) A novel method for automatic localization of joint area on knee plain radiographs
41. Van Noord N, Postma E (2017) Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recogn* 61:583–592
42. Wang J, Ke S, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, Liu W, Xiao B (2020) Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Anal Mach Intell*, pp 1–1
43. Woo S, Park J, Lee J-Y, Kweon I (2018) Cbam: Convolutional block attention module. p 07
44. Yadav SS, Jadhav SM (2019) Deep convolutional neural network based medical image classification for disease diagnosis. *J Big Data* 6:113
45. Yong C, Teo K, Murphy B, Hum Y, Tee Y, Xia K, Lai KW (2021) Knee osteoarthritis severity classification with ordinal regression module. *Multimed Tools Appl*, 01
46. Zhang B, Tan J, Cho K, Chang G, Deniz CM (2020) Attention-based cnn for kl grade classification Data from the osteoarthritis initiative, pp 731–735

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.