Automatically Quantifying Radiographic Knee Osteoarthritis Severity

Final Report - CS 229 - Machine Learning

Suhas Suresha SUHAS17@STANFORD.EDU
M.S. Student in Institute for Computational and Mathematical Engineering (ICME)

Akash Mahajan AKASHMJN@STANFORD.EDU
M.S. Student in Management Science and Engineering (MS&E)

Nathan Dalal NATHANHD@STANFORD.EDU
B.S. Student in Computer Science (CS)

Abstract

In this paper, we implement machine learning algorithms to automatically quantify knee osteoarthritis severity from X-ray images according to the Kellgren & Lawrence (KL) grades. We implement and evaluate the performance of various machine learning models like transfer learning, support vector machines and fully connected neural networks based on their classification accuracy. We also implement the task of automatically extracting the knee-joint region from the X-ray images and quantifying their severity by training a faster region convolutional neural network (R-CNN).

1. Introduction and Motivation

Knee osteoarthritis (OA) is the most common cause of limited mobility in adults. One in two individuals are at a risk of developing knee OA by age 85. Current treatments for knee OA are limited to symptoms management. Doctors have observed that radiographic (X-ray) and symptomatic OA status do not correlate. Hence identifying knee OA severity from X-ray images and MRI scans is a hot topic in the osteoarthritis research community.

In the current report, we attempt to identify and classify OA severity from X-ray images. We tackle the problem using various machine learning techniques such as support vector machines (SVM), transfer learning and region convolutional neural networks (R-CNN) and compare the results.

The ultimate goal is to identify features in the radiographic images that directly correlate with pain. Identifying such features would have great clinical utility as they could be used as targets for drug development. We believe that solving the OA classification problem is a good first step towards achieving the ultimate goal.

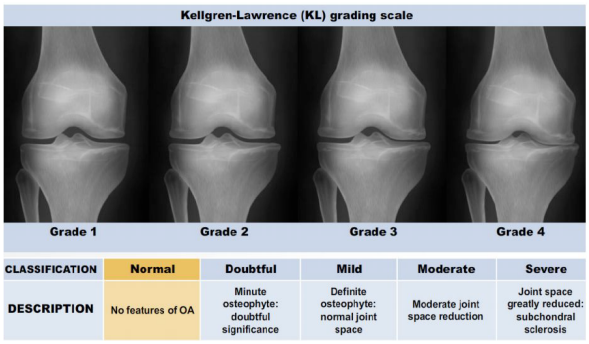

Figure 1. Kellgren and Lawrence (KL) grading system

2. Osteoarthritis Staging

Osteoarthritis staging is done according to the Kellgren and Lawrence (KL) grading system, which falls on a five-point scale (0 to 5). Grade 0 indicates no knee OA condition. Grade 4 indicates severe knee OA condition and the intermediate grades indicate the progressive stages. The grading system is based on joint space narrowing, presence of osteophytes, sclerosis and bone deformity. The K-L scoring system is explained in more detail in figure 1.

3. Objectives

Our objective is to automatically quantify knee osteoarthritis (OA) severity from X-ray images, labeled using the Kellgren and Lawrence (KL) grading system. In order to achieve our objective, we explore various machine learning techniques like support vector machines (SVM), trans-

fer learning and region convolutional neural networks (R-CNN). We analyze and compare the results from the various techniques and identify the model giving the best classification accuracy.

4. Related Work

Previous work on classifying OA severity based on the KL score have used a multipurpose bio-medical image classifier known as Wndchrm. Wndchrm uses hand-crafted features based on characteristics like polynomial decomposition, contrast, pixel statistics, textures and features extracted from image transforms to perform classification. Such techniques failed to achieve a classification accuracy greater than 29 %.

In a more recent arXiv paper by Antony et al., they apply transfer learning and deep learning models to manually extracted knee-joint portion of the X-ray images to automatically learn feature representations and classify OA severity. They report a significant improvement in accuracy compared to the ones obtained by the Wndchrm classifier. They obtain an classification accuracy of 57.6 % to achieve a state of the art accuracy.

5. Dataset

The dataset consists of 4214 X-ray images of the knees (including both right and left leg) which are labeled according to the Kellgren and Lawrence (KL) grading system. The images were obtained from the Osteoarthritis Initiative (OAI) dataset courtesy of the Mobilize Center at Stanford University. Figure 2 shows a few examples of knee OA X-ray images in the dataset. The distribution of the images (right and left leg knees considered separately) based on the KL grading scheme in the dataset is provided in the table below. We observe that our dataset is imbalanced, i.e, 40.45 % of the dataset is stage 0, whereas only 2.29 % of the dataset is stage 4.

Stage	Number of images	% of total
0	3054	40.45
1	1384	18.34
2	1978	26.20
3	960	12.71
4	173	2.29

6. Methodology

We implement two methodologies in this paper. In the first methodology, we use a pretrained convolutional neural network as a feature extractor (transfer learning) and then apply various machine learning models to classify the images based on KL score. In the second methodology, we train a faster region convolutional neural network (R-CNN) to first segment the knee-joint region from the X-ray images and then to classify them based on KL score. We discuss both of these methodologies in detail in the following subsections.

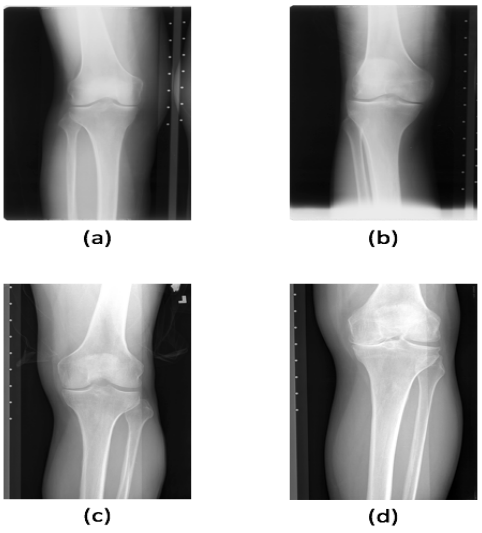

Figure 2. Examples of knee OA X-ray images in the dataset. KL grades for the images are: **(a)** Stage 1, **(b)** Stage 2, **(c)** Stage 3 and **(d)** Stage 4.

6.1. Transfer Learning

In the transfer learning methodology, we have two stages. In the first stage, we extract features from the raw images using a pretrained convolutional neural network and various pooling techniques. In the second stage, we train a machine learning model on the extracted features to classify the images exploring the use of different loss functions. The training pipeline for the current methodology is clearly explained in figure 3.

6.1.1. Feature Extraction

We first pre-process the images to have a fixed size such that loss in resolution and aspect ratio is minimal. After observing a histogram of the aspect ratios of all images in the dataset, we fixed our processed images' aspect ratio to be 1.67 (median value). The size of the processed images was fixed to be 1280x768.

We then extract features from the processed images using a ImageNet pre-trained VGG-16 network. In the report, we compare features extracted from the final pooling layer (pool-5) of VGG-16 network. The input dimension for a VGG-16 network is fixed to be 224x224. Hence we generate an activation map for the input image by convolving a 224x224 window over the processed image with a stride length of 32. The resulting activation map is of size 40x24x512 when we use pool-5 features. In order to re-

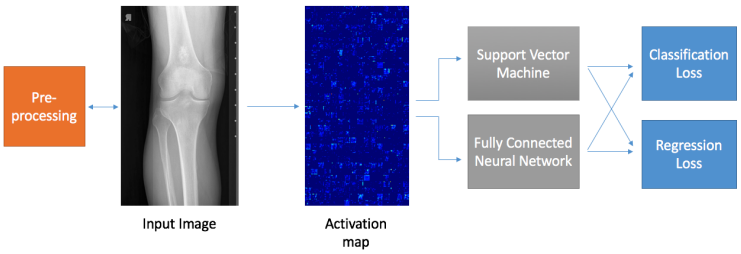

Figure 3. Training pipeline for Methodology 1

duce the dimensions of the activation map, we do a max pooling operation with varying window sizes and strides to down-sample the activation feature size. Another method we use to reduce dimensionality of the features is to apply principal component analysis (PCA) on the activation maps.

6.1.2. Model Selection

We implement two types of machine learning models to train on the extracted features to classify the X-ray images based on KL score.

- **Support vector machine** : We train linear kernel support vector machines (SVM) on the extracted features associated with the KL score. Since our dataset is imbalanced, we applying weight balancing to have class weights to be inversely proportional to the class frequencies in the input data.

- **Fully connected neural network** : We train a two-layer fully connected neural network with dropout on the extracted features associated with the KL score.

We also try out two different loss functions:

- **Classification Loss** : We use a categorical cross-entropy loss function, which is the most common type of loss function used in multi-class classification problems.

- **Regression Loss** : We use a mean-squared error loss function by treating the classification problem as a regression problem during training. The intuition behind using a regression loss function is that our knee OA grading is actually more of a progression from Stage 0 to Stage 4 rather than a strict classification.

6.2. Faster R-CNN

We train a faster region convolutional neural network (R-CNN) with our dataset to achieve the following two objectives: 1. **extract the knee joint regions** from the X-ray images and 2. **classify the extracted knee-joint regions** based on KL score. Faster R-CNN achieves both of these objectives together, where it first learns to identify potential knee-joint regions using the region proposal network and then classifies the knee-joint regions using a object classification network. An overview of the faster R-CNN model is provided in figure 4.

In order to train a faster R-CNN model, we need to label a bounding box around the knee-joint region for each training image. The training is divided into the following 4-step procedure:

- **Step 1** : Train a **region proposal network** with weights initialized from pre-trained ZF network.

- **Step 2** : Train the **object classification network** using proposals from step 1.

- **Step 3** : Re-train **region proposal network** with initialized weights learnt in step 2.

- **Step 4** : Re-train **object classification network** using proposals from step 3.

During testing, the trained faster R-CNN provides 300 region proposals for knee-joint regions for each test X-ray image with a predicted label and confidence score. We choose the region proposal and the label with the highest confidence score.

7. Results and Discussion

In this section, we discuss and analyze the results obtained using both transfer learning and faster R-CNN.

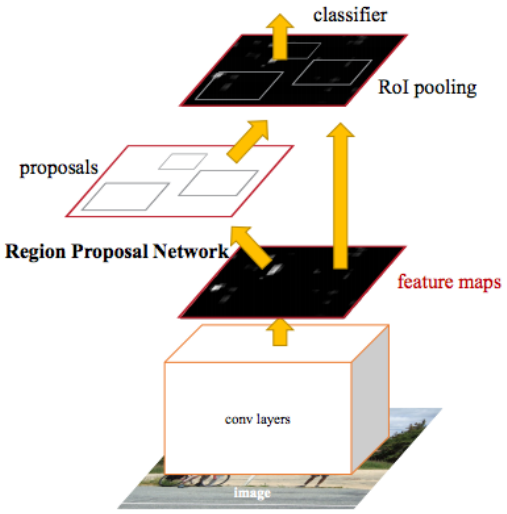

Figure 4. Faster R-CNN model

7.1. Transfer Learning

In this section, we discuss the results obtained using both linear SVMs and fully connected neural networks.

7.1.1. Linear SVM

We first try out a simple SVM on the activation. Interestingly, we were able to completely overfit the training data, hence tuned the regularization parameter $C = 0.01/n$ where n is the number of training samples. This gave us predictions in Table 1.

As the labels follow a successive order, we use the **Pearson Correlation Coefficient** between the predictions and the labels to quantify how 'clean' the predictions are. A higher correlation means that misclassifications are made for adjacent labels as opposed to those far apart. Such a model would be more useful with its predictions.

The ability to overfit to training data indicated a variance problem, while the low accuracy and correlation with labels indicated a bias. We try a more complex model to see if we can solve these issues.

7.1.2. Fully Connected Network (FCN)

We trained a fully-connected network with 2 hidden layers of dimension 2048 and 256, before adding either a classification layer or a regression layer. The best results are reported in Tables 2 - 5. While Table 2, had the best overall classification accuracy, we can see considerable off-diagonal predictions with class 3 and 4 being predicted for class 0.

In Table 3, we confirm our hypothesis that a regression loss would lead to an improvement in prediction as can be seen via the higher correlation score. The overall classification accuracy is slightly lower, however a model that misclassifies successive grades is more useful than one whose predictions do not correlate well.

Label	0	1	2	3	4	**Precision**	**Recall**
0	157	30	36	28	18	0.56	0.58
1	56	19	28	20	10	0.28	0.14
2	60	16	69	39	13	0.40	0.35
3	9	3	29	35	9	0.28	0.41
4	0	0	12	5	1	0.02	0.06
Overall accuracy: 0.40							
Prediction correlation: 0.35							

Table 1. Predictions by linear SVM model

Label	0	1	2	3	4	**Precision**	**Recall**
0	186	30	34	11	8	0.54	0.69
1	64	28	31	6	4	0.29	0.21
2	77	28	62	21	7	0.36	0.32
3	20	9	36	17	3	0.28	0.20
4	0	1	10	6	1	0.04	0.06
Overall accuracy: **0.42**							
Prediction correlation: 0.36							

Table 2. Best FCN classification model

We investigate the source of error in Fig. 5 and 6. We can see in Fig. 5 that there is clearly a bias towards making predictions of a lower grade. While we have resampled our data to overcome the skew in distribution, the model might need a more detailed set of features to make this distinction. After tuning the learning rate and other hyperparameters, in Fig. 6 we can see the learning curves for the best model that indicates overfitting. Either a larger dataset, or lower-dimensional data from segmented images would help reduce this problem.

7.1.3. Fully Connected Network (FCN) on grouped labels

On the hypothesis that our features from the entire image are not detailed enough to distinguish between such granular grades, we group the grades together as 0, (1,2) , (3,4) and then try to examine performance. This is seen in Table 4, 5. We get a much better classification performance with the classifier performing better since the 'graded' effect of the data is reduced with just 3 classes. Also, handling the skewed class weights is more effective in the cross-entropy loss which probably leads to a better performance.

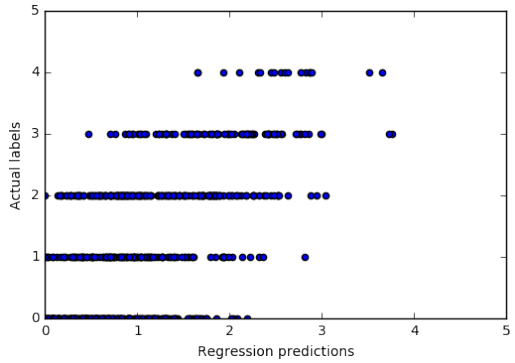

Figure 5. FCN Regression model predictions indicating bias

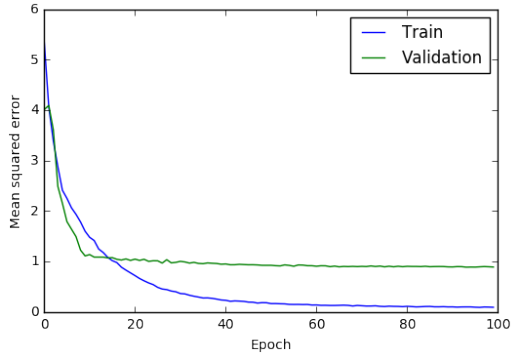

Figure 6. Learning curves for FCN regression model indicating variance

Label	0	1	2	3	4	**Precision**	**Recall**
0	130	118	21	0	0	0.68	0.48
1	35	77	20	1	0	0.24	0.58
2	24	102	61	8	0	0.39	0.31
3	1	22	48	12	2	0.41	0.14
4	0	0	8	8	2	0.50	0.11
Overall accuracy: 0.40							
Mean squared error: 1.96							
Prediction correlation: **0.58**							

Table 3. Best FCN regression model

Label	0	1	2	**Precision**	**Recall**
0	175	90	4	0.56	0.65
1	124	187	17	0.57	0.57
2	14	50	39	0.65	0.38
Overall accuracy: **0.57**					
Prediction correlation: **0.44**					

Table 4. Best FCN classification model on grouped labels

Label	0	1	2	**Precision**	**Recall**
0	162	107	0	0.58	0.60
1	114	211	3	0.52	0.64
2	2	86	15	0.83	0.15
Overall accuracy: 0.55					
Mean squared error: 0.67					
Prediction correlation: **0.44**					

Table 5. Best FCN regression model on grouped labels

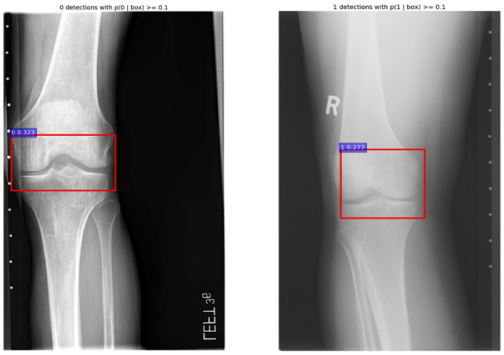

Figure 7. Examples of correctly classified X-ray images (with knee joint region extracted) using faster R-CNN

7.2. Faster R-CNN

We labeled a bounding box around the knee-joint region for 1000 images. We chose 700 images for training (200 of that for validation) the faster R-CNN and 300 images for testing. A few correctly classified examples, along with the knee-joint regions are shown in figure 7. The precision and recall matrix for faster R-CNN model on the test data is shown in table 5.

We observe that the faster R-CNN has a 98 % accuracy in predicting the knee-joint region if the intersection over union (IoU) of the predicted box with the labeled box is at least 0.7. Hence it is extremely accurate in predicting the knee-joint regions from the X-ray images. But we observe that the overall classification accuracy is only 41 % even for faster R-CNN. This is because the number of training images that we used was only 700. Labeling the bounding boxes for the knee-joint region manually is a very time-consuming task and hence we had only a limited number of training images.

8. Conclusion

Hence we conclude with our current resolution of features, distinguishing between more coarse grades is feasible, while finer grades would require better features. From

Label	Precision	Recall
0	0.60	0.55
1	0.22	0.18
2	0.44	0.31
3	0.40	0.49
4	0.22	0.05
Overall accuracy: 0.41		

Table 6. Results from faster R-CNN model

transfer learning, we observe that our best model gave us an overall classification accuracy of 42% for the knee OA grading problem, which is still shy of the state of the art accuracy of 58%. We believe that this is mainly due to the fact that the state of the art paper has used knee-joint segmented images rather than the whole image. We were able to compare results from quite a few machine learning approaches, but our accuracy failed to improve beyond 42%. Using the faster R-CNN method, we were able to achieve near precision accuracy in segmenting the knee-joint regions from the X-ray images. But here again, we failed to get good classification accuracy due to the fact that our training set for faster R-CNN was small.

9. Future Work

We believe that working with the knee-joint segmented images will lead to significant increase in classification accuracy of knee OA grading. In this regard, we believe that training the faster R-CNN with more training images would definitely lead to equal or better than than the state of the art accuracy. For our other approach, improving the resolution of our activation features using segmented images, or Spatial Pyramidal Pooling could prove beneficial. Further, visualizing the features learned by these networks as in Zeiler et. al. would prove useful to understand and improve performance.

10. Bibliography

[1] Joseph Antony, Kevin McGuinness, Noel E OConnor and Kieran Moran, *ArXiv Report*, 2016.

[2] L. Shamir, S. M. Ling, W. Scott, M. Hochberg, L. Ferrucci, and I. G. Goldberg, "Early detection of radiographic knee osteoarthritis using computer-aided analysis" *Osteoarthritis and Cartilage*, vol. 17, no. 10, pp. 13071312, 2009.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *NIPS*, 2015.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", *ArXiv Report*, 2015.

[5] Jonathan Krause1, Hailin Jin ,Jianchao Yang, Li Fei-Fei, "Fine-Grained Recognition without Part Annotations", *CVPR*, 2015.

[6] Matthew D. Zeiler and Rob Fergus, "Visualizing and Understanding Convolutional Networks", *ECCV*, 2014.