Computerized Medical Imaging and Graphics 75 (2019) 84–92

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss

Pingjun Chen[a], Linlin Gao[b], Xiaoshuang Shi[a], Kyle Allen[a], Lin Yang[a,*]

[a] J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL, United States
[b] Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China

ARTICLE INFO

Article history:
Received 12 September 2018
Received in revised form 23 May 2019
Accepted 12 June 2019

Keywords:
Knee osteoarthritis
Kellgren and Lawrence grading
Ordinal loss
Convolutional neural network

ABSTRACT

Knee osteoarthritis (OA) is one major cause of activity limitation and physical disability in older adults. Early detection and intervention can help slow down the OA degeneration. Physicians' grading based on visual inspection is subjective, varied across interpreters, and highly relied on their experience. In this paper, we successively apply two deep convolutional neural networks (CNN) to automatically measure the knee OA severity, as assessed by the Kellgren-Lawrence (KL) grading system. Firstly, considering the size of knee joints distributed in X-ray images with small variability, we detect knee joints using a customized one-stage YOLOv2 network. Secondly, we fine-tune the most popular CNN models, including variants of ResNet, VGG, and DenseNet as well as InceptionV3, to classify the detected knee joint images with a novel adjustable ordinal loss. To be specific, motivated by the ordinal nature of the knee KL grading task, we assign higher penalty to misclassification with larger distance between the predicted KL grade and the real KL grade. The baseline X-ray images from the Osteoarthritis Initiative (OAI) dataset are used for evaluation. On the knee joint detection, we achieve mean Jaccard index of 0.858 and recall of 92.2% under the Jaccard index threshold of 0.75. On the knee KL grading task, the fine-tuned VGG-19 model with the proposed ordinal loss obtains the best classification accuracy of 69.7% and mean absolute error (MAE) of 0.344. Both knee joint detection and knee KL grading achieve state-of-the-art performance. The code, dataset, and models are released at https://github.com/PingjunChen/KneeAnalysis.

Published by Elsevier Ltd.

1. Introduction

Knee osteoarthritis (OA) is the most common form of arthritis and the major cause of activity limitation and physical disability in older adults (Conaghan et al., 2015). More than half of the older Americans over 65 have radiological evidence of OA in at least one joint (Neogi, 2013). It is estimated that more than 20% of US residents will pass their 65th birthday by 2030 and be at risk for OA (Ortman et al., 2014). Pain and other symptoms caused by knee OA have severe effects on the quality of the elderly's life. Worse still, no treatment can inhibit the degenerative structural changes that are responsible for knee OA progression. However, early detection and treatment can help the elderly slow down OA progression and improve their quality of life. The hallmarks of knee OA include joint space narrowing (JSN), subchondral sclerosis, and osteophyte formation. MRI can reflect the 3D structures of knee joints. However, MRI is only available at large medical centers and the expensive exam makes MRI inappropriate for routine knee OA diagnosis. Because of the characteristics of safety, cost-efficiency, and wide accessibility, X-ray has been taken as the gold standard for the knee OA screening. The Kellgren and Lawrence (KL) grading system, accepted by WHO in 1961, is the most commonly used knee OA severity grading system (Kellgren and Lawrence, 1957). KL system splits knee OA severity into 5 grades from grade 0 to grade 4. The sample and criterion of each grade are shown in Fig. 1.

Physicians usually inspect a scanned knee X-ray image and then give KL grades to both knee joints in a very short time period. The diagnostic accuracy is highly relied on physicians' experience and carefulness. In addition, the criterion of KL grading is very ambiguous. For example, possible osteophytic lipping and doubtful JSN are used as the criterion for KL grade 1. Even the same physician may give different KL grades for the same knee joint when inspecting at different time points. The KL intra-rater reliability ranges from 0.67 to 0.73 in a study conducted by Culvenor et al. (2015). We suppose this low reliability of physicians' grading to be rooted in misclassifying the knee joint's KL grade to its nearby grades because of the ambiguous criterion. In clinical diagnosis, misclassifying the grade of a knee joint to its nearby grade (e.g., grade 1 to grade 2) is far less serious than misclassifying the grade to be faraway (e.g., grade 1 to grade 4). Therefore, it is insufficient to consider the grading accu-

[*] Corresponding author.
E-mail address: lin.yang@bme.ufl.edu (L. Yang).

https://doi.org/10.1016/j.compmedimag.2019.06.002
0895-6111/Published by Elsevier Ltd.

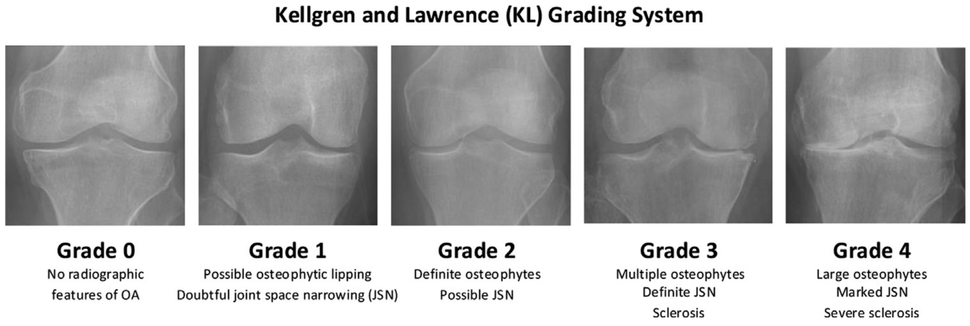


Fig. 1. Knee joint samples of all KL grades and their corresponding criterion.

racy as the only evaluation metric. To address this issue, we employ mean absolute error (MAE) to be another metric for knee KL grade classification evaluation, inspired by MAE used as the evaluation metric in age estimation (Niu et al., 2016).

Currently, there is an urgent need to accurately detect the presence and quantify the severity of knee OA because of its high prevalence. Fully automatic knee severity grading can provide objective, reproducible prediction, and won't have the fatigue problem after long hours of diagnosis. There are mainly two steps in predicting knee OA severity from the raw screened knee X-ray image: knee joint detection and classifying the detected knee joint into one of the five KL grades.

Several methods have been developed for the knee joint detection and KL grade classification in the last decade. In the knee joint detection, Shamir et al. (2009) propose templates matching technique with a sliding window strategy to locate the knee joints. Euclidean distances between 20 pre-defined knee joint images of size 15 × 15 and the current shifted window on the down-scaled knee X-ray image are computed. The window with the smallest Euclidean distance is taken as the detected knee joint. Also based on the sliding window strategy, Antony et al. (2016) apply linear support vector machine (SVM) to detect the knee joints using Sobel horizontal image gradients as features, motivated by the horizontal edges are contained in knee joint images. In Tiulpin et al.'s (2017) study, they first generate region proposals for knee joints. As they notice there is an intensity increase due to the presence of the patella, and then a sharp intensity drop because of the space between knee joints. Then linear SVM classifier is applied on these knee joint proposals using HoG features (Dalal and Triggs, 2005). Antony et al. (2017) later on propose a fully convolutional network (FCN) based approach to detect knee joints and achieve the state-of-the-art performance. However, FCN based model is mainly designed for semantic segmentation (Long et al., 2015). Specialized CNN detection architecture could be more appropriate for the knee joint detection task.

As to knee KL grade classification, early in 2009, Shamir et al. (2009) propose a weighted nearest neighbors algorithm using manual-crafted features including texture, Chebyshev statistics, haralick features, etc. In the past 3 years, Antony et al. (2016, 2017) attempt deep learning based methods on this task. In Antony et al. (2016), they formulate the classification of KL grades as a regression problem and use mean squared loss to fine-tune BVLC CaffeNet for knee KL grade classification. In Antony et al. (2017), they design a new CNN model and optimize a weighted combination of cross-entropy loss and mean squared loss. Tiulpin et al. (2018) present a new Deep Siamese CNN model to measure knee OA severity. In their method, random seeds are used to select different knee joint sub-regions and then fuse the predictions of selected regions.

With the rapid improvement of the computation ability and large available datasets, deep learning based methods achieve the state-of-the-art performance in numerous vision tasks, including image classification, object detection, and segmentation (Long et al., 2015; LeCun et al., 2015; Redmon et al., 2016; Roth et al., 2018). In the past 5 years, deep learning is widely used in medical image analysis (Litjens et al., 2017; Lu et al., 2017; Cai et al., 2016; Mahbod et al., 2019), like cell detection and segmentation (Su et al., 2015; Höfener et al., 2018), mitosis detection (Saha et al., 2018), white matter lesion segmentation (Manjón et al., 2018), and retinal blood vessel segmentation (Jiang et al., 2018). Deep learning based methods are also applied on knee OA analysis in several studies (Antony et al., 2016, 2017; Tiulpin et al., 2018). However, the performance of knee analysis still need to be improved. Considering the ordinal essence of the KL grading task, a better loss function can improve the knee KL grading performance.

In this paper, we successively apply two CNN models to automatically grade knee OA severity and achieve state-of-the-art performance. We first customize a one-stage detection architecture YOLOv2 (Redmon and Farhadi, 2019) to detect knee joints. Then a novel ordinal loss is proposed as the replacement for cross-entropy loss to fine-tune KL grade classification model. Most popular CNN models are tested on the knee KL grading task and experiments show that the VGG-19 model with the proposed ordinal loss obtains the best knee severity grading performance. The proposed ordinal loss demonstrates better performance than cross-entropy on all compared CNN models on knee KL grading. Fig. 2 shows the pipeline of knee severity grading on the knee X-ray image.

2. Materials and methods

2.1. Dataset and preprocessing

Knee X-ray images used for evaluation are obtained from the osteoarthritis initiative[1] (OAI), which is a multi-center, longitudinal, prospective observational study of knee osteoarthritis (OA) aiming to identify biomarkers for OA onset and progression (Nevitt et al., 2006). There are 4796 participants with age ranging from 45 to 79 enrolled in this program. In our study, we use knee bilateral PA fixed flexion X-ray images from the baseline cohort to evaluate our proposed method.

As OAI is a multi-center study, the physical resolution and dimension of these knee X-ray images collected from baseline cohort are not consistent. Preprocessing is needed before the knee OA quantification. Firstly, all raw screened X-ray images are resized to own the same physical resolution. The value is chose to be close to the median of all images' physical resolution, and 0.14 mm/pixel is used. To ensure all processed images to have the same size, we then crop the central region from the resized image with the height of 2048 pixels and the width of 2560 pixels. As we would predict KL grades of both left and right knee joints from one X-ray image,

[1] https://oai.epi-ucsf.org

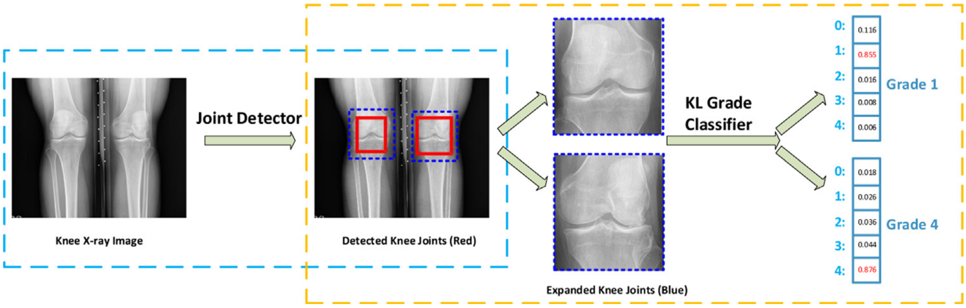


Fig. 2. The pipeline of knee joint severity grading, which includes knee joint detection and knee KL grade classification. Detected knee joints (red) are expanded by a certain ratio (1.3 used) to cover broad knee joint region (dashed blue) for KL grading. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we only keep those X-ray images with available KL grades on both knee joints.

After the preprocessing and filtering, 4130 X-ray images with 8260 knee joints are remained. We randomly split all knee X-ray images into training, validation, and testing sets with a ratio of 7 : 1 :2. This splitting is conducted in a grade-wise manner according to the KL grade of the left knee joint in knee X-ray image to ensure relatively stable grade distribution among training, validation, and testing sets. After the splitting, the knee joint testing set contains 828 X-ray images and 1656 knee joints, in which 639 knee joints of grade 0, 296 of grade 1, 447 of grade 2, 223 of grade 3, and 51 of grade 4.

Since there is no ground truth for knee joint detection, we implement a labeling software to manually annotate knee joints under the guidance of physicians. As we plan to utilize this annotation for joint space segmentation to measure joint space width in our future study, we annotate the knee bone region in a strict manner, which mainly covers the inner part of knee joint (e.g. the red bounding box in Fig. 2). We would expand the annotated knee joint by a certain ratio (1.3 used) for the knee KL grading (e.g. the dashed blue bounding box in Fig. 2). In this manner, the annotated knee joints can be used for both knee joint detection and knee KL grade classification.

2.2. *Knee joint detection*

There are several general CNN detection architectures available, including R-CNN series (Girshick et al., 2014; Ren et al., 2015, 2017), YOLO series (Redmon et al., 2016; Redmon and Farhadi, 2019), SSD (Liu et al., 2016), etc. For natural object detection, different scales and aspect ratios are key challenges. Faster R-CNN would first generate 2000 region proposals of different aspect ratios using Region Proposal Network (RPN). In SSD, default bounding boxes with different aspect ratios at each location of feature map in different convolution layers are used as object proposals. When detecting knee joints of OA participants, their knee size is much less varied because of the following two reasons. Firstly, all knee X-ray images are preprocessed to have the same physical resolutions. Secondly, the participants are mostly the elderly. Although there are differences in people's physique, the difference of knee joint size is limited.

Here we consider customizing YOLOv2 for knee joint detection by setting the initial knee joint size to be close to the real knee joint size. The default initial knee size can be obtained through clustering on all available training knee joints. In the YOLOv2 detection architecture, object detection is formulated as a regression problem that refines the height, width, center coordinates, and confidence score for each of the preset bounding boxes located in all divided grid centers. The regression becomes easier to fit when the initial bounding box is close to the real bounding box. YOLOv2 is a one-stage detector that can be optimized in an end-to-end manner and does not require explicit region proposal generation. The grid-by-grid way to produce knee joint bounding box proposals, in essence, is the same as the sliding window strategy. Most possible locations of the image are checked to locate knee joints. The advantage of CNN network is that the features of all proposals are calculated all at once in one forward operation.

YOLOv2 improves YOLO in various aspects. In this knee joint detection, we adopt the following improvements: (1) Batch normalization is added to regularize the model and assist the convergence in training. (2) New direct location prediction, which restrains the object center to be located inside the preset grid cells, is applied to stabilize the detection. (3) Better bounding box initialization is used through the K-means clustering. To note that we use YOLOv2 only for detection and get rid of the classification part. Integrated knee joint detection and KL grading is a more elegant framework compared with separately training knee detection and KL grade classification models. However, experiments show worse detection performance and much lower classification accuracy when integrated training. In medical applications, like knee analysis in this study, we give priority to model's performance rather than model's elegance.

2.3. *Adjustable ordinal loss*

The essence of KL grade prediction is an ordinal regression problem. Because confusing two distant grades (e.g., predicting grade 0 to grade 4) is more serious than confusing two close grades (e.g., predicting grade 0 to grade 1). The cross-entropy loss, as the default loss in deep learning based classification model for object classification, treats all categories equally. It doesn't consider the closeness between different categories. The softmax layer in CNN classification model would output probabilities $[p_0, p_1, \ldots, p_{n-1}]^T$ for n categories, with $\sum_{i=0}^{n-1} p_i = 1.0$. In ordinal classification task like the knee KL grading, given an image with grade m, we expect the output probability distribution to satisfy following two properties: (1) p_m is as close to 1.0 as possible; (2) For $k \in \{0, \ldots, n-1\} \setminus \{m\}$, if $|k-m|$ is larger, then p_k should be even smaller. The cross-entropy only satisfies the first property. We propose a new ordinal loss to meet these two properties.

We first devise an adjustable ordinal matrix W to denote the penalty weights between the predicted grade and the real grade. W is a $n*n$ square matrix, where $w_{i,j} \in W$ stands for the penalty weight of predicting grade j to grade i, here $i, j \in \{0, 1, \ldots, n-1\}$ and $n = 5$ in the knee KL grading task. The matrix on bottom left of Fig. 3 is an example of the adjustable ordinal matrix. In this design, $w_{:,m}$ stands for the penalty vector of grade m. We set the penalty weight of each grade to itself fixed as 1 and to the rest grades to be higher if the rest grades are faraway. Based on this ordinal matrix and the

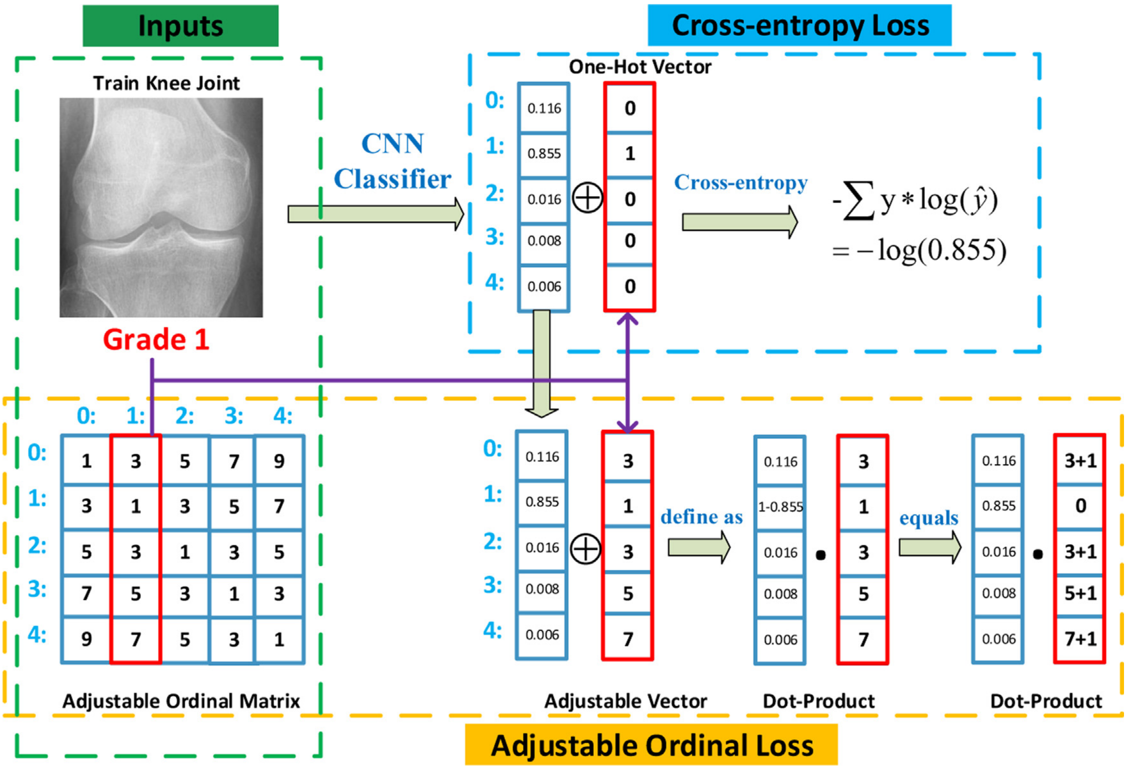


Fig. 3. Demo of the proposed ordinal loss calculation process and a comparison with the cross-entropy loss. The upper right part shows the cross-entropy loss optimization, which tries to push the probability of ground truth category to 1.0, but without considering the rest categories. The bottom right shows the proposed ordinal loss calculation. It tries to push the probability of ground truth to be 1.0 and the probabilities of faraway grades to be even smaller using larger penalty weight.

output probabilities by the softmax layer, the proposed ordinal loss is defined as:

$$loss = \sum_{i=0}^{n-1} w_{i,m} \cdot q_i, \tag{1}$$

where m is the true KL grade of the input image; $q_i = p_i$ if $i \neq m$, otherwise $q_i = 1 - p_i$. Minimizing loss in Eq. (1) requires p_m to be close to 1.0 and faraway grade to have even smaller probability due to its corresponding high penalty weight.

To simplify the implementation, we revise the adjusted ordinal matrix $\bar{W}$ to rewrite the loss, in which $\bar{w}_{i,j} = 0$ if $i = j$, otherwise $\bar{w}_{i,j} = w_{i,j} + 1$. As $\sum_{i=0}^{n-1} p_i = 1.0$ and $w_{m,m}$ is fixed to be 1, the loss in Eq. (1) is equal to the following form:

$$loss = \sum_{i=0}^{n-1} \bar{w}_{i,m} \cdot p_i, \tag{2}$$

Here in Eq. (2), probability output from the softmax layer can be used directly to calculate the proposed loss. Fig. 3 shows a demo on the computation process of the proposed ordinal loss and a comparison with the cross-entropy loss. As non-diagonal entry in the penalty matrix W can be adjusted. We name the penalty matrix as adjustable ordinal matrix and the loss as adjustable ordinal loss. The square of the proposed ordinal loss is used in the actual CNN classifier fine-tuning process because of its better performance.

2.4. CNN classifiers in KL grading

There are several popular CNN classification architectures along the deep learning development in the past 5 years. ResNet (He et al., 2016) is demonstrated to be easier to optimize and shows better generalization performance on recognition tasks. VGG networks (Simonyan and Zisserman, 2019) are very simple and elegant, with multiple 3 × 3 kernel-sized filters one after another. InceptionV3 (Szegedy et al., 2016) is widely used in medical classification tasks, including skin cancer (Esteva et al., 2017), diabetic retinopathy detection (Gulshan et al., 2016), etc. The latest DenseNet (Huang et al., 2017) is designed to strengthen feature propagation and to encourage feature reuse. DenseNet obtains better classification performance over previous network structures on four highly competitive object recognition benchmark tasks (CIFAR-10, CIFAR-100, SVHN, and ImageNet).

In this study, we fine-tune all these popular CNN classification networks on the knee KL grading task to find the best CNN model for the knee KL grading. Moreover, we would compare the performance of the proposed ordinal loss with the cross-entropy loss on multiple CNN models with different architectures to test the proposed loss's generalization ability.

3. Results and discussion

3.1. Experimental setting

3.1.1. YOLOv2 implementation details

After preprocessing, the size of a knee X-ray image is 2048 × 2560, which is too large for YOLOv2. We resize all these images to be of size 256 × 320 as the input for knee joint detection. The annotated bounding boxes are resized accordingly. Instead of directly feeding the gray X-ray image into the CNN model, we first normalize the gray image by subtracting the mean and dividing by the standard deviation. The mean and standard deviation are calculated from training images. Then we concatenate three same normalized gray knee X-ray images to a three-channel image and take it as the model's input. K-means is applied to cluster knee bounding boxes (also called anchor box) for the knee detection initialization. The number of bounding boxes from 1 to 6 is tested in

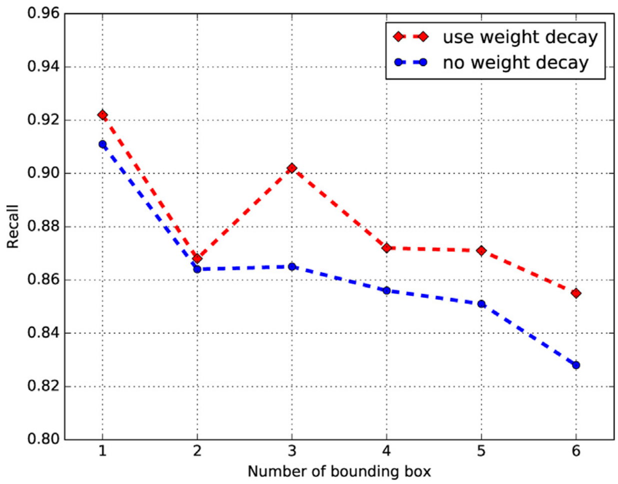


Fig. 4. The performance of knee joint detection using YOLOv2 with different number of bounding boxes and weight decay.

the experiments. We also compare the effects of using weight decay or not.

In the testing stage, knee X-ray image is resized, normalized, and then concatenated, following the same preprocessing procedure as training. After the model's forward pass on the input image, we first remove those bounding boxes with confidence score smaller than 0.12, and then apply non-maximum suppression (NMS) with an overlap ratio of 0.7 to prune the remaining ones. The bounding box with a higher confidence score is taken as the final detection result. At last, the detected bounding box is mapped back to higher resolution for detection evaluation (2048×2560) and knee joint cropping for the following knee KL grading.

3.1.2. CNN classifiers fine-tuning details

The knee joint images used in CNN classifier fine-tuning for the KL grading are cropped from annotated bounding box for knee joint detection task with an expanding ratio of 1.3. Mean and standard deviation of training knee joint images are calculated for normalization. Cropped images are resized to 299×299 for InceptionV3 and to 224×224 for ResNet, VGG, and DenseNet. They are normalized and then are concatenated to be three-channel images as inputs for CNN classifier training and testing. The adjustable ordinal matrix used is the demo matrix as shown in the bottom left of Fig. 3. We compare the proposed ordinal loss with the cross-entropy loss on all fine-tuned CNN classifiers.

3.2. Knee joint detection evaluation

Intersection over Union (IoU) between detection results and manual annotations, also known as the Jaccard index, is used as the fundamental metric to evaluate knee detection performance, which is also adopted in the studies of Antony et al. (2016) and Tiulpin et al. (2017). Based on IoU, following metrics are used to evaluate knee joint detection performance: knee joint recall with IoU threshold of 0.75 and mean IoU of detected knee joints.

Fig. 4 compares the performance of YOLOv2 on knee joint detection with different number of anchor boxes as well as the usage of weight decay (5.0e−4 used as the coefficient). In contrast to our expectation that more bounding boxes would gain better result, the best detection recall is obtained when the number of bounding box is 1. The recall performance would decrease as more bounding boxes is used. This should be caused by the less variance of the knee joint size among different images. The usage of weight decay can improve the detection performance when other settings are fixed.

Table 1
Comparison of the knee joint detection methods on the OAI dataset. Fine-tuned YOLOv2 model achieves the best performance on three main knee detection evaluation metrics.

Methods	Recall$_{IoU \geq 0.75}$	Mean IoU	Time [s]
FCN (Antony et al., 2016)	0.892	0.830	–
HOG-SVM (Tiulpin et al., 2017)	–	0.840	0.169
Proposed	**0.922**	**0.858**	**0.0105**

Table 1 shows the comparison results between the proposed method and two start-of-the-art methods, namely FCN based method (Antony et al., 2016) and SVM based method (Tiulpin et al., 2017). The customized YOLOv2 achieves the best performance in terms of all used metrics. In the testing set, all 1656 knees are detected with no false positive and false negative. 1526 (92.2%) knee joints have the IoU value larger than 0.75 and 1654 (99.9%) knee joints have the IoU value larger than 0.50.

Fig. 5 contains some examples of knee detection results. The knee images with high contrast in column (A) show good detection performance. The detected bounding box of knee joint is very close to the ground truth when the IoU is larger than 0.90. Column (B) shows three examples of knee images with detection IoU close to 0.75, upon which the detected bounding box is moderately consistent with ground truth. Considering that we would expand the detected knee joint by a ratio of 1.3 to cover broad knee joint region, detection IoU of 0.75 should be good enough for the following knee KL grading. Column (C) presents two knee joints with detection IoU smaller than 0.5. The low detection IoU value in image (7) may be caused by the low contrast that leading to a much broad knee joint region to be detected. Low detection IoU of the left knee joint in image (8) can be caused by the implant. From these detection examples, we can see that knee joint detection using YOLOv2 achieves very promising results.

3.3. Knee KL grade classification evaluation

As suggested in the preprocessing part, knee joint images for KL grading are cropped based on the detected bounding box with certain expansion from testing knee X-ray images. We evaluate the KL grade classification on both manually annotated and automatically detected knee joints. Two metrics, including classification accuracy and MAE, are used to evaluate classification results. The best model selected in the training is based on the classification accuracy on the validation set.

Table 2 shows the classification performance of all fine-tuned models with the cross-entropy loss and the proposed ordinal loss based on the ordinal matrix as shown in Fig. 3 on manually and automatically detected knee joints. Among all compared CNN architectures on knee KL grading task, VGG classifiers obtain better performance than the rest, with VGG-19 achieving the best accuracy and VGG-16 achieving the smallest MAE on manually annotated knee joints. InceptionV3 achieves competitive performance with VGG classifiers and is superior to the variants of DenseNet and ResNet. InceptionV3's performance validates its broad usage in medical applications. ResNet variants show the worst performance among all compared models. The performance of VGG and ResNet does not conform to our expectation. Considering their's performance on large-scale image recognition, we expect that ResNet to be the most favorable CNN model to attain the best performance and VGG to be the least to obtain good accuracy. As there is lack of understanding on the underlying mathematical principle of deep neural networks, the behaviors of these networks on the knee KL grading suggest that the performance of CNN classification models is highly dependent on the recognition task, the classification may not get the optimal accuracy if assuming certain network architecture, like ResNet-34, to be the best candidate.

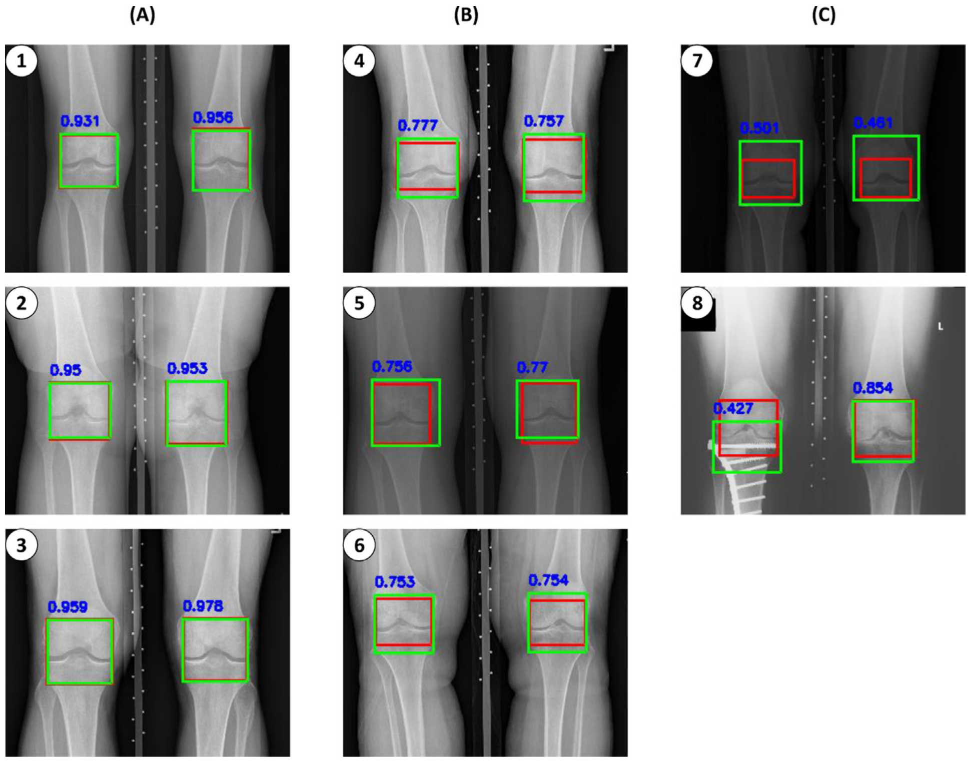


Fig. 5. Knee detection examples. Red bounding box stands for the annotated knee joint region and green bounding box stands for the detected knee joint region. The blue score over the bounding box stands for the IoU value between the detected knee joint and the annotated knee joint. Column (A) includes 3 knee X-ray images obtaining good detection results with IoU larger than 0.90. Column (B) includes example images of IoU value close to 0.75. Column (C) includes two knee joints with IoU smaller than 0.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison of different classifiers on the knee KL grading on both manually and automatically detected knee joints. The main variants of ResNet, VGG, DensetNet and Inception are compared with both cross entropy loss and our proposed ordinal loss. VGG-19 with proposed ordinal loss achieves the best performance among all compared models on the knee KL grading.

	Methods	Knee joints	Accuracy	MAE	Knee joints	Accuracy	MAE
ResNet	Antony et al. (2017)	Manual	63.6%	0.503	Auto	61.9%	0.541
	ResNet-18-CE		65.8%	0.441		65.7%	0.448
	ResNet-18-Ordinal		66.7%	0.423		65.7%	0.442
	ResNet-34-CE		66.2%	0.426		66.5%	0.420
	ResNet-34-Ordinal		67.8%	0.402		67.6%	0.403
	ResNet-50-CE		64.6%	0.437		64.3%	0.437
	ResNet-50-Ordinal		66.2%	0.403		65.8%	0.408
	ResNet-101-CE		64.9%	0.425		65.2%	0.431
	ResNet-101-Ordinal		65.5%	0.408		66.7%	0.391
	ResNet-152-CE		66.2%	0.441		66.6%	0.429
	ResNet-152-Ordinal		67.7%	0.390		65.7%	0.422
VGG	VGG-16-CE		67.1%	0.413		67.4%	0.414
	VGG-16-Ordinal		68.5%	**0.356**		69.1%	**0.358**
	VGG-16bn-CE		68.1%	0.386		68.5%	0.391
	VGG-16bn-Ordinal		68.4%	0.376		67.5%	0.389
	VGG-19-CE		69.3%	0.370		69.2%	0.370
	VGG-19-Ordinal		**69.6%**	0.360		**70.4%**	**0.358**
	VGG-19bn-CE		67.1%	0.403		66.6%	0.416
	VGG-19bn-Ordinal		68.3%	0.369		67.8%	0.370
DenseNet	DenseNet-121-CE		67.3%	0.400		67.4%	0.402
	DenseNet-121-Ordinal		68.2%	0.381		67.8%	0.389
	DenseNet-169-CE		66.8%	0.395		66.6%	0.399
	DenseNet-169-Ordinal		67.1%	0.388		65.4%	0.401
	DenseNet-201-CE		65.7%	0.410		66.3%	0.411
	DenseNet-201-Ordinal		67.3%	0.389		66.3%	0.395
Inception	InceptionV3-CE		68.1%	0.400		67.7%	0.406
	InceptionV3-Ordinal		68.4%	0.366		66.6%	0.385

Note: *CE* stands for cross-entropy and *Ordinal* stands for ordinal loss.

Comparing the performance between using the cross-entropy loss and the proposed ordinal loss, the ordinal loss obtains higher accuracy and lower MAE in manually cropped knee joints in all compared classifiers. In automatically detected knee joints, most classifiers (except ResNet-152, VGG-16bn, InceptionV3) get higher accuracy. But all classifiers obtain lower MAE. These results demon-

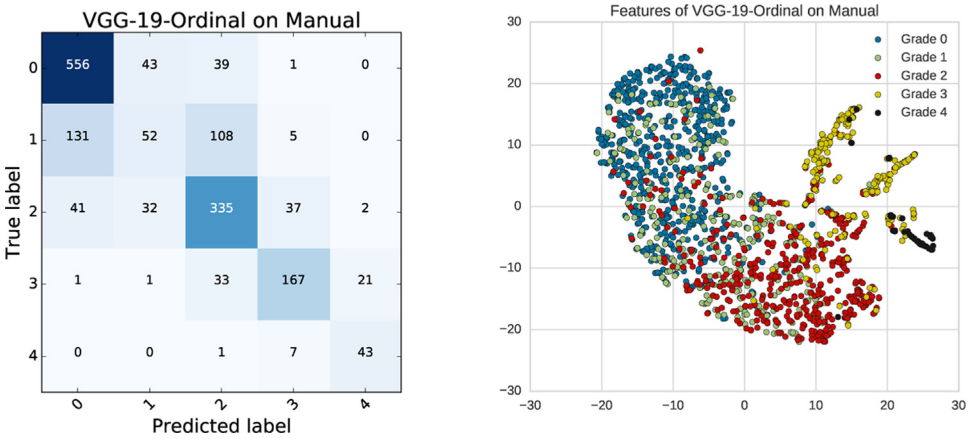


Fig. 6. The left figure shows the confusion matrix of VGG-19-Ordinal model on manually cropped knee joints. The right figure shows the t-SNE dimension reduction on features from the penultimate layer of VGG-19-Oridinal model on manually cropped knee joints.

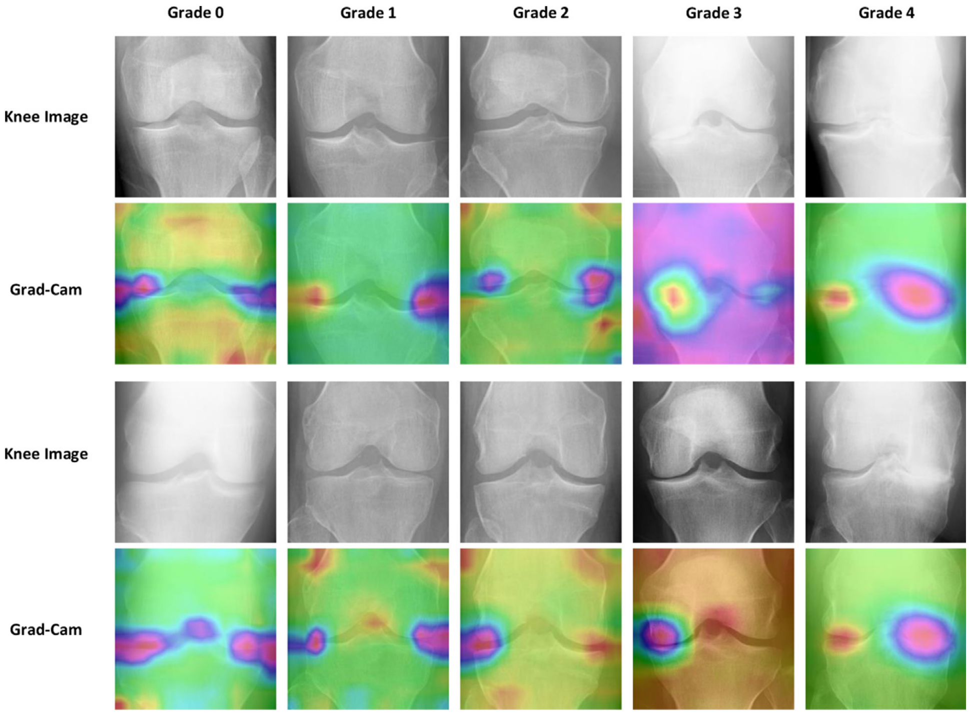


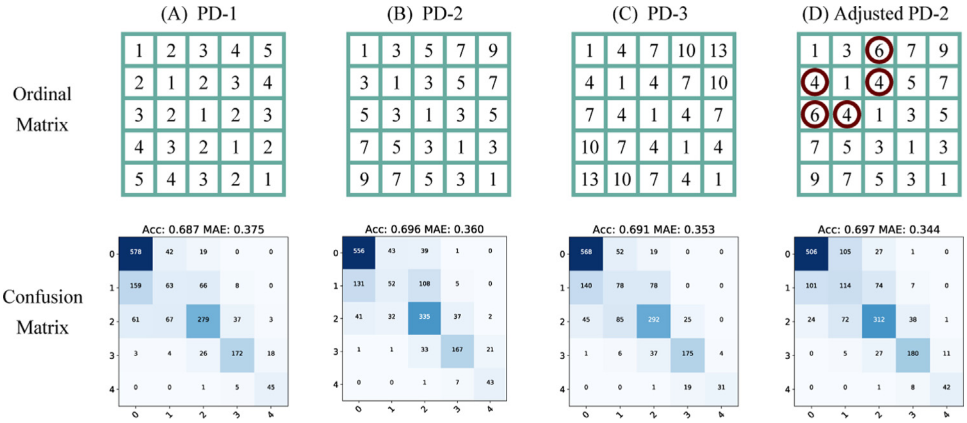
Fig. 7. Localizations achieved by Grad-CAM technique based on the VGG-19 model on testing knee joint images. From left to right are knee joints with grade 0 to 4. The first and third row are examples of knees images and the second and fourth row are their corresponding overlaid classification activation maps. The hallmarks of knee OA, including joint space narrowing (JSN), subchondral sclerosis, and osteophyte formation, can be accurately localized.

strate the superior effects of the proposed ordinal loss in the knee KL grading task. Comparing the accuracy and MAE between manually cropped knee joints and automatically detected knee joints, their results are very close. Some classifiers, like ResNet-101 and VGG-16, even obtain higher accuracy and lower MAE on automatically detected knee joints. The results certify the usability of automatically detected knee joints for KL grade classification.

Fig. 6 displays the confusion matrix of classification and dimension reduction results using t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008) on penultimate layer features of VGG-19-Ordinal model on manually cropped knee joints. Among five KL grades, the grade 1 has the lowest accuracy with lots of samples predicted to grade 0 and grade 2. The accuracy of the rest grades are much higher. According to the t-SNE mapping of features extracted from testing samples, samples of grade 0/1/2 are mixed together, which validates why it's hard to accurately classify grade 1. One reason may lie in the ambiguous criterion of grade 1, which can be easily misclassified to be either grade 0 or grade 2. There is trend of gradual change of grade 0 to 4 from left to right in the t-SNE mapping results, which also illustrate the ordinal nature of the knee KL grading.

Table 2 shows that all fine-tuned models achieve better performance than the method used by Antony et al. Tiulpin's method, which uses average multiple-class accuracy as the evaluation metric and achieves accuracy of 66.71%. Based on the confusion matrix in Fig. 6, VGG-19-Oridinal achieves average multiple-class accuracy of 67.70%, which is superior to that of Tiulpin's method.

We apply Gradient-weighted Class Activation Mapping (Grad-CAM) technique (Ramprasaath et al., 2016) on VGG-19 model to locate the most significant areas in the image for classification. Fig. 7 shows two knee images of each KL grade. The most critical regions for KL grading are distributed around the knee joint space. The hall-



(A) PD-1

1	2	3	4	5
2	1	2	3	4
3	2	1	2	3
4	3	2	1	2
5	4	3	2	1

Acc: 0.687 MAE: 0.375

	0	1	2	3	4
0	578	42	19	0	0
1	159	63	66	8	0
2	61	67	279	37	3
3	3	4	26	172	18
4	0	0	1	5	45

(B) PD-2

1	3	5	7	9
3	1	3	5	7
5	3	1	3	5
7	5	3	1	3
9	7	5	3	1

Acc: 0.696 MAE: 0.360

	0	1	2	3	4
0	556	43	39	1	0
1	131	52	108	5	0
2	41	32	335	37	2
3	1	1	33	167	21
4	0	0	1	7	43

(C) PD-3

1	4	7	10	13
4	1	4	7	10
7	4	1	4	7
10	7	4	1	4
13	10	7	4	1

Acc: 0.691 MAE: 0.353

	0	1	2	3	4
0	568	52	19	0	0
1	140	78	78	0	0
2	45	85	292	25	0
3	1	6	37	175	4
4	0	0	1	19	31

(D) Adjusted PD-2

1	3	6	7	9
4	1	4	5	7
6	4	1	3	5
7	5	3	1	3
9	7	5	3	1

Acc: 0.697 MAE: 0.344

	0	1	2	3	4
0	506	105	27	1	0
1	101	114	74	7	0
2	24	72	312	38	1
3	0	5	27	180	11
4	0	0	1	8	42

Fig. 8. Comparing the KL grading performance using different ordinal matrices. The first row contains the four compared ordinal matrices and the second row shows the corresponding KL grading confusion matrix. Here PD denotes penalty distance.

marks of knee OA also are located in joint space. The features found by CNN through learning from large samples are consistent with the grading criterion. The JSN and subchondral sclerosis are accurately localized by the CNN model in two examples of grade 3 and 4 in the first row.

Fig. 8 shows the confusion matrices on the test dataset of manually detected knee joints under different ordinal matrix settings. Here we compare four ordinal matrices in the experiment. Penalty distance is abbreviated as PD and PD-1/2/3 denote that the loss penalty weight between neighbor grade is 1, 2, and 3, respectively. PD-2 is the ordinal matrix used to compare the performance of different CNN models. As we can see Fig. 8, PD-2 achieves the best accuracy among PD-1/2/3, PD-3 gets the best mean absolute error (MAE), and PD-1 attains the worst MAE. These results demonstrate that selecting proper penalty distance can obtain better classification accuracy, and large penalty distance can reduce MAE.

Based on the grading confusion matrix obtained by applying PD-2, we adjust some entries in PD-2, namely, those circled entries of Fig. 8(D), and obtain the adjusted PD-2. We obtain slightly better accuracy under this setting, while much better MAE. What's more, comparing the corresponding circled items in the confusion matrix, most of the misclassification in these items is reduced. The result obtained by the adjusted PD-2 further validates the capability of the proposed ordinal loss on boosting grading performance.

Although the adjusted PD-2 can achieve better performance than the other three ordinal matrices, the model's performance can be further improved if we spend more time and computation on looking for a better ordinal matrix. To note that, there is no restriction on the value of the entries in ordinal matrix to be integers. The penalty vector can also be set as [1, 3.2, 5.6, 7.1, 8.8] for grade 0, provided that the classification performance can be further improved.

Here we empirically provide a general way to choose an optimal or sub-optimal ordinal matrix. Firstly, we can investigate using ordinal matrices with fixed penalty distance, e.g., PD-1/2/3. Then based on the best-performed candidate, adjusting a few entries inside the matrix to obtain a better one. Whereas it would be best if all entries in the ordinal matrix can be automatically learned during the training process. Therefore, further research will investigate how to automatically learn the ordinal matrix in the proposed ordinal loss function.

4. Conclusion

In this paper, we apply a customized YOLOv2 model for the knee joint detection and fine-tune CNN models with a novel ordinal loss for knee KL grading. State-of-the-art performance are achieved on both knee joint detection and knee KL grading. Based on YOLOv2's performance on the knee joint detection, one-stage detector YOLOv2 is well fitting to detection tasks with less varied object size. The proposed ordinal loss improves the classification accuracy and reduces the MAE between prediction and ground truth compared with using the cross-entropy among all popular CNN classification models on the knee KL grading task, suggesting its potential in ordinal classification tasks. Compared to the variants of ResNet or DenseNet, the fine-tuned VGG-19 model achieves the best classification performance, validating the performance of CNN models highly dependent on the recognition task.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by National Institutes of Health R01 AR065479-02.

References

Antony, J., McGuinness, K., O'Connor, N.E., Moran, K., 2016. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. International Conference on Pattern Recognition, 1195–1200.

Antony, J., McGuinness, K., Moran, K., O'Connor, N.E., 2017. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. Internaltional Conference on Machine Learning and Data Mining in pattern recognition, 376–390.

Cai, J., Lu, L., Zhang, Z., Xing, F., Yang, L., Yin, Q., 2016. Pancreas segmentation in MRI using graph-based decision fusion on convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 442–450.

Conaghan, P.G., Porcheret, M., Kingsbury, S.R., Gammon, A., Soni, A., Hurley, M., Rayman, M.P., Barlow, J., Hull, R.G., Cumming, J., et al., 2015. Impact and therapy of osteoarthritis: the arthritis care oa nation 2012 survey. Clin. Rheumatol. 34 (9), 1581–1588.

Culvenor, A.G., Engen, C.N., Øiestad, B.E., Engebretsen, L., Risberg, M.A., 2015. Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Knee Surg. Sports Traumatol. Arthrosc. 23 (12), 3532–3539.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 886–893.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580–587.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316 (22), 2402–2410.

H&rdquo;ofener, H., Homeyer, A., Weiss, N., Molin, J., Lundstr&rdquo;om, C.F., Hahn, H.K., 2018. Deep learning nuclei detection: a simple approach can deliver state-of-the-art results. Comput. Med. Imaging Graph. 70, 43–52.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L., 2017. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 3.

Jiang, Z., Zhang, H., Wang, Y., Ko, S.-B., 2018. Retinal blood vessel segmentation using fully convolutional network with transfer learning. Comput. Med. Imaging Graph. 68, 1–15.

Kellgren, J., Lawrence, J., 1957. Radiological assessment of osteo-arthrosis. Ann. Rheum. Dis. 16 (4), 494.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: single shot multibox detector. European Conference on Computer Vision, 21–37.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440.

Lu, L., Zheng, Y., Carneiro, G., Yang, L., 2017. Deep Learning and Convolutional Neural Networks for Medical Image Computing. Springer.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Pitiot, A., Wang, C., 2019. Fusing fine-tuned deep features for skin lesion classification. Comput. Med. Imaging Graph. 71, 19–29.

Manjón, J.V., Coupé, P., Raniga, P., Xia, Y., Desmond, P., Fripp, J., Salvado, O., 2018. MRI white matter lesion segmentation using an ensemble of neural networks and overcomplete patch-based voting. Comput. Med. Imaging Graph. 69, 43–51.

Neogi, T., 2013. The epidemiology and impact of pain in osteoarthritis. Osteoarthr. Cartil. 21 (9), 1145–1153.

Nevitt, M., Felson, D., Lester, G., 2006. The Osteoarthritis Initiative: Protocol for the Cohort Study 1.

Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G., 2016. Ordinal regression with multiple output CNN for age estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4920–4928.

Ortman, J.M., Velkoff, V.A., Hogan, H., et al., 2014. An Aging Nation: The Older Population in the United States.

Ramprasaath, R., Abhishek, D., Ramakrishna, V., Michael, C., Devi, P., Dhruv, B., 2016. Grad-CAM: why did you say that? Visual explanations from deep networks via gradient-based localization. CVPR.

Redmon, J., Farhadi, A., 2019. Yolo9000: better, faster, stronger, arXiv preprint 1612.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 91–99.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Roth, H.R., Oda, H., Zhou, X., Shimizu, N., Yang, Y., Hayashi, Y., Oda, M., Fujiwara, M., Misawa, K., Mori, K., 2018. An application of cascaded 3D fully convolutional networks for medical image segmentation. Comput. Med. Imaging Graph. 66, 90–99.

Saha, M., Chakraborty, C., Racoceanu, D., 2018. Efficient deep learning model for mitosis detection using breast histopathology images. Comput. Med. Imaging Graph. 64, 29–40.

Shamir, L., Ling, S.M., Scott Jr., W.W., Bos, A., Orlov, N., Macura, T.J., Eckley, D.M., Ferrucci, L., Goldberg, I.G., 2009. Knee X-ray image analysis method for automated detection of osteoarthritis. IEEE Trans. Biomed. Eng. 56 (2), 407–415.

Simonyan, K., Zisserman, A., 2019. Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

Su, H., Xing, F., Kong, X., Xie, Y., Zhang, S., Yang, L., 2015. Robust cell detection and segmentation in histopathological images using sparse reconstruction and stacked denoising autoencoders. International Conference on Medical Image Computing and Computer-Assisted Intervention, 383–390.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826.

Tiulpin, A., Thevenot, J., Rahtu, E., Saarakkala, S., 2017. A novel method for automatic localization of joint area on knee plain radiographs. Scandinavian Conference on Image Analysis, 290–301.

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