Predict Total Knee Replacement among Osteoarthritis Patients using Supervised and Semi-supervised Networks with Magnetic Resonance Imaging

Hong Gao
New York University
New York, NY
hg1196@nyu.edu

Mingsi Long
New York University
New York, NY
ml5893@nyu.edu

Yulin Shen
New York University
New York, NY
ys2542@nyu.edu

Jie Yang
New York University
New York, NY
yangj14@nyu.edu

ABSTRACT

This project[1] aims to learn representations from 3D knee Magnetic Resonance Imaging(MRI) and predict total knee replacement in the next eight years through two deep learning techniques: a supervised convolution neural network (GlosyNet) using labelled images and a semi-supervised ladder network using both labeled and unlabeled images for training. Although the training time increased significantly, we found that by incorporating more unlabeled images in the training using the LadderNet, the performance improved. The accuracy increased from 0.58 to 0.65, and the AUC improved from 0.53 to 0.62.

KEYWORDS

Osteoarthritis, knee replacement, Magnetic Resonance Imaging, image classification

1 INTRODUCTION

Osteoarthritis (OA) is a disease of the synovial joint tissues which leads to pain and long term disabilities [4][5][13]. The prevalence of symptomatic knee OA has been increasing over the past several decades in the US, concurrent with an aging population and the growing obesity epidemic. There are 14 million individuals in the US who have symptomatic knee osteoarthritis. A lot of the patients have had sufficient progression such that they are eligible for total knee replacement (TKR) [2][4][5][7][16]. Knowing ahead of time which patient will progress and eventually end up with knee replacement becomes important. This can help clinicians and researchers to target the right patients and start with early intervention including clinical trials. Since the pathology is not perfectly understood, there

[1] This project is a Capstone project at NYU Center for Data Science in Fall 2018. It is supervised by Professor Cem M. Deniz, PhD at NYU Langone Health, Department of Radiology. The codes for this project are available at https://github.com/NYU-CDS-Capstone-Project/OA_TKR_imaging.git

is a lack of biomarkers of OA development and progression, making the development of effective medical treatment difficult [2]. One way is to use deep learning in Magnetic Resonance Imaging (MRI) to learn representations that can early signal the likelihood of getting total knee replacement.

Problem Statement: In this study, we would like to use 3D MRI to predict the probability of TKR within eight years.

2 LITERATURE REVIEW AND RELATED WORK

MRI has become a central tool to quantify relaxometry and morphology of knee [1]. Sharma L. et al. found that MRI tissue lesions help improve prediction of mild to moderate disease among people with higher risk for knee OA with normal X-rays [19] Barr A.J. et al. constructed three-dimensional bone measurements from MRI and detected their association with total knee replacement [1]. In constructing risk calculator for development of OA over 8 years, Joseph G.B. et al. demonstrated the inclusion of MRI-based morphological abnormalities and cartilage T_2 significantly improved model performance [8].

Besides traditional measures, deep neural networks, a machine learning technique, is used in processing MRI images and extract meaningful features. Norman B. et al. used U-Net convolutional network to perform automatic segmentation on MRI [14].

Some new deep learning techniques are introduced in image processing field, which can be applied in OA MRI to help increase accuracy and interpretability. Ronneberger O. et al. proposed U-Net Convolutional Network which outperforms the prior best methods on imaging tasks[18]. In their recent paper, Hjelm R.D. et al. considered the problem of building a more proper encoder method to encode the image input as a digital vector and experimented a deep INFOMAX network on image classification problems[6]. Based on the mutual information neural estimation, the authors showed that good representations can be learned without the generator/decoder. They also found that a simple classifier based on the Jensen-Shannon divergence is a more stable and efficient discriminator since it leverages local structure in the input to improve suitability of representations for classification. In 2015, Rasmus et al. proposed Ladder Network, which combines supervised and unsupervised learning. The model tries to minimize both supervised and unsupervised loss simultaneously, avoiding the need for layer-wise pre-training [17].

3 DATA

3.1 Data Source

We use data from the NIH Osteoarthritis Initiative (OAI) database[2]. It is a multicenter, prospective, longitudinal observational study for knee OA. 4,796 participants ages 45-79 were recruited during Feb 2004 to May 2006. The study collects demographic and clinical information, including X-rays and 3D MRI. These patients are invited back to visit every six to twelve months for eight years. Both left and right knee conditions are recorded until they drop out of the study. Three orientations of the sagittal are acquired from the MRI: (1) the entire patella and as much of the suprapatellar bursa as possible; (2) orthogonal to the coronal acquisitions and sagittal to the joint; and (3) perpendicular to a line tangent to the posterior cortices of the femoral condyles[16]. This paper only focuses on the first set of MRIs. Figure 1 shows an example of the MRI. Since the MRI is 3D, Figure 1 only shows the center slice along the thrid dimension (more will be explained in later sections). Overall, about 10% of the patients got total knee replacement.

[2] https://data-archive.nimh.nih.gov/oai

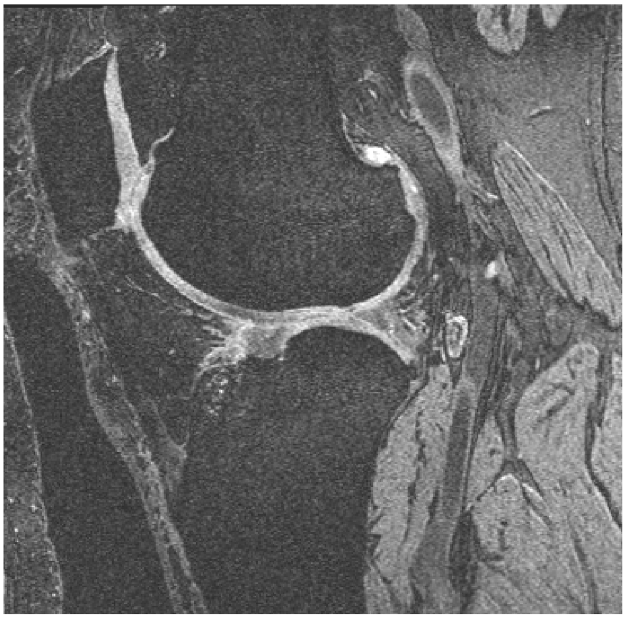

Figure 1: A sagittal slice of 3D Knee MR Image

3.2 Outcome of Interest

We define the OA progression as a binary outcome. It represents the development of moderate to severe radiographic or symptomatic knee OA of the subjects with a baseline Kellgren-Lawrence (K-L) grade ≤ 2 [11] and no symptoms in the knee (the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) [15] pain score ≤ 1). This outcome is defined as 1 if any of the following during the studied eight years: receiving a TKR confirmed by medical records and/or knee radiographs; worsening to K-L grade 3 or 4; and a WOMAC pain score ≥ 5 at any two follow-up time points [9]. This composite outcome is chosen to obtain a broader clinical significance or applicability compared to single outcome measures such as receiving a necessary TKR only.

3.3 Study Cohort

This study implements both supervised and semi-supervised learning to identify OA-relevant imaging biomarkers. Supervised learning uses images labelled by the outcome of interest. Semi-supervised learning falls between supervised learning and unsupervised learning, which uses unlabeled images. In a nutshell, the selected case-control pairs are based on demographics and clinical information collected. These patients have a binary outcome, and therefore are considered to be in the supervised cohort. Those who are not selected in the case-control pairs are considered in the unsupervised cohort.

3.3.1 Supervised Cohort. There are 1,170 patients in the supervised cohort. Cases are defined as individuals who have a positive OA progression outcome in either knee after the baseline enrollment date; individuals who receive a partial knee replacement in either knee or have a knee replacement before baseline are removed. Controls are defined as individuals who appear at the 96-month visit and do not have a positive OA progression outcome in either knee by that visit. Lastly, each case and control must have baseline information for the confounding variables age, body mass index (BMI), and gender. Each case was matched to a control such that each matched pair consisted of subjects with the 'exact same' gender and age. In the one to one matching process[3], an additional constraint was added that the matched subjects have the same BMI within a specific tolerance, that is, no more than a 10% BMI difference between subjects. There were no significant differences between case and control groups with regard to age, height, weight, and BMI. The dataset from case-control pairs contain either the left or right knee images from each subject. In case a subject had a positive outcome in both knees during the data collection, the knee with a positive outcome first recorded into the image dataset was included.

3.3.2 Unsupervised Cohort. The subjects who are not selected for the supervised cohort will form unsupervised cohort. This cohort includes 3,626 subjects. The details of both cohorts are in the table 1.

4 MODEL

4.1 Supervised Learning: GlosyNet

Our supervised learning model is designed based on the Very Deep Convolutional Networks (VGG) which was published by Simonyan K. and Zisserman A.[20]. The architecture and parameters are tuned best to our current capacities.

4.1.1 Architecture. Our inputs are 1,170 fixed-size 448 × 448 × 37 3-dimensional MRI images. We use 70% as training set, 10% as validation set, and the rest 20% as test set. During preprocessing, all the images are cropped to size of 348 × 348 × 32 and then go through standard normalization where the mean is subtracted to the image and divided by its standard deviation. After that a various of augmentation mechanisms are applied. Images in the training set are randomly cropped and then randomly flipped whereas validation and test images are center cropped.

Processed images are passed through a stack of convolutional (conv.) layers, where we use small 3 × 3 × 3 kernels. The first two conv. layers have stride size as 2 × 2 × 1, and stays at 1 × 1 × 1 as we would like to extract features from the first two dimensions while preserve the information in the third dimension. In order to drastically reduce spatial dimension to decrease computation cost and prevent overfitting, spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers. Two types of max-pooling are applied, where the first one uses kernel size and stride size of 2 × 2 × 2, and the second one has both sizes of 2 × 2 × 1. The former downsamples in all three dimensions, whereas the latter only does so in the first and second dimension. After a stack of conv. layers, a fully connected layer (FC) with 9,216 channels is passed following an average pooling layer. In our model, the average pooling layer has kernel size of 1 × 1 × 1, which preserves all the information from previous layers. All hidden layers are equipped with the rectification nonlinearity [12].

The detailed configuration of GlosyNet is shown in Figure 2. The number of channels of conv. layers start from 32 in the first layer and then increase by a factor of 2 after each max-pooling layer starting from the second max-pooling until it reaches 128. There are 1,209,698 trainable parameters in this model.

4.1.2 Training. There are about 411 batches for training data with two images in each batch. During training, the learning rate is set at 0.00005. We use cross entropy

[3] without replacement

	Matched Cohort (supervised)		Un-matched Cohort (unsupervised)
Characteristic	**Controls**	**Cases**	**All**
Age (y)	61.00 (8.82)	61.21 (8.84)	61.15 (9.31)
Height (m)	1.67 (.090)	1.68 (.088)	1.68 (.093)
Weight (kg)	80.67 (14.57)	82.70 (15.28)	81.09 (16.80)
BMI (kg/m2)	28.71 (3.98)	29.24 (4.38)	28.51 (5.03)
Male (%)	0.36	0.36	0.43
White (%)	0.82	0.82	0.78

Table 1: Demographics Information for Cohorts

to estimate the empirical loss and Adam as the optimizer, which is a popular gradient-based optimization method. We perform 10 epochs on GlosyNet. It takes about five hours to train 10 epochs. The entire training is done on Skynet HPC in NYU Langone Health with Pytorch 0.4.1.

4.2 Semi-supervised Learning: Ladder Network

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data [3]. As we know, there is usually a lot more unlabelled data, and labelled data requires more resource to find. For example, in this study there are 1,170 labelled MR images compared to more than 7,000 unlabelled images. In addition, it has been discovered that the use of unlabeled data together with a small amount of labeled data can improve the accuracy considerably [3]. Therefore we decide to use Ladder Network, which combines supervised learning with unsupervised learning in deep neural networks, to do semi-supervised learning. We would like to see if Ladder Network could improve the performance in our project.

4.2.1 Architecture. Similar to most of neural network models, Ladder Network is composed of encoder and decoder. Unsupervised learning is often used to pre-train models. What's special about LadderNet is that it is designed to do both unsupervised learning and supervised learning at the same time. The overarching architecture of LadderNet is illustrated in Figure 3. The encoder in the figure is based on the GlosyNet. The decoder is the corresponding upconv-net, where 3D transposed convolution operators are applied to upscale the images following the inverse direction in Figure 2. The decoder can be considered as the "inverse" net of the encoder. Our LadderNet model is relatively "big", since the decoder has almost the same number of parameters as the encoder. In our case, the total number of parameters of the Ladder is 2,176,099, which is almost two times of the total parameters of GlosyNet.

As shown in Figure 3, the model consists of two different encoders, which are a clean encoder and a corrupted encoder. They support supervised learning and unsupervised learning respectively. The two encoders share the same network architecture. The only difference between them is that the corrupted encoder adds Gaussian noise to the inputs of the all layers. The decoder reconstructs the corrupted outputs of the each encoder layer. The denoising function is a de-convolution function with batch normalization. The reconstructions of decoder are used in unsupervised learning.

Ladder network is trained to do unsupervised and supervised learning simultaneously, by minimizing the sum of supervised and unsupervised cost. Therefore the LadderNet avoids the need for layer-wise pre-training

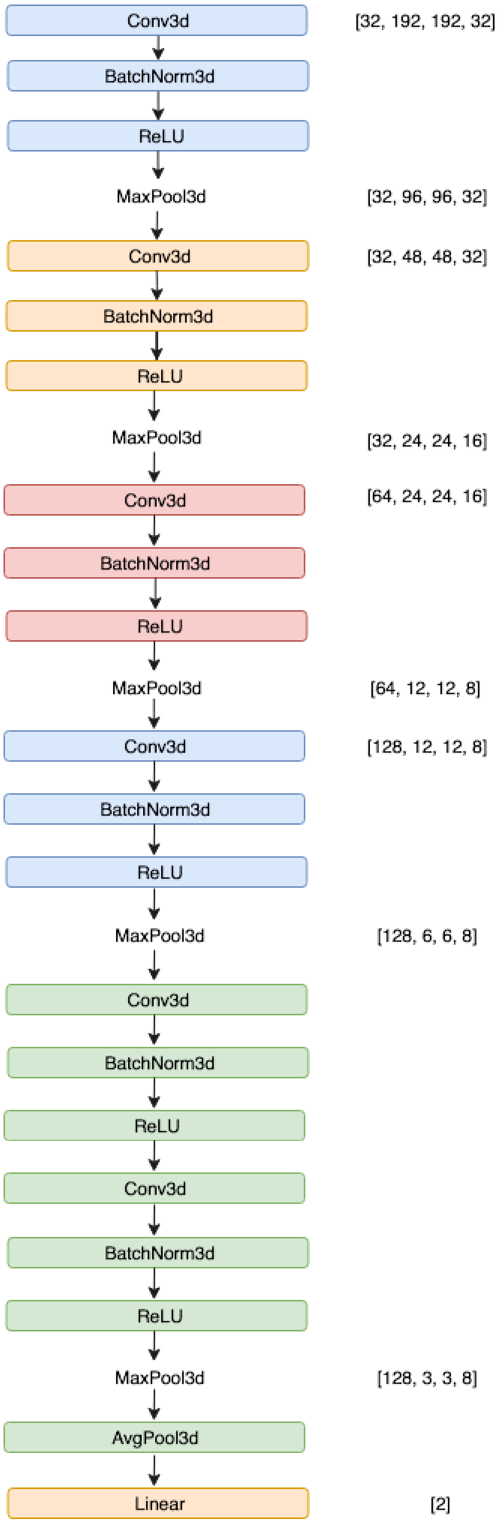

Figure 2: GlosyNet

[3]. The total cost is the weighted sum of supervised loss and unsupervised loss, where the supervised loss is the cross entropy error of the prediction made by the encoder, the unsupervised loss is the mean square error of the original image presentation and the image presentation reconstructed from decoder of the image with noise [17]. Specifically, the loss calculations are in the following equations.

For each unlabelled image, the unsupervised loss of the layer

$$\text{Unsupervised cost of } i\text{−th layer} = ||Y^i - \widetilde{Y}^i||^2,$$

where $||\cdot||$ means the L^2 norm of tensors, Y^i is the clean image representation of the $i-$th encoder layer, and $\widetilde{Y}^i$ is the reconstructed representation of the image with noise in the $i-$th encoder-decoder layer. The total unsupervised loss of the unlabelled image is

$$\text{Unsupervised cost}$$
$$= \sum_{i=1}^{N} \text{u\_cost}[i] * \text{Unsupervised cost of } i\text{−th layer}$$

where N is total number of layers in the LadderNet, u_cost is the weights assigned to the unsupervised loss of each encoder-decoder layer.

For each labelled image,

$$\text{Supervised cost} = y \log \widetilde{y} + (1 - y) \log(1 - \widetilde{y})$$

where y and $\widetilde{y}$ are the true and predicted labels of the image. Notice that here the predicted label is the output of the clean encoder, which is the exact GlosyNet.

The total cost of LadderNet for a dataset which contains both labelled and unlabelled images is

$$\begin{aligned}\text{Total cost} &= \text{Average Supervised cost}\\ &+ \text{Average Unsupervised cost}\end{aligned}$$

We refer the code of Abhishek Kadian[10] in PyTorch, which implemented LadderNet based Linear Neural Network.

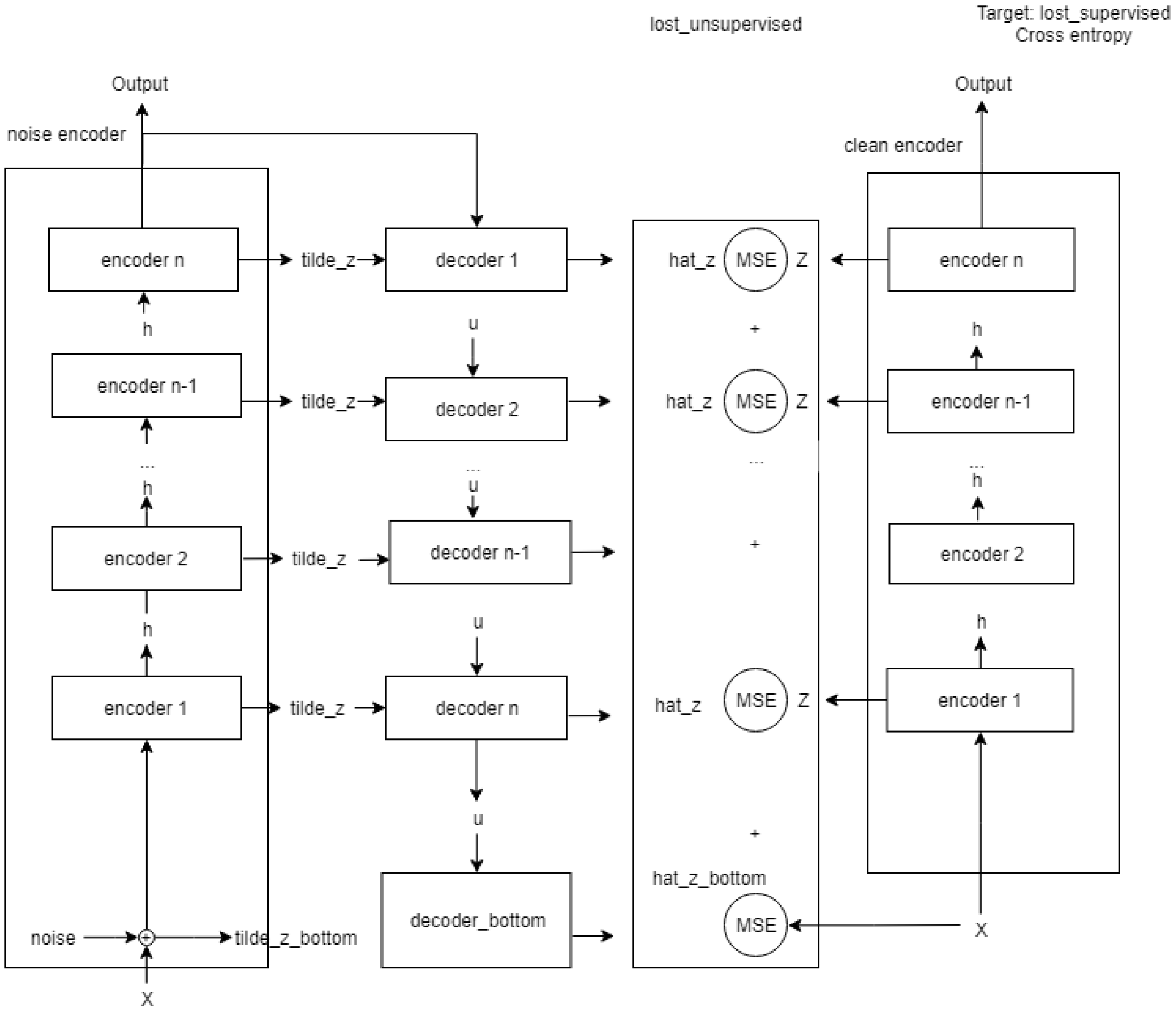

Figure 3: LadderNet Architecture

4.2.2 Training. Since the semi-supervised learning model's encoder is based on the GlosyNet, in order to compare the two networks' performance, we use the same labelled image dataset, which contains 1,170 labelled 3D MR images. In addition, We use 2,000 unlabelled 3D MR images in the training. These 2,000 images are selected from about 7,000 total images we have in order to reduce training time. All the images have fixed-size 448 × 448 × 37 , center cropped them to size of 348 × 348 × 32.

In this model, we load two unlabelled images in each batch, and the standard deviation for the Gaussian noise is set to 0.2. The seed is set as 42 to produce random value. In the first layer of the encoder, its in channel is 1. After that in the next couple of layers the channel becomes 32, 32, 64, 128, 128. In the decoder, its in channel is 128 which is the same as the last layer in the encoder. After that the channel becomes 128, 128, 64, 32, 32, respectively. Because the training takes a significant amount of time (almost 1 day per epoch), we choose to run 8 epochs in the interest of time.

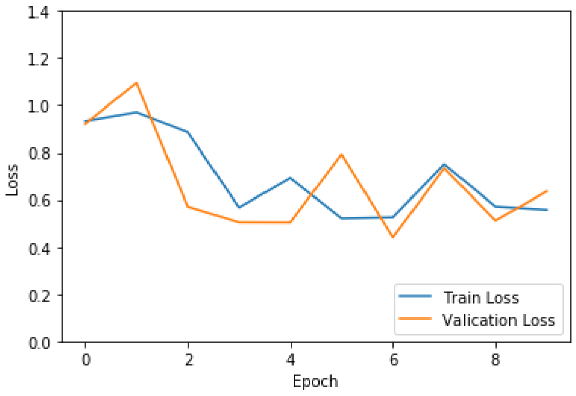

Figure 4: Loss for GlosyNet

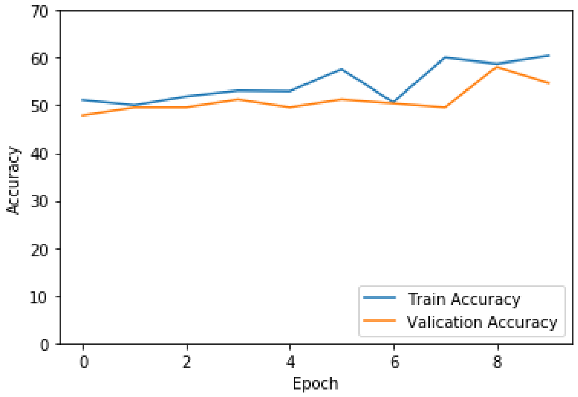

Figure 5: Accuracy for GlosyNet

5 RESULT

5.1 GlosyNet

The training and validation loss decreased significantly in the first epoch and stayed at a relatively low level (about 0.55 for training set and 0.6 for validation set) after training for three epochs (Figure 4). Both train and validation accuracy show smooth and steady increase during the training process. The validation accuracy improved from 50% in the beginning to a range of 54% to 58% after the training for 10 epochs (Figure 5). The best accuracy it reaches is 58.1. The Area Under the Curve (AUC) performance of GlosyNet is 0.53 (Figure 6).

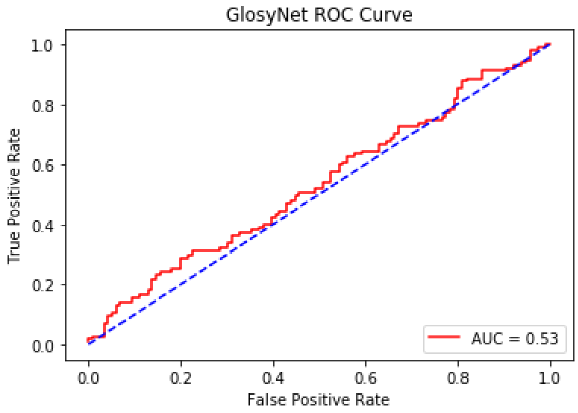

Figure 6: AUC for GlosyNet

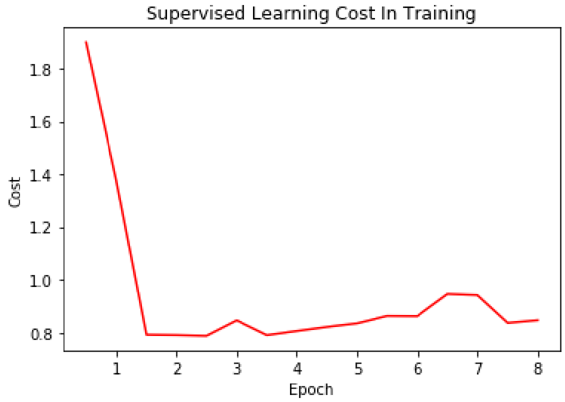

Figure 7: Supervised Loss for LadderNet

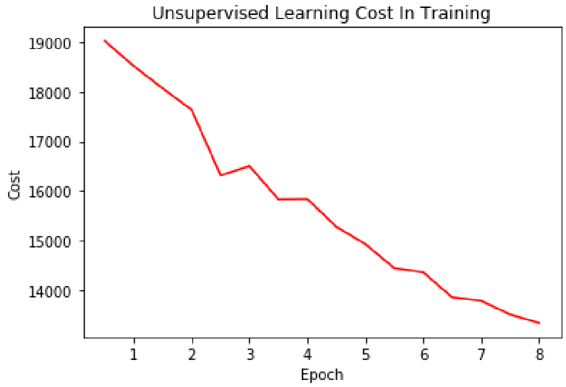

Figure 8: Unsupervised Loss for LadderNet

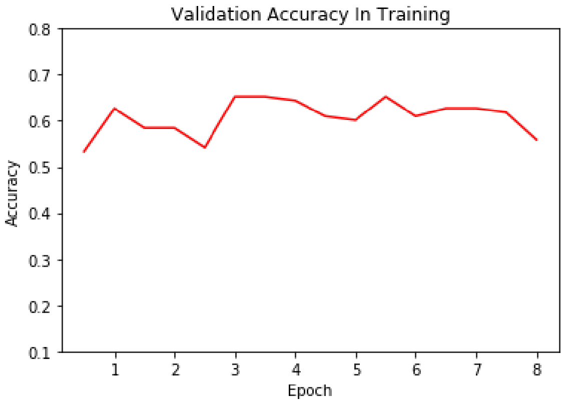

Figure 9: Validation accuracy for LadderNet

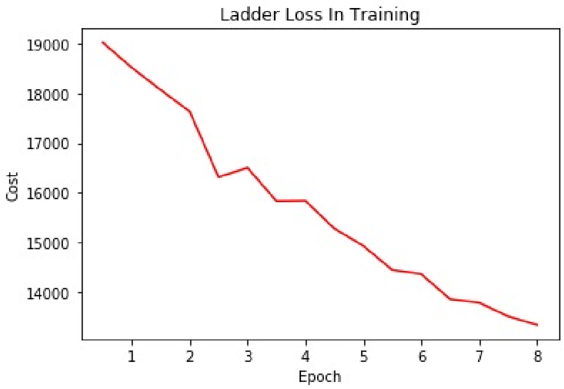

Figure 10: Loss for LadderNet

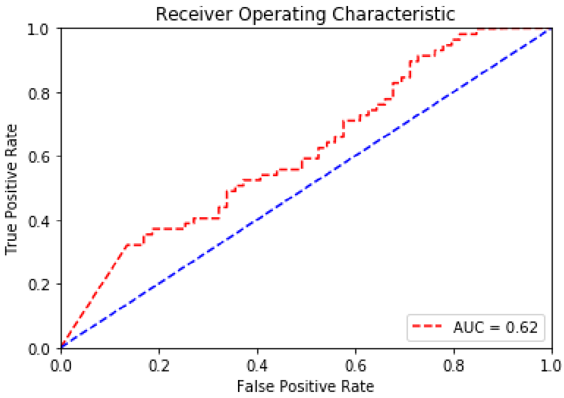

Figure 11: AUC for LadderNet

5.2 Ladder Network

In figure 10, we can see the total loss which includes supervised loss and unsupervised loss decreases significantly during the training process. This implies that the model has been learning well. The reason why the loss is so large is that the unsupervised loss is calculated from both the difference with original images and reconstructed images. Each cost from a pair of them are the sum of 348 × 348 × 32 values. Also we adjust u_cost parameters from 0.000001 to 10 to change the scale of unsupervised cost. We found that when u_cost = $[0.1^6, 0.1^6, 0.1^6, 0.1^4, 0.1, 1]$, the model performs the best. In Figure 8 and 7, it is obvious that the unsupervised cost is much larger than the supervised cost. Although we think the learning process is successful, the validation accuracy is still not optimal, and seems unstable overall. The best accuracy reached is around 0.65. The accuracy does not increase significantly after certain point as the loss keeps decreasing. Sometimes the accuracy even decreases. We can see it in Figure 9. Two ROCs in Figure 11 and 6 show that although LadderNet really improves upon the supervised learning in GlosyNet as AUC increases a lot, we have not reached the desired prediction performance yet based on the current training time, which is only about a week. Since the loss looks still very big, we can assume the result may get better when the loss is trained into a relatively small scale.

6 CONCLUSION

6.1 GlosyNet

Using small-sized kernels in a very deep convolution networks have already demonstrate to improve performance on image classification problems [20]. With the stack of conv. layers, GlosyNet proved that this structure is able to learn effectively from 3D images and reduce loss in a relatively fast speed. The trends in train and validation accuracy also indicate that the network is stable and is indeed learning different features from the images gradually. In terms of generalization, GlosyNet is able to make predictions relatively well on the test set, yielding an AUC at 0.53.

Other than GlosyNet, we also explored a few other networks with different structures by tuning the parameters on the convolution layers and kernel sizes. The GlosyNet so far has the best performance.

However, by looking at Figure 5 closely, both train and validation accuracy has a sign of increasing at the end of 10 epochs and we don't observe a clear divergence between them. This might suggest that although the network has been learning, it has not finished the full search and converge to the best state. The conclusion can be made from looking at Figure 4 as well as the loss still exhibits a decreasing trend at the end of the training. Given more time and computation power, GlosyNet is expected to reach an optimized state and perform better in terms of accuracy.

6.2 Ladder Network

From the result we can see:

(1) Compared with GlosyNet, the AUC improves from 0.53 to 0.62.
(2) Although the validation accuracy has been improving, it is not stable. The loss decreases very fast but the accuracy doesn't increases with the same speed during training, as shown in the figures 9 and 10.

Figures 7 and 8 show that the supervised loss is much lower than the unsupervised loss, which suggests that the LadderNet focuses on denoising during these training epochs. This potentially explains the unstable status and the slow improving speed in the validation accuracy.

To improve the performance of LadderNet, we propose the following potential ways:

- Train the model for more epochs until the unsupervised loss decreases to small enough until the supervised loss is trained;
- Tune the u_cost hyperparameters to make the LadderNet more efficiently;
- Use more unlabelled images with slow training process to give more information on unsupervised learning.

7 DISCUSSION

Applying deep neural networks to MRI to solve classification problems in OA has not yet become fully studied. Efforts in the literature involved extensive domain knowledge and feature extractions. This is among one the first a few attempts to identify the sensitivity and specificity of MRI. Although image classification problem in general has many encouraging development in recent years, it is always a challenging work to apply in the medical field. One consideration is that the model has to ensure a confident true positive rate and high accuracy. Particularly in our case, we are hoping to identify patients with high risk who progress fast and intervene before they need knee replacement. Since medical intervention is expected to slow down the progression, it is important to have a model with minimum false negative rate so it doesn't exclude any patients that would benefit from the interventions.

Another consideration is the deployment of models. Training models is just the first step in applying models in real-world settings. Many aspects need to be considered before production. For example, a machine learning engineer system needs to be built to host the model and efficiently take in new medical record and produce results that are readable and explainable for doctors. The system should also be seamlessly integrated with the high secure medical record system to ensure patients' personal information to be protected at all times.

7.1 General Issues

We have encountered several issues during exploration and development of different models.

The biggest challenge is the size of the images. It requires a long time to load and preprocess even before passing to the model. It takes longer time to go through all the conv. layers, making it very difficult to train and tune in a limited time frame.

Another challenge is the computation power. Even with multiple GPUs on HPC, deeper than GlosyNet networks always hit memory errors, limiting our options to explore deeper and larger networks. We tried another convolution structure with about 23 million parameters, the accuracy of which got to 53% after five epochs. However, it hit memory limit right after that. One guess is our code fails to reset all the parameters so memories were used in augmenting matrices. Another guess is that the model will continue training with larger GPU nodes. This needs more time to research.

7.2 Recommended Future Steps

Overall, by taking more unlabeled images into account, LadderNet yields to better performance than GlosyNet. The time for each epochs in training though increased significantly at the same time. We believe the following steps will help improve the model:

7.2.1 Incorporating Demographic and Clinical Data. The main focus of this paper is to explore the utility of MRI and deep neural networks in identifying OA progression, which is why only MR Imaging are used. The NIH Osteoarthritis Initiative (OAI) collected comprehensive demographic and clinical data, such as pain level and BMI. It would be interesting to explore whether adding this information to the model would increase performance. Specifically, for GlosyNet, it might be helpful to add the personal data before the last fully connected layer. And for LadderNet, these information could be added to the encoder output.

7.2.2 More Images. One biggest road block of this analysis is the time limit. If more time is given, it is recommended to take advantage of more image data collected by the OAI. In LadderNet, we only used 2,000 images for the unsupervised part, where 7,000 images are available.

7.2.3 More Angles. As mentioned briefly in Section 3.1, the OAI collects MRI in three orientations. Incorporating the other two angles might help improve model performance. It is recommended to develop new models or just to include the images in the existing models as a starting point.

7.2.4 Try and Error. With more time and computation power, exploring more model structures and tuning hyperparameters are good exercises. For example in GlosyNet, there could be many different ways to construct conv. layers and spatial pooling layers. Deeper neural networks are also recommended to see if more channels and parameters help improve model performance. Hyperparameters in LadderNet including the standard deviation of the noise injection and the denoising weights at each layer can also be explored more in detail.

7.2.5 More Time. Another suggestion is simply training the model for a longer period of time. Deep neural networks always require relatively longer time to learning all the features and parameters. Especially for a complex network like LadderNet. In the result section, we discussed that GlosyNet still has potential to further decrease loss and increase accuracy. There is no surprise if LadderNet reaches a much higher accuracy training after training for a few more epochs.

8 ACKNOWLEDGEMENT

This work is fully supported by the Center for Data Science Capstone faculties. We thank our project advisor, Professor Cem M. Deniz, PhD who provided valuable insight, expertise, data and the computational platform that made everything we did in this research possible.

REFERENCES

[1] Andrew J. Barr, Bright Dube, Elizabeth M. A. Hensor, Sarah R. Kingsbury, George Peat, Mike A. Bowes, Linda D. Sharples, and Philip G. Conaghan. 2016. The relationship between three-dimensional knee MRI bone shape and total knee replacementâĂŤa case control study: data from the Osteoarthritis Initiative. *Rheumatology* 55, 9 (2016), 1585–1593. https://doi.org/10.1093/rheumatology/kew191

[2] Johannes WJ Bijlsma, Francis Berenbaum, and Floris PJG Lafeber. 2011. Osteoarthritis: an update with relevance for clinical practice. *The Lancet* 377, 9783 (2011), 2115 – 2126. https://doi.org/10.1016/S0140-6736(11)60243-2

[3] Rino Boney. 2016. Introduction to Semi-Supervised Learning with Ladder Networks. http://rinuboney.github.io/2016/01/19/ladder-network.html

[4] Marita Cross, Emma Smith, Damian Hoy, Sandra Nolte, Ilana Ackerman, Marlene Fransen, Lisa Bridgett, Sean Williams, Francis Guillemin, Catherine L Hill, Laura L. Laslett, Graeme Jones, Flavia Cicuttini, Richard Osborne, Theo Vos, Rachelle Buchbinder, Anthony Woolf, and Lyn March. 2014. The global burden of hip and knee osteoarthritis: estimates from the Global Burden of Disease 2010 study. *Annals of the Rheumatic Diseases* 73, 7 (2014), 1323–1330. https://doi.org/10.1136/annrheumdis-2013-204763 arXiv:https://ard.bmj.com/content/73/7/1323.full.pdf

[5] Bhushan R. Deshpande, Jeffrey N. Katz, Daniel H. Solomon, Edward H. Yelin, David J. Hunter, Stephen P. Messier, Lisa G. Suter, and Elena Losina. 2016. Number of Persons With Symptomatic Knee Osteoarthritis in the US: Impact of Race and Ethnicity, Age, Sex, and Obesity. *Arthritis Care & Research* 68, 12 (2016), 1743–1750. https://doi.org/10.1002/acr.22897 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acr.22897

[6] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv e-prints*, Article arXiv:1808.06670 (Aug. 2018), arXiv:1808.06670 pages. arXiv:stat.ML/1808.06670

[7] Felson DT, Lawrence RC, Dieppe PA, and et al. 2000. Osteoarthritis: New insights. part 1: the disease and its risk factors. *Annals of Internal Medicine* 133, 8 (2000), 635–646. https://doi.org/10.7326/0003-4819-133-8-200010170-00016 arXiv:/acp/content$_{p}ublic/journal/aim/19967/0000605 - 200010170 - 00016.pdf$

[8] Gabby B. Joseph, Charles E. McCulloch, Michael C. Nevitt, Jan Neumann, Alexandra S. Gersing, Martin Kretzschmar, Benedikt J. Schwaiger, John A. Lynch, Ursula Heilmeier, Nancy E. Lane, and Thomas M. Link. 2018. Tool for osteoarthritis risk prediction (TOARP) over 8 years using baseline clinical data, X-ray, and MRI: Data from the osteoarthritis initiative. *Journal of Magnetic Resonance Imaging* 47, 6 (2018), 1517–1526. https://doi.org/10.1002/jmri.25892 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.25892

[9] Gabby B. Joseph, Charles E. McCulloch, Michael C. Nevitt, Jan Neumann, Alexandra S. Gersing, Martin Kretzschmar, Benedikt J. Schwaiger, John A. Lynch, Ursula Heilmeier, Nancy E. Lane, and Thomas M. Link. 2018. Tool for osteoarthritis risk prediction (TOARP) over 8 years using baseline clinical data, X-ray, and MRI: Data from the osteoarthritis initiative. *Journal of Magnetic Resonance Imaging* 47, 6 (2018), 1517–1526. https://doi.org/10.1002/jmri.25892 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.25892

[10] Abhishek Kadian. 2017. ladder. https://github.com/abhiskk/ladder

[11] J. H. Kellgren and J. S. Lawrence. 1957. Radiological Assessment of Osteo-Arthrosis. *Annals of the Rheumatic Diseases* 16, 4 (1957), 494–502. https://doi.org/10.1136/ard.16.4.494 arXiv:https://ard.bmj.com/content/16/4/494.full.pdf

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[13] L. Menashe, K. Hirko, E. Losina, M. Kloppenburg, W. Zhang, L. Li, and D.J. Hunter. 2012. The diagnostic performance of MRI in osteoarthritis: a systematic review and meta-analysis. *Osteoarthritis and Cartilage* 20, 1 (2012), 13 – 21. https://doi.org/10.1016/j.joca.2011.10.003

[14] Berk Norman, Valentina Pedoia, and Sharmila Majumdar. 2018. Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. *Radiology* 288, 1 (2018), 177–185. https://doi.org/10.1148/radiol.2018172322 arXiv:https://doi.org/10.1148/radiol.2018172322 PMID: 29584598.

[15] Bellamy NW, W.W. Buchanan, Charlie Goldsmith, J Campbell, and Larry Stitt. 1989. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in

patients with osteoarthritis of the hip or knee. *The Journal of rheumatology* 15 (01 1989), 1833–40.

[16] C.G. Peterfy, E. Schneider, and M. Nevitt. 2008. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and Cartilage* 16, 12 (2008), 1433 – 1441. https://doi.org/10.1016/j.joca.2008.06.016

[17] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. 2015. Semi-Supervised Learning with Ladder Network. *CoRR* abs/1507.02672 (2015). arXiv:1507.02672 http://arxiv.org/abs/1507.02672

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR* abs/1505.04597 (2015). arXiv:1505.04597 http://arxiv.org/abs/1505.04597

[19] L. Sharma, M. Hochberg, M. Nevitt, A. Guermazi, F. Roemer, M.D. Crema, C. Eaton, R. Jackson, K. Kwoh, J. Cauley, O. Almagor, and J.S. Chmiel. 2017. Knee tissue lesions and prediction of incident knee osteoarthritis over 7 years in a cohort of persons at higher risk. *Osteoarthritis and Cartilage* 25, 7 (2017), 1068 – 1075. https://doi.org/10.1016/j.joca.2017.02.788

[20] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556 http://arxiv.org/abs/1409.1556