Predicting Future Knee Osteoarthritis Using Baseline Knee Radiographs

Matt Titchenal, mtitch@stanford.edu
Nishant Pandit, nish1519@stanford.edu
Stephanie Young, sryoung@stanford.edu

Abstract

Osteoarthritis (OA) is the leading cause of disability in the United States.[1] It is characterized by a loss of cartilage in the knee and the development of abnormal bony growths called osteophytes.[2] Radiologists rely on radiographs of the knee in order to classify Kellgren Lawrence (KL) grades for knee OA progression. These grades determine the severity of the knee OA. Knee OA is difficult to study because its symptoms of pain, stiffness, and abnormal joint function are relatively absent until the disease has already reached an advanced state. Patients suffer from late diagnosis after irreversible damage to the cartilage has already occurred. Therefore, detecting progression of knee OA is very important for clinical treatment and there is a need for diagnostic techniques to detect early signs of the disease to identify patients at risk for future development of OA. The goal of this paper is to demonstrate the efficacy of using deep learning techniques on knee radiograph images in order to predict patient outcomes of knee OA. Our dataset has data up to eight years after the radiograph was taken. In this paper, we implement machine learning algorithms in order to predict progression of knee OA from knee radiograph images. We implement a variety of techniques including transfer learning, feature extraction, batch normalization, and hyperparameter tuning and evaluate them based on their performance metrics, namely precision and recall.

Introduction and Motivation

Knee OA is a common disease with one in two individuals at risk for developing knee OA by the age of 85.[3] Currently, there are no reliable methods for accurately predicting knee OA development for a patient with a healthy knee, likely because the features that predict knee OA progression are highly complex. This paper outlines a novel technique for the prediction of future knee OA development in patients that do not yet have OA at the time of radiograph collection. Such a prognostic tool serves a useful clinical purpose of identifying patients likely to suffer from advanced knee OA in the future. This study utilizes transfer learning from a deep convolutional network designed for large scale image recognition[4] (VGG-16) combined with a two layer neural network that is then trained to predict which patients progress to clinical OA. Furthermore, a variety of techniques are investigated such as overfitting training images, perturbing the training and test image size, using Adam optimization vs mini-batch gradient descent, implementing batch normalization, perturbing the learning rate, L2 regularization, dropout, and finally perturbing the number of layers in order to tune the model to knee OA progression prediction.

Kellgren and Lawrence (KL) Grading

The Kellgren and Lawrence (KL) system is a way of quantifying the severity of knee

[1] J. Antony, K. McGuinness, N. E. O'Connor, K. Moran, N. E. O. Connor, and K. Moran, "Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks," ArXiv e-prints, vol. 1609, no. ICPR 2016 proceedings, pp. 1195–1200, 2016.

[2] https://www.arthritis-health.com/types/osteoarthritis/what-knee-osteoarthritis

[3] S. Suresha, A. Mahajan, N. Dalal, "Automatically Quantifying Radiographic Knee Osteoarthritis Severity," CS229

[4] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition

OA. It is split into five grades (0 to 4).[5] See Figure 1 for a description of the grades.

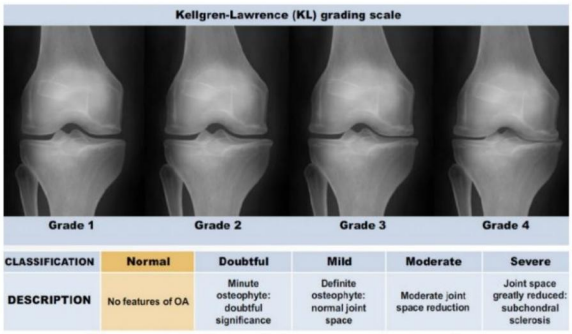


Figure 1: Kellgren & Lawrence Grading System

Related Work

There has been other work using deep learning applied to knee OA. *Antony et al* in "Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks"[6] uses a deep CNN pre-trained on ImageNet data and tuned on knee OA data to improve mean squared error. Another paper, *Suresha et al* in "Automated staging of knee osteoarthritis severity using deep neural networks"[7] again aims to predict KL grades based on X-ray images. The work demonstrated that very deep neural nets were able to significantly increase the performance compared to shallow networks with hand engineered features of predicting KL grades based on X-ray images.

We extend these models to look at knee radiograph X-ray images collected at 12 months, 24 months, 36 months, 48 months, 72 months, and 96 months with radiologist labeled KL grades to determine whether an image taken at baseline (or early on in the progression) is able to be predictive of knee OA progression (KL two or greater) at 96 months.

Dataset

We use a dataset from the Osteoarthritis Initiative (OAI) which consists of X-ray images of both the left and right knees and labeled over an eight year period according to their KL grade. Our dataset has 4794 patients with knee X-rays at baseline (9588 knees) and associated KL grades at baseline, 12 months, 24 months, 36 months, 48 months, 72 months, and 96 months after baseline. We excluded all knees with OA at baseline (KL $\geq$ 2), and were left with 8597 "healthy" knees (KL Grade $\leq$ 1). These were classified into progressors and non-progressors. A patient was considered a progressor if they had clinical OA by the conclusion of the study (KL $\geq$ 2). Otherwise, they were classified as a non-progressor.

Classification	Number of Images
Progressor (KL >= 2 at 96 months)	1057 knees
Non-Progressor (KL < 2 at 96 months)	7540 knees

The input to our algorithm are the baseline knee radiographs for patients that appear healthy according to a radiologist's KL grade (KL Grade $\leq$ 1). We then use a convolutional neural network (CNN) to output a prediction for whether or not these patients progress to arthritis (KL $\geq$ 2) within the 8 year follow-up period.

[5] https://radiopaedia.org/articles/kellgren-and-lawrence-system-for-classification-of-osteoarthritis-of-knee
[6] J. Antony, K. McGuinness, N. E. O. Connor, K. Moran, "Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks"
[7] S. Suresha, L. Kidzinski, E. Halilaj, G.E. Gold, S.L. Delp, "Automated staging of knee osteoarthritis severity using deep neural networks"

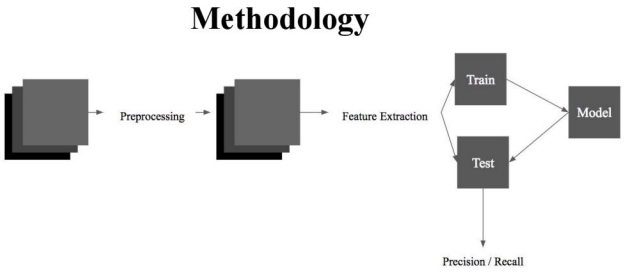


Figure 2. Generalized diagram of CNN model

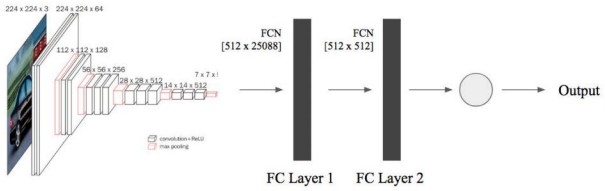


Figure 3. VGG-16 with two fully-connected layers to output binary classification[8]

Preprocessing Data

Normalizing Left and Right Knees

Patient X-xray images had to be normalized such that the left knee and right knee appeared the same. Images were split down the middle and then the right knee image was mirrored in order to mimic the left knee. This data preprocessing helped standardize images as determining between left and right knees is not meaningful towards predicting progression.

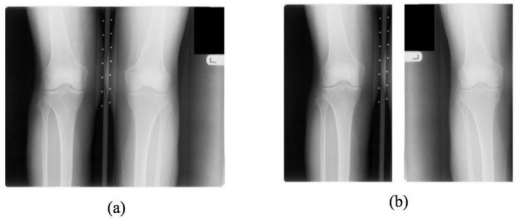


Figure 4. Examples of image mirroring to standardize left and right knee. (a) Left and right knee from a patient (b) Mirrored X-rays to convert all images to left knee equivalents

[8]https://www.researchgate.net/figure/VGG16-architecture-16_fig2_321829624

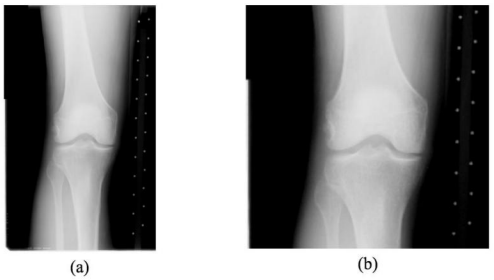


Figure 5. Examples of data resizing / cropping. (a) Before preprocessing (b) After preprocessing

Data Scaling and Centering

We also perform image scaling and cropping on the images. The images are of varying dimensions so we resize and crop the images to become numpy arrays of size 224 x 224. This has the additional effect of identifying and extracting only the knee area of the X-ray.

Training / Test Set Selections

We randomly sampled images from our patient pool in order to compose our training, development, and test sets. Our image data contains many more non-progressor images than progressor images. We therefore decided to randomly sample proportionally from both the progressor and non-progressor sets (e.g. 70% of progressors for training and 70% of non-progressors for training) in order to balance the number of progressors and non-progressors in our training set. Even so, when running our initial model, we found our precision and recall to be very low (~10%). We therefore decided to un-bias our data further and randomly selected a fixed number (700 images from progressors and 700 images from non-progressors for experimentation and 1500 images from progressors and 1500 images from

non-progressors for our final model) in order to train. We decided to forgo a development set as we did all model tuning on the training set.

Transfer Learning / Feature Extraction (VGG-16 & Maxpool-5)

We utilize a very deep convolutional neural net pre-trained on ImageNet data. This pre-trained deep CNN has several features. It utilizes small (3x3) convolutional filters in combination with a very deep network.[9] The authors thus hypothesize that larger filters should be able to perform even better when used in conjunction with current models. Most importantly, the model is also shown to generalize well to other datasets.[10]

For our dataset, we extracted features from our X-ray images using VGG-16 with a maxpool-5 filter. We chose to use this pooling layer based on work done in Project XVision which applied VGG-16 to chest X-rays.[11] We then used a two layer neural net on top of our pre-trained model in order to make predictions.[12] The final output of feature extraction is a content feature matrix of shape 7x7x512 which we then put into a two-layer neural net.

Model

2-Layer Fully Connected Neural Network

We trained a two layer fully connected CNN on the extracted features to predict KL score progression. See the table below for our parameters, we used the same parameters as XVision as a starting point.[13] Optimizer: Mini-Batch Gradient Descent, Mini-Batch Size: 20, Loss Function: Binary Cross-Entropy, Epoch Number: 20.

Experiments

From here, we executed a number of experiments in order to increase precision and recall. See our table below for a summary of our experiments:

Experiment	Parameters	Train	Test
#1: 700 progressors / 700 non-progressors (90% / 10%)	No Batch Normalization Learning Rate = 0.01 2 Layer	P: 58% R: 58%	P: 47% R: 47%
#2: Same as #1	No Batch Normalization Learning Rate = 0.001 2 Layer	P: 75% R: 71%	P: 53% R: 53%
#3: Same as #1	Batch Normalization Learning Rate = 0.01 2 Layer	P: 99% R: 99%	P: 61% R: 61%
#4: Same as #1	Batch Normalization Learning Rate = 0.01 2 Layer Adam Optimizer	P: 97% R: 98%	P: 60% R: 60%
#5: Same as #1	Batch Normalization Learning Rate = 0.01 2 Layer L2 Regularization	P: 99% R: 99%	P: 60% R: 60%
#6: Same as #1	Batch Normalization Learning Rate = 0.01 2 Layer Dropout = 0.9	P: 99% R: 99%	P: 59% R: 59%

Discussion

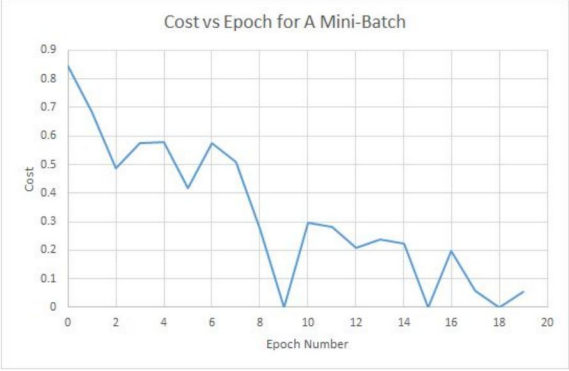
We found from our experimentation that batch normalization by far improved our precision / recall the most. We found that stochastic gradient descent was slightly better than Adam and that regularization (L2) had little effect. Dropout at 90% had little effect but we later find that dropout at 50% has a positive effect. We also tried Maxpool-4 instead of Maxpool-5 and a batch size of 100 instead of 20 to try to reduce overfitting to the training data with very little difference in test precision / recall (not shown in table).

[9] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition
[10] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition
[11] https://github.com/ayush1997/Xvision
[12] https://github.com/ayush1997/Xvision
[13] https://github.com/ayush1997/Xvision

Adding In Time Series Information
Finally, we augmented the data set by including images of progresors / non-progressors from different time points (e.g. not just images from baseline but also from 12 months, 24 months and so on). If a patient was KL <= 1, we used all the images during which they were classified as KL <=1 not just their baseline time = 0 image. The following table shows our final models' precision and recall after augmenting the data:

Final Models	Parameters	Train	Test
Model #1: 1500 progressors / 1500 non-progressors (90% / 10%)	Batch Normalization Learning Rate = 0.01 **2 Layer Network** **No Dropout**	P: 99% R: 99%	P: 67% R: 67%
Model #2: same as #1	Batch Normalization Learning Rate = 0.01 **2 Layer Network** **Dropout = 0.9**	P: 99% R: 99%	P: 66% R: 66%
Model #3: same as #1	Batch Normalization Learning Rate = 0.01 **1 Layer Network** **Dropout = 0.9**	P: 99% R: 99%	P: 68% R: 68%
Model #4: same as #1	**Batch Normalization** **Learning Rate = 0.01** **2 Layer Network** **Dropout = 0.5**	**P: 99%** **R: 99%**	**P: 70%** **R: 70%**

(a)



(b)

Figure 6. Data for our final models. (a) Model #4 is the best model with the highest precision / recall (b) Mini-Batch descent makes learning jumpy

Discussion

Our study demonstrated a proof-of-concept technique for use deep learning to detect future knee OA progression based on X-ray radiographs of currently healthy knees. This represents a significant improvement on the status quo since radiologists are limited to features detected by the human eye. Our best model demonstrates 70% precision / recall on detecting progression (KL >= 2) by the end of eight years. However, we know that our model greatly overfits the training data and we were unable to lower our model's variance. However, it is remarkable that the algorithm was able to perform as well as it did, given that radiographs only provide information on bony structures, while soft tissue information from cartilage and other structures in the knee may be important in predicting knee OA.

Future Work

The prevalence of OA is biased towards women who have a higher rate of OA than men.[14] Therefore, including demographic information as features may help our model predict OA progression by using image features as well as demographic signals. In addition, to try to fix the overfitting problem, data augmentation or adding noisy data may be able to decrease our variance. In addition, we currently treat our model as time-invariant. Using resnets that include sequential information from each time sample of images (e.g. 12 months > 24 months) we anticipate may improve model prediction since previous time slot images are predictive of future time images.

[14] S.M. Hussain, F.M. Cicuttini, B. Alyousef, Y. Wang, "Female hormonal factors and osteoarthritis of the knee, hip and hand: a narrative review"

Contributions

Matt: Researched available datasets and completed data acquisition from OAI, wrote code for dataset labeling and preprocessing, drafted project proposal, milestone report and project poster, helped edit final report.

Nishant: Resolve software dependencies to create a Ubuntu workstation (& AWS). Wrote code for image data redistribution, porting dataset into model, various model configurations and led effort in running experiments. Helped edit project milestone, project poster, and final report.

Stephanie: Wrote final project report, wrote code for dataset redistribution, training / test / dev set selection and splitting, and assisted with experimental design. Helped edit project poster and project milestone and led in figure creation.