Automatic Grading of Knee Osteoarthritis from Plain Radiographs Using Densely Connected Convolutional Networks

Alexey Mikhaylichenko and Yana Demyanenko

Institute of Mathematics, Mechanics and Computer Science, Southern Federal University, Rostov-on-Don, Russia
alexey.a.mikh@gmail.com, demyanam@gmail.com

Abstract. In this paper, we consider densely connected convolutional networks and their applicability to the problem of assessment of knee osteoarthritis (OA) severity in the five-point Kellgren-Lawrence scale. First, we use trained from scratch Single Shot Detector (SSD) to localize knee joint areas in radiographs. Then, we apply DenseNets to quantify OA stages in the images of detected knee joints. We consider networks of different depths, trained both from scratch and pre-trained on the ImageNet dataset and fine-tuned in the images from Osteoarthritis Initiative dataset (OAI). Also, different loss functions are examined to understand which one gives the best training results. In the knee joint localization task, we obtain an accuracy of 94.03% under the Jaccard index threshold of 0.75. Also, our classifier outperforms the current state-of-the-art with accuracy of 71% in the classification task.

Keywords: Knee osteoarthritis · Kellgren and Lawrence grading · Convolutional neural network · Classification

1 Introduction

Osteoarthritis (OA) is one of the most common diseases of the musculoskeletal system. Such diseases reduce the patient’s working capacity and lead to large socio-economic losses. A significant part of the costs appears due to the late diagnostics of disease. Currently, there is no effective treatment for osteoarthritis, except for the replacement of the entire joint with an artificial one after its complete destruction. Therefore, the only way to reduce the costs is the early diagnosis when it is still possible to slow down the process of joint destruction.

The most common method of knee osteoarthritis diagnostics relies on X-rays images analysis. The main symptoms of OA are the degeneration and wear of the articular cartilage. However, cartilage tissue is not directly visible in radiographs. Therefore, with using X-rays, the progression of OA is assessed mainly by indirect signs—the narrowing of joint space or the appearance of osteophytes.
Osteoarthritis is a phased disease. Despite the existence of various systems for assessing the stages of osteoarthritis progression, the diagnosis is highly dependent on the subjectivity and experience of the expert. In addition, the widely used Kellgren-Lawrence grading scale [1] is very ambiguous [5,6]. The use of automated analysis could reduce the influence of the subjectivity factor and increase the accuracy and the reliability of the diagnosis.

2 Related Works

In the paper [7] OA stage classification is examined using Decision Trees, Naive Bayes classifier, Bayesian (Probabilistic) networks, and logistic regression by various texture characteristics (histogram features, Haralick features, etc.). There are 130 X-ray images applied in this study.

Minciullo et al. [8] present an original approach to classify OA using shape information from lateral knee radiographs. In this work, the authors use a statistical shape model for key points of knee contour detection, which are then utilized to the OA classification with random forests. The achieved performance for OA classification to 5 grades is 47.9% in the set of 300 images. The authors declare that it is comparable with similar techniques applied to the frontal view.

Antony et al. [9] suggest simple, trained from scratch convolutional neural network (CNN) which consists of 5 convolutional layers and a fully-connected layer. They optimize a weighted combination of cross-entropy loss and mean squared error, and achieve classification accuracy at about 60%. In the study [10], the authors use pre-trained CNNs with various architectures for classifying the OA grade, which are fine-tuned in the OAI dataset.

Tiulpin et al. [14] present a new CNN model not only for OA classification but also for the assessment of the joint space narrowing and the osteophytes presence. Their approach is based on the ensembling of two neural networks, each of which also consists of two parts: the pre-trained on the ImageNet convolutional layers, and 7 independent fully-connected layers. To connect these convolutional and fully connected layers, they utilize an average pooling layer after the convolutional block. They evaluate various network backbones from Resnet family with squeeze-excitation (SE) blocks [13]. The reached average multi-class accuracy is 66.68% for the OA grading task. In [15] the Deep Siamese CNNs are examined. They use the symmetry in the image, and their network consists of two branches, each of which works with different parts of the knee joint. The classification accuracy is 67.49%.

Wahyuningrum et al. [18] suggest a three-step method: preprocessing, features extraction using simple VGG-based CNN, and classification of the knee OA severity using LSTM. For the analysis, the work uses a 400 × 100 images of the knee joint area. The LSTM network is trained by optimizing Stochastic Gradient Descent (SGD). The average classification accuracy is 75.28%, but the accuracy of KL grade 1, presented in the article, is very different from the corresponding accuracy of all previous works and results of our experiments and needs to be rechecked.
Liu et al. [21] use the Faster R-CNN [20] for the simultaneous location of the knee joint area and classification OA severity grade for this area. The focal loss function is used for training to address the class imbalance. The own dataset containing 1385 X-ray images is used as a dataset for the analysis instead of the OAI dataset. The achieved classification accuracy is 74.3%.

Pingjun et al. [11] propose an adjustable ordinal loss instead of cross-entropy loss, to assign a higher penalty to misclassification with a larger distance between the real and the predicted KL grades. Several networks of well-known architectures, such as ResNet, VGG, etc., were trained in this work. The best result (69.7%) was obtained on the fine-tuned VGG-19 model with the proposed ordinal loss.

In this paper, we study the applicability of DenseNets to the problem of assessment knee osteoarthritis severity. We consider networks of different depths, both trained from scratch and pre-trained on the ImageNet, and evaluate different loss functions to improve the training process. Different ensembling methods are used for evaluation.

DenseNets is also used in [19], but for the case of the abbreviated Kellgren-Lawrence scale—KL grades of 0–1 were grouped since the clinical response for these two grades are usually similar. U-Net network is utilized for knee joints localization. Also, the authors considered only ensembles of different model checkpoints and only cross-entropy loss function for training.

3 Materials and Methods

3.1 Osteoarthritis Grades

There are various systems for assessing the progression of knee osteoarthritis. One of the standards is the Kellgren-Lawrence (KL) grading scale [1,2]. There are 5 grades of OA severity in this scale. A brief description of each stage is given in the Table 1, samples of knee joint images for each grade are shown in Fig. 1.

| Grade | Verbal description                                      |
|-------|--------------------------------------------------------|
| Grade 0 | None: no JSN or reactive changes                       |
| Grade 1 | Doubtful: doubtful JSN, possible osteophytic lipping   |
| Grade 2 | Minimal: definite osteophytes, possible JSN            |
| Grade 3 | Moderate: moderate osteophytes, definite JSN, some sclerosis, possible bone-end deformity |
| Grade 4 | Severe: large osteophytes, marked JSN, severe sclerosis, definite bone ends deformity |

The Kellgren-Lawrence scale has been criticized for the emphasis on the osteophytes, the overall grading of OA from normal to severe, and insensitivity
to changes [3]. Osteoarthritis of the minimum severity is diagnosed in grade 2 [4], therefore, in some studies, stage 0 and 1 are combined into one (normal).

![Fig. 1. Knee joint samples of significant KL grades](image)

### 3.2 Data

The Osteoarthritis Initiative (OAI) dataset is a standard dataset for research on knee osteoarthritis using radiographs. It contains an archive of clinical data from 4,796 patients aged 45 to 79 years. In addition, each radiograph contains information about the measurements, osteophytes, joint space narrowing, the age of the patient, the stage of OA (including Kellgren-Lawrence grades) taken from few experts, etc.

In this study, we used bilateral posterior-anterior (PA) fixed-flexion knee radiographs for our experiments. Since the OAI dataset does not contain ground truth for knee joint detection, we manually annotate knee joints on all images.

After filtering, when low-quality images (very blurry, overexposed, etc.) were removed from the dataset, 4130 X-ray images with 8260 knee joints were remained. We randomly split all radiographs into train, validation, and test sets with a ratio of 7:1:2. The distribution of the knee joint images with OA on the Kellgren-Lawrence scale is given in Table 2:

| Group     | KL-0 | KL-1 | KL-2 | KL-3 | KL-4 | Images |
|-----------|------|------|------|------|------|--------|
| Train     | 2295 | 1051 | 1504 | 752  | 175  | 6604   |
| Validation| 319  | 148  | 223  | 111  | 25   | 826    |
| Test      | 639  | 296  | 447  | 223  | 51   | 1656   |

The OAI dataset is a highly unbalanced dataset, therefore we utilized data augmentations during models training. We applied random horizontal flip, random rotation and scale, saturation, and brightness. After all of these augmentations, we were rescaling the images to $224 \times 224$. All transformations were performed in random order on the fly.
3.3 Knee Joints Localization

Mainly, there are two steps in assessing knee osteoarthritis severity from raw X-ray images: knee joint localization and classifying the detected knee joint into OA grades. Several methods have been developed for knee joint detection in modern works.

One of the simplest methods is given in [17]—localization of the knee joint area there is performed by calculating the center of mass of the histogram, calculated by the sum of intensities of image rows. Similar histograms are also used in [12]. In [10], it is proposed to use the SVM classifier to search for ROI. In paper [16] for knee joint detection template matching is used. Solutions using deep learning have also been proposed: trained from scratch fully convolutional network (FCN) [9], and customized YOLOv2 network [11].

In this paper, we consider Single Shot Detector (SSD) for knee joint detection. This architecture has higher accuracy and performance with a comparison with the previous approaches like YOLO, R-CNN, etc. [22]. This network composes of two parts—features maps extraction from the input image, and applying convolution filter to detect objects. For features extraction mainly used pre-trained well-know architectures—VGG-16, Resnet-101, Inception [23] etc. We utilized MobileNet [24], pre-trained on COCO (Common Objects in Context) dataset. This architecture allows achieving high accuracy comparable with the state-of-the-art approaches while having significantly less number of parameters and high speed.

3.4 Quantifying Osteoarthritis Severity

In this study, we investigate the use of networks from DenseNet family [25] for assessing the severity of knee OA through classification. We were interested in the claim that dense connections have a regularizing effect, which reduces overfitting on tasks with the smaller training sets [25]. It means, that DenseNets work well on small datasets, which is important in our case. We consider networks of various depth, different loss functions, and different training processes.

As our initial approach, we trained from scratch DenseNet-121 and DenseNet-161, where the top fully connected layer was replaced with the layer with 5 outputs of the Kellgren-Lawrence grades.

Next, we examined the use of pre-trained on the ImageNet networks to solve the problem. In various options, we used both the training only the last fully-convolutional layer (freezing weights on convolutional layers), and fine-tuning the entire network.

Then we considered the effectiveness of applying various loss functions to train networks in both cases. We used the cross-entropy loss as the main loss function. We also experimented with adjustable ordinal loss [11].

In addition to the experiments, we also investigated using the convolution layer with window size equal to the layer size between convolutional and fully-connected layers instead of a simple average pooling. We assumed that some
information is lost in this transition, and the result can be improved by chang-
ing the transition layer. However, this modification did not lead to significant
improvements. Therefore, we present only the results with the average pooling
in the paper.

3.5 Models Ensemble

A neural network ensemble combines the outputs of multiple individually trained
neural networks in order to improve generalization ability and reduce dispersion
in model predictions. In most cases, the best results of large machine learning
competitions are achieved by an ensemble of models, rather than by a single
model.

There are many different types of ensembles. One of them is stacking, which
involves training a new learner to combine the predictions of several models.
We used the most common ensemble approach (the simplest form of stacking)—
unweighted averaging [26]. Subsequently, we picked the models, summed their
predictions, and propagated them through the softmax layer. Eventually, the
class probability of the KL grade \(k\) for given image \(x\) was inferred as follows:

\[
P(y = k|x) = \frac{\exp \left[ \sum_{m=1}^{M} P_m(y = k|x) \right]}{\sum_{j=1}^{K} \exp \left[ \sum_{m=1}^{M} P_m(y = j|x) \right]},
\]

where \(M = 3\) is the number of models in the ensemble, \(K = 5\) is number
of classes and \(P_m(y = k|x)\) an individual network output before the softmax
layer (unnormalised probability distribution). We considered different sets of
models as parts of ensembles: of same network with different training check-
points, of the same network with different random seeds (21, 42, and 84), and
an ensemble of networks with different structures (a combination of DenseNet-
121 and DenseNet-161). In the last two cases, we chose models that the best
performed on the validation set.

4 Experiments and Results

4.1 Knee Joints Detection

A comparison of the accuracy of the proposed method and methods from pre-
viously published works is given in Table 3. We use the Intersection over Union
(IoU), also known as the Jaccard index, between manual annotations and detec-
tion results to evaluate localization accuracy:

\[
IoU(A, B) = \frac{|A \cap B|}{|A \cup B|},
\]

where A and D are the manually annotated and the automatically detected
bounding boxes of knee joint, respectively. The numbers in the table indicate
the percentage of correctly detected knee joints, where the Jaccard index values
Table 3. A comparison of the accuracy of the proposed method and methods from previously published works, based on the Jaccard index (J). YOLOv2 and SSD based networks are fine-tuned models.

| Methods                      | $J \geq 0.25$ | $J \geq 0.5$ | $J \geq 0.75$ | Mean |
|------------------------------|---------------|---------------|---------------|------|
| Linear SVM + Sobel [10]      | ~81.8%        | 38.6%         | -             | 0.36 |
| Region proposal [12]         | -             | -             | -             | 0.8399 |
| FCN [9]                     | 100%          | 99.9%         | 89.2%         | 0.83 |
| YOLOv2 [11]                 | -             | -             | 92.2%         | 0.858 |
| SSD, our method             | 100%          | 100%          | 94.03%        | 0.844 |

are greater than 0.25, 0.5 and 0.75, and the mean value of Jaccard index is given for all samples. A dash means that the corresponding data is not available.

A comparison of the results shows that methods knee joint localization based on convolutional neural networks reach higher detection accuracy than other approaches. Moreover, simpler methods do not have any significant advantages over neural networks. Besides, the experiments have shown that using a pre-trained network in the localization problem allows greater accuracy compared to a network trained from scratch.

4.2 The Knee Osteoarthritis Classification on the Kellgren-Lawrence Scale

Training Process. All the models were trained using the adaptive moment estimation (Adam) optimizer [27] with a learning rate of $1e^{-3}$ and a weight decay (L2-norm regularization) of $1e^{-4}$. During the training, the learning rate was decayed by 5% every 5 epochs. The batch size was fixed to 64. In total, we trained every model for 75 epochs.

First, we re-implemented the model described in the paper [9]. This network produces two outputs—one for regression and the other one for classification. In our implementation, we used only classification output and 224 × 224 pixel images as the network input (instead of 300 × 200 from original work).

Secondly, we trained DenseNet-121 and DenseNet-161 from scratch on the OAI dataset. Next, we performed a fine-tuning of DenseNets that were pre-trained on the ImageNet dataset. We executed different experiments: both, we trained only the FC layer with a frozen convolution part, and we unfroze the convolutional layers and trained the full network. We found that training the full model gives a much better result than the training of the network with frozen convolution layers. Besides, the network with frozen convolution layers is very quickly overfitted. The results of testing only the fully trained models are described below. We used this procedure three times with different random seeds (21, 42, and 84) and selected model snapshots with the best accuracy.

We investigated training our DenseNets with two loss functions: cross-entropy loss and adjustable ordinal loss [11]. We implemented an adjustable ordinal
matrix, described in the original paper, to denote the penalty weights between the predicted and the real grade, and use it for the training process. However, validation performance in the multi-class average accuracy was comparable with the case using cross-entropy loss.

Finally, we used our model snapshots to make ensembles of these models and check them on the test set. The final results are presented in Table 4.

In the case of unbalanced datasets, a common practice is using various weighted data sampling strategies like under-sampling (removing samples from the majority class) or over-sampling (adding more examples from the minority class). We tested these strategies, however, they did not lead to an improvement in the final scores.

**Evaluation.** In our experiments, we got multi-class accuracy 61.78% for Antony’s CNN [9], which we trained only for classification using stochastic gradient descent and Nesterov momentum to reproduce the original results (60.3%). The improved accuracy may be due to the fact that we applied the L2-norm weight regularization for all layers, not just for the last two convolutional layers and the fully connected layer, as in [9].

**Table 4.** Comparison of different classifiers on the Kellgren-Lawrence grading task. We trained multiple networks and selected the best snapshots for each model. The main variants of DenseNet are compared with both cross entropy loss (CE) and ordinal loss (Ordinal) from [11]

| Method                | Accuracy | Precision | Recall | $F_1$ |
|-----------------------|----------|-----------|--------|-------|
| Antony et al. [9]     | 61.78%   | 0.58      | 0.62   | 0.55  |
| Tiulpin et al. [14]   | 66.68%   | –         | –      | –     |
| Siamese CNNs [15]     | 67.49%   | –         | –      | –     |
| VGG-19 [11]           | 69.70%   | –         | –      | –     |
| CNN-LSTM [18]         | 75.28%   | –         | –      | –     |
| Faster R-CNN [20]     | 74.30%   | –         | –      | –     |
| DenseNet-121-CE       | 68.98%   | 0.66      | 0.69   | 0.66  |
| DenseNet-121-Ordinal  | 67.39%   | 0.67      | 0.67   | 0.67  |
| DenseNet-161-CE       | 67.69%   | 0.67      | 0.68   | 0.66  |
| DenseNet-161-Ordinal  | 67.94%   | 0.66      | 0.68   | 0.66  |
| 3× DenseNet-121-DRS   | **71.08%** | 0.68      | **0.71** | 0.68  |
| 3× DenseNet-161-DRS   | 70.17%   | **0.69**  | 0.70   | **0.69** |
| 3× DenseNet-121-DCH   | 69.69%   | 0.66      | 0.70   | 0.66  |
| 3× DenseNet-161-DCH   | 70.10%   | 0.68      | 0.70   | 0.68  |
| 4× DenseNet-121/161   | 70.47%   | 0.68      | 0.70   | 0.67  |

We trained DenseNet-121 and DenseNet-161 with cross-entropy and adjustable ordinal losses (DenseNet-[121/161]-CE, DenseNet-[121/161]-Ordinal). The
ordinal loss showed better result than cross-entropy only for DenseNet-161. Despite this, the enhancement of this approach, probably, may provide better results than using cross-entropy loss.

Training such complex networks from scratch, as expected, did not bring any improvements and the accuracy of the classification was significantly less than in the case of pre-trained models, therefore, these results are not presented in the table above.

As we mentioned before, we experimented with different sets of models for ensembles: ensemble from the same model with different random seeds (3 × DenseNet-[121/161]-DRS), ensemble the same model from different checkpoints (3 × DenseNet-[121/161]-DCH) and ensemble from combination of two DenseNet-121 and two DenseNet-161 (4 × DenseNet-121/161). The best result—average multiple-class accuracy of 71.08%—was achieved by the ensemble of three DenseNets, trained with different random seeds (21, 42, and 84). Classification metrics and the confusion matrix with ROC curves for the best ensemble are presented in Table 5 and Fig. 2 respectively.

Table 5. Classification metrics achieved by the ensemble of three models, trained with different random seeds (3 × DenseNet-121-DRS)

| KL Grade | Accuracy  | Precision | Recall | F1-score | AUC  |
|----------|-----------|-----------|--------|----------|------|
| 0        | 92.02%    | 0.70      | 0.92   | 0.80     | 0.92 |
| 1        | 16.22%    | 0.44      | 0.16   | 0.24     | 0.78 |
| 2        | 72.26%    | 0.72      | 0.72   | 0.72     | 0.91 |
| 3        | 83.41%    | 0.82      | 0.83   | 0.82     | 0.98 |
| 4        | 62.75%    | 0.97      | 0.63   | 0.76     | 0.99 |
| Mean     | 71.08%    | 0.68      | 0.71   | 0.68     | –    |

Fig. 2. Confusion matrix and ROC curves of KL grading for the best ensemble of 3 × pre-trained DenseNet-121 with different random seeds. Average multi-class accuracy is 71.08%

The classification accuracy reported in [20] and [18] is higher than in our research, however, it is incorrect to compare our results with these cases directly.
To train and test the model in [20], a completely different dataset was used, several times smaller in volume (1385 base images) and less balanced (especially for grade 1, which gives the biggest error) than the OAI dataset.

The results of classification metrics for grade 1 in [18] differ fundamentally from the observations in both our and previous studies—the accuracy for grade 1 is comparable to the accuracy of other grades, while in all other studies this is not the case (see Table 5 for details). Therefore, the claimed results need to be rechecked and confirmed.

5 Discussion

Analysis of the confusion matrix and classification metrics for each class shows that the main misclassification occurs between neighboring classes. Moreover, the biggest error is noted for KL grade 1, due to the small variations between grade 0 and grade 1. This fact was also noted in previous works.

The analysis of the saliency map, calculated by the Grad-CAM algorithm [28] for the final DenseNet-121, shows that classification does not depend much on the accuracy of the joint area localization—the main pixels that make the greatest contribution to decision making are found directly in the area of osteophytes and in the joint space (Fig. 3). This hypothesis is supported by experiments that are used true knee joint images, shifted by a random number of pixels vertically and horizontally—the final accuracy was comparable to the results obtained over true areas.

Due to the high requirements for accuracy in the healthcare sector, the apply of fully automatic classification methods can be dangerous at this stage. However, the described method can be used to help the expert make a decision by providing additional information about the radiograph, such as the probability distribution of OA grades, the attention map, etc. (Fig. 3).

Fig. 3. Example of applying the Grad-CAM algorithm to the best DenseNet-121 and probability distribution for the presence of OA by the Kellgren-Lawrence scale.
There are several ways to improve the approach described in the work. First, it is to explore more complicated ensemble methods (e.g., super learner, stochastic weight averaging, etc.). Based on the results obtained, it can be concluded that the ensembles can significantly increase the final accuracy. Secondly, the study of training methods, which will reduce the classification error for KL grade 1, is seen as a perspective. And finally, it is important to decrease the false-negative rate of the network, which is especially important in the field of diagnostics of diseases.

6 Conclusion

In this paper, we applied the SSD model for knee joint localization and investigated the applicability of DenseNets and their ensembles to quantifying knee osteoarthritis from plain radiographs.

In the detection task the SSD model achieved the results comparable with the current state-of-the-art. The mean Jaccard index is 0.844, and the percentage of the images with $J > 0.75$ is 94.03%, which is better than with previous methods.

In the grading knee OA task DenseNets also outperformed previous approaches with average multiple-class accuracy 71.08% in an ensemble of three DenseNet-121, trained with different random seeds. DenseNet-161 as a whole did not show better results than DenseNet-121, which may be due to insufficient data for training. Also, the usage of ordinal loss did not lead to a significant increase in classification accuracy. However, after the improvements aimed at reducing the classification errors of neighboring grades, or errors for grade 1 (doubtful in KL scale), the prediction accuracy might significantly increase.

We also provide trained models and source codes for all analyzed models. All of this is publicly available on Github: https://github.com/almikh/grading-knee-oa-using-densenets.

References

1. Kellgren, J., Lawrence, J.: Radiological assessment of osteo-arthrosis. Ann. Rheum. Dis. 16, 494–502 (1957)
2. Schiphof, D., Boers, M., Bierma-Zeinstra, S.: Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. Ann. Rheum. Dis. 67, 1034–1036 (2008)
3. Altman, R., Gold, G.: Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthr. Cartil. 15(Suppl A), A1–A56 (2007)
4. Altman, R., Asch, E., Bloch, D., Bole, G., Borenstein, D., Brandt, K., et al.: Development of criteria for the classification and reporting of osteoarthritis: classification of osteoarthritis of the knee. Arthritis Rheum. 29, 1039–1049 (1986)
5. Culvenor, A., Engen, C., Engebretsen, L., Risberg, M.: Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Osteoarthr. Cartil. (22), 265 (2014)
6. Sheehy, L.: Validity and sensitivity to change of three scales for the radiographic assessment of knee osteoarthritis using images from the multicenter osteoarthritis study (MOST). Osteoarthr. Cartil. 23, 1491–1498 (2015)
7. Chan, S., Dittakan, K.: Osteoarthritis stages classification to human joint imagery using texture analysis: a comparative study on ten texture descriptors. In: Santosh, K.C., Hegadi, R.S. (eds.) RTIP2R 2018. CCIS, vol. 1036, pp. 209–225. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-9184-2_19
8. Minciullo, L., Cootes, T.: Fully automated shape analysis for detection of Osteoarthritis from lateral knee radiographs. In: 23rd International Conference on Pattern Recognition (ICPR), pp. 3787–3791 (2016)
9. Antony, J., McGuinness, K., Moran, K., O’Connor, N.: Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. arXiv:1703.09856 [cs.CV], 29 March 2017
10. Antony, J., McGuinness, K., Moran, K., O’Connor, N.: Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. arXiv:1609.02469 [cs.CV], 8 September 2016
11. Pingjun, C., Linlin, G., Xiaoshuang, S., Kyle, A., Lin, Y.: Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. Comput. Med. Imaging Graph. 75, 84–92 (2019). https://doi.org/10.1016/j.compmedimag.2019.06.002
12. Tiulpin, A., Thevenot, J., Rahtu, E., Saarakkala, S.: A novel method for automatic localization of joint area on knee plain radiographs. In: Sharma, P., Bianchi, F.M. (eds.) SCIA 2017. LNCS, vol. 10270, pp. 290–301. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59129-2_25
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
14. Tiulpin, A., Saarakkala, S.: Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. arXiv:1907.08020, Image and Video Processing (2019)
15. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., Saarakkala, S.: Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. arXiv:1710.10589 [cs.CV], 29 October 2017
16. Norman, B.D., Pedoia, V., Noworolski, A., Link, T.M., Majumdar, S.: Automatic knee Kellgren Lawrence grading with artificial intelligence. Osteoarthr. Cartil. 26, S436–S437 (2018)
17. Anifah, L., Purnama, I.K., Hariadi, M., Purnomo, M.H.: Osteoarthritis classification using self organizing map based on Gabor kernel and contrast-limited adaptive histogram equalization. Open Biomed. Eng. J. 7, 18–28 (2013). https://doi.org/10.2174/1874120701307010018
18. Wahyuningrum, R.T., Anifah, L., Purnama, I.K., Purnomo, M.H.: A new approach to classify knee osteoarthritis severity from radiographic images based on CNN-LSTM method. In: 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), pp. 1–6 (2019). https://doi.org/10.1109/ICAwST.2019.8923284
19. Norman, B.D., Pedoia, V., Noworolski, A., et al.: Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. J. Digit. Imaging 32, 471–477 (2019). https://doi.org/10.1007/s10278-018-0098-3
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
21. Liu, B., Luo, J., Huang, H.: Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN. Int. J. Comput. Assist. Radiol. Surg. 15, 457–466 (2020). https://doi.org/10.1007/s11548-019-02096-9
22. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

23. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., et al.: Going deeper with convolutions. arXiv:1409.4842 [cs.CV], 17 September 2014

24. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, Computer Vision and Pattern Recognition (2017)

25. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. arXiv:1608.06993 [cs.CV], 28 January 2018

26. Cheng, J., Aurélien, B., Mark, L.: The relative performance of ensemble methods with deep convolutional neural networks for image classification. J. Appl. Stat. 45, 2800–2818 (2018). https://doi.org/10.1080/02664763.2018.1441383

27. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 [cs.LG] (2014)

28. Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74