# Disentanglement of Latent Factors of Radiographic Knee Osteoarthritis Severity Classification

Seung-woo Ko[1]

[1]Korea International School, Jeju Campus

## ABSTRACT

Knee osteoarthritis is the most common form of the musculoskeletal disorder that usually happens to the elderly. The diagnosis and treatment for knee osteoarthritis are causing a huge economic burden to society. Traditionally, the diagnosis for knee osteoarthritis was done by analyzing MRI (Magnetic Resonance Imaging). However, this method has problems in the way that it is costly and has limited access since they are only available in specialized medical institutions. There have been many research studies that attempt to use x-ray images that are safe, cost-efficient, and commonly available. Their method often fails and shows poor results due to a lack of training dataset. In this paper, I proposed a novel autoencoder-based knee osteoarthritis classification system. Unlike the previous deep learning-based research, the proposed method disentangles the osteoarthritis-related latent factors from knee x-ray images. These latent factors are then trained to predict the severity of osteoarthritis. The proposed method achieves accuracy of 78.9% on the publicly available dataset. Our method produces accurate results without having a large dataset while successfully avoiding the curse of dimensionality. Throughout the comprehensive experiment, I have shown that the proposed method outperforms the existing state-of-the-art methods by a great accuracy margin.

## Introduction

Knee osteoarthritis is well known as one of the common musculoskeletal disorders. It causes not only physical inconvenience but also a big economic burden to society. It has been reported that the average lifetime treatment cost of knee osteoarthritis is €19,000/year [1]. The reason why it is expensive is due to the lack of a systematic and efficient diagnosis system. The current diagnosis method that is widely being used is analyzing MRI (Magnetic Resonance Imaging). However, this is inefficient due to its high cost and poor accessibility as MRI devices are often located in specialized medical institutions. Modern society is an aging era, in which elderly people increase, and Knee Osteoarthritis is a common disorder that mostly happens to the elderly. However, the elderly have a lack of access to treatment due to aforementioned problems.

Radiographs or x-ray images are not only cost-efficient but also have better accessibility compared to MRI. In this circumstance, there have been many researches that attempt to utilize the machine learning approach to the radiographs to diagnose knee osteoarthritis. Their methods have shown that it is feasible to apply convolutional neural networks to classify the severity of osteoarthritis. These methods yield poor results due to the lack of knee osteoarthritis dataset. The accuracy of the learning-based algorithm is heavily based on the amount of data samples. However, it is practically difficult to obtain such x-ray images for knee osteoarthritis and the labeling process is very time consuming.

In order to solve this problem, I proposed a novel autoencoder-based knee osteoarthritis classification system. I separate the training process into two steps of manifold learning and classifier training. By finding the latent space of the knee x-ray images, the proposed system successfully classifies the severity of osteoarthritis. I also propose a novel data augmentation to make the trained model see a wide range of sample characteristics during the training process. The proposed method achieves accuracy of 78.9% on the knee osteoarthritis dataset which is publicly available online.

## Objective

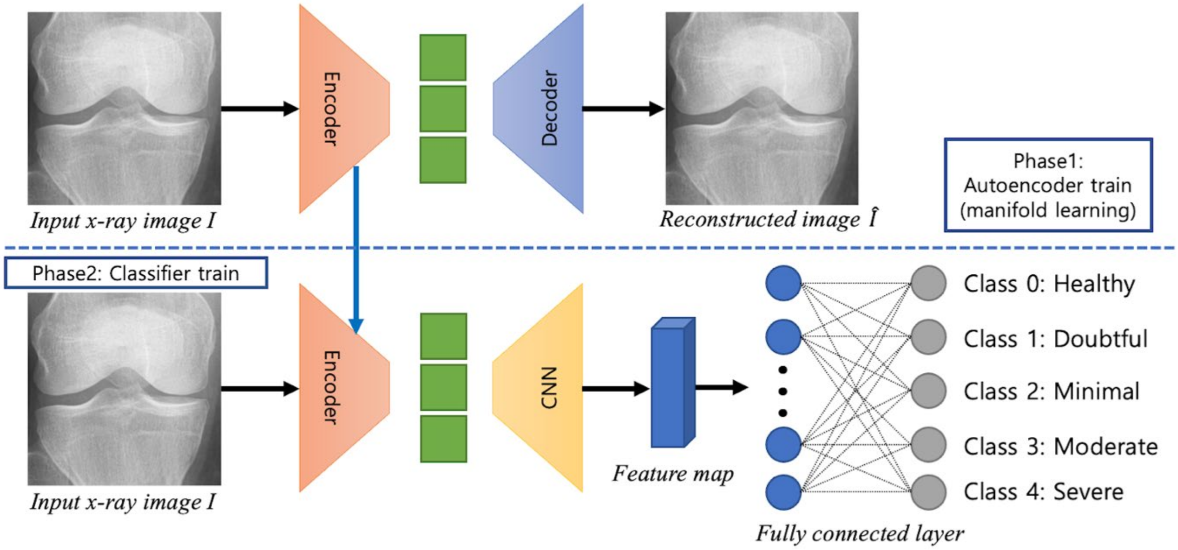

Fig. 1. Overall architecture of the proposed system.

Figure 1 shows the flow chart of the proposed system. The training process of the proposed system is divided into two phases. In the first phase, I conduct manifold learning to find important knee osteoarthritis-related latent space using autoencoder architecture composed of encoder-decoder networks. The purpose of the first phase is to train the encoder as a robust and accurate feature extractor without having a large dataset by avoiding the curse of dimensionality. In the second phase, with a pre-trained encoder network, the system performs classification training to classify diagnosis categories about knee osteoarthritis. In conclusion, the proposed system *F* takes *I* as input and predicts class score vector $\mathrm{P} = \{C_0, C_1, C_2, C_3, C_4\}$. Where $C_0, C_1, C_2, C_3,$ and $C_4$ are assigned to one of a fixed set of diagnosis categories as shown in Fig 1. I classify the input x-ray image *I* to $C_K$ which is the maximum value in the class score vector. I define the classification process as $F(I) = P$. The detailed process of the first and second phase is explained in chapter 3.

## Method

In general, it is necessary to acquire a large volume of datasets to make the trained model perform well. However, practically, it is very difficult to obtain a sufficient amount of knee osteoarthritis samples due to its restriction. Thus, the density of the collected knee osteoarthritis dataset is relatively low and it causes the curse of dimensionality problem which degrades the accuracy of the trained model. To avoid the curse of dimensionality problem, I conduct manifold learning which works as a dimensionality reducer in many computer vision fields [2-4]. By applying a manifold learning scheme to the proposed system, the trained network is able to extract the important knee osteoarthritis disease-related feature in low dimensions. This feature is fed to the classifier in future steps.

As aforementioned, I first conduct manifold learning to find important knee osteoarthritis-related latent space. For the architecture I exploit the autoencoder which is often composed with a encoder and a decoder. The encoder takes input knee x-ray image *I* and produces latent variable *z*. Given the input latent variable *z*, the decoder outputs the reconstructed knee x-ray image $\hat{\imath}$. I define an encoder $E : I \rightarrow z$ and a decoder $D : z \rightarrow \hat{\imath}$.

In the second phase, I set the pre-trained encoder $E$ to freeze and exploit it as a feature extractor. The CNN (Convolutional Neural Network) takes the extracted knee osteoarthritis feature as input and produces the feature map. Finally, this feature map is fed to the fully connected layer and class score vector $P$ is predicted.

## 3.1 Architecture

To develop the autoencoder architecture, I exploit UNet [5] that has comparable performance in various computer vision tasks. I remove their skip-connection operations that transfer the features extracted from the encoder to the decoder. I empirically found that this removal yields better accuracy. For CNN, I choose Resnet [6] as the backbone and single linear layer for the proposed fully connected layer. Especially, I use Resnet18 among other variants as its layer depth is deep enough to yield comparable results.

## 3.2 Loss Function

To train the proposed system I use two types of loss functions. For the autoencoder architecture used in the first phase, I calculate the pixel-wise difference between the original input x-ray image and the reconstructed image.

Eq. (1) shows how this operation is calculated.

$$L_{Phase1} = |i - \hat{\mathrm{i}}|_1$$

Where $I$ and $\hat{\imath}$ denote the original input image and reconstructed image, respectively. Concretely, if the decoder reconstructs the $\hat{\imath}$ to be identical as $I$, the $L_{Phase1}$ becomes 0 which means there is no penalty to the network during the train process. Thus, $L_{Phase1}$aims to make the decoder to reconstruct image as same as possible to the its input image while encoder produces useful knee-related feature for decoder. For the knee osteoarthritis classifier, I use cross-entropy loss function which often used to train classifier models in general [7, 8].

Eq. (2) shows how the cross-entropy loss is being calculated.

$$L_{Phase2} = -P\log\hat{P}$$

## 3.3 Data Augmentation

Data augmentation is often applied to the training sample to allow the trained model to see a wide range of sample characteristics during the training process. In general, random translation and horizontal flip are often used for classification trains. However, these augmentations degrade the accuracy of the trained model. I assume there is no location-related variance in the training dataset as the knee x-ray image is obtained in a very restricted environment. Instead, there is a huge pixel intensity variance in the dataset as shown in Fig. 2.

As their pixel distribution diverse, the trained network often fails to predict accurate results. To solve this problem, I proposed a novel intensity manipulation augmentation technique. The proposed data augmentation method randomly manipulates the brightness of input images. The effectiveness of the proposed data augmentation is explained in detail in chapter 4.3.

# Results and Discussion

## 4.1 Dataset

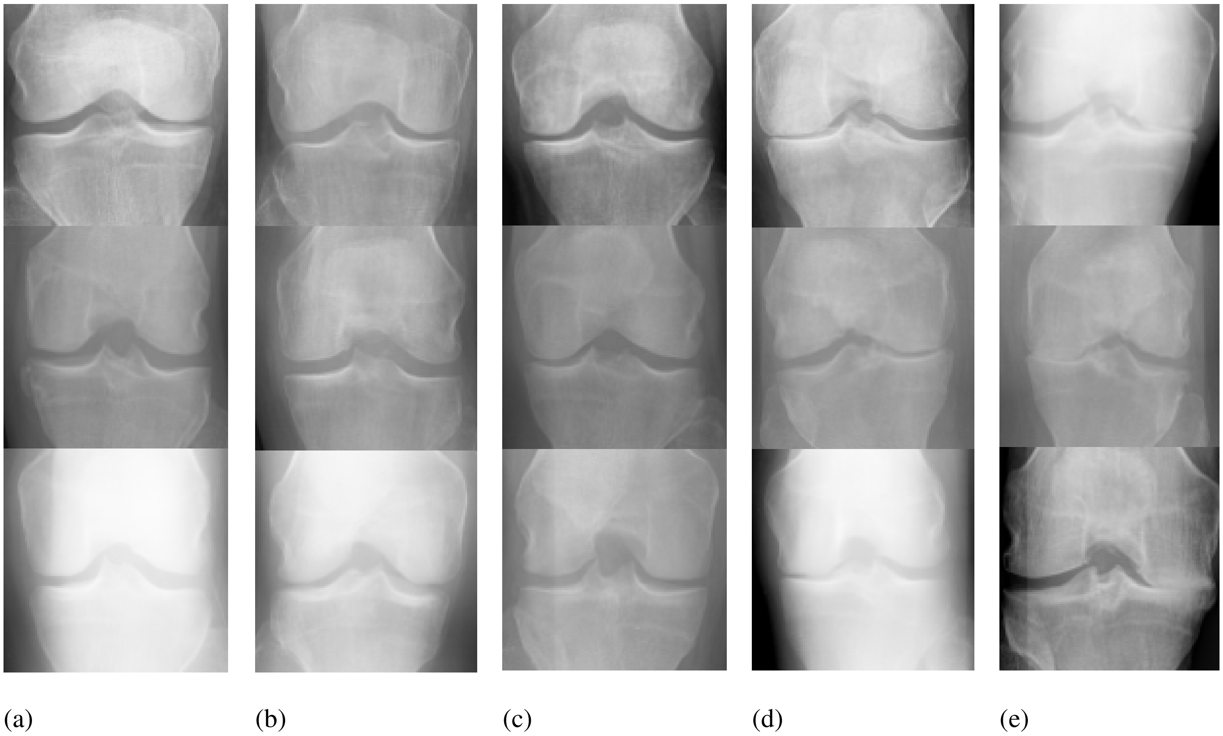

Fig. 2. Example of samples for each category (a) Class 0 (healthy knee), (b) class 1 (doubtful), (c) class 2 (minimal), (d) class 3 (moderate), and (e) class 4 (severe case).

Fig. 2 showcases the images for each category in the knee osteoarthritis dataset [9] I used in this research. The dataset is publicly available online. The dataset consists of 9,786 samples and they are labeled based on 5 different classes; Class 0 to 4. Images that are assigned to class 0 are healthy knee images, while images in class 1 are doubtful images and the images in class 2 have minimal problems. The images in class 3 have moderate amounts of problems and the images in class 4 have severe problems that require treatment. As different individuals have different amounts of problems in each of their knees, the dataset contains a wide range of variety. As shown in Fig. 2, images have a diverse intensity distribution due to various device characteristics and different x-ray exposure time. This variety caused the classification task to be more challenging. For accurate experiments, I split the dataset into trainset and testset individually. Trainset and testset occupies 90% and 10% of the total data samples, respectively.

## 4.2 Comparison with Previous Research

Table 1. Comparison with previous research

| **Years Inducted** | **Accuracy (%)** |
|---|---|
| AlexNet [10] | 64.7 |
| VGGNet [11] | 68.1 |
| Resnet [6] | 72.5 |
| Ours | 78.9 |

Table 1 shows the result of the comparison with the previous state-of-the-art methods. The purpose of this comparison is to identify the difference between the accuracy of the proposed method and the existing methods. This comparison verifies whether or not the proposed method outperforms the other methods. As shown in Table 1, the Resnet [6] achieves accuracy of 72.5%, while the proposed method achieves accuracy of 78.9%. This result clearly shows that the proposed method outperforms the existing Resnet with a performance gap of 6.4%.

Compared to the VGGNet [11] which achieves accuracy of 68.1%, the proposed method produces greater accuracy by achieving 10.8% more accuracy. Last comparison method AlexNet [10] achieves 14.2% smaller accuracy compared to the proposed method.

In Conclusion, the proposed method surpasses all of the existing state-of-the-art methods while having a great performance gap. I attribute this to the manifold learning in the proposed training pipeline. This process allows the trained model to consistently extract robust knee osteoarthritis related features without having a large dataset. Also, the proposed novel data augmentation technique helps the model with dealing with the input images that are affected by noises. I will explain how each proposed method affects the final performance in chapter 4.3. in detail.

### 4.3 Ablation Study

Table 2. Ablation study result

| Ablation Model | Accuracy (%) |
|---|---|
| Ablation model 1 (trained without manifold learning) | 72.5 |
| Ablation model 2 (trained without proposed data augmentation) | 77.1 |
| Proposed method | 78.9 |

In this chapter, I conduct an ablation study to examine how each proposed method affects the final accuracy of the full model by comparing the accuracy of each ablation model. Firstly, I train the full model using both manifold learning and proposed data augmentation techniques. To verify the effectiveness of manifold learning and the proposed data augmentation technique, I train two ablation models without each proposed method.

To train the first ablation model (ablation model 1), I removed the manifold learning phase from the proposed train pipeline. As shown in table 2, the first ablation model achieves an accuracy of 72.5% which degrades 6.4% from the proposed full model. This result clearly shows that the proposed manifold learning pipeline successfully helps the trained model to produce better results without having a large dataset by avoiding the curse of dimensionality.

The second ablation model (ablation model 2) is trained without the proposed data augmentation. It achieves an accuracy of 77.1%. The observed accuracy gap for this comparison is 1.8%. It shows the proposed data augmentation contributes the performance boost by enforcing the trained model perform more robustly against a wide range of intensity knee x-ray images.

## Conclusion

In this research, I proposed a novel autoencoder-based knee osteoarthritis classification system. To avoid the curse of dimensionality, I separate the training process into two phases. First training phase aims to train an encoder as a good knee-related feature extractor. With the pre-trained encoder, the CNN and fully connected layer is trained to predict the severity of osteoarthritis. The proposed system yields accurate results without having a large dataset. Throughout the comprehensive experiments, I have shown that the proposed method outperforms the previous state-of-the-art methods with a noticeable performance gap. I also conducted an ablation study to examine how each proposed method contributes to the performance boost. In conclusion, the proposed method achieved accuracy of 78.9% on the knee osteoarthritis dataset which is publicly available online. In the future, I plan to apply attention mechanisms to the proposed system to obtain more accurate results.

## References

Losina, Elena, et al. "Lifetime medical costs of knee osteoarthritis management in the United States: impact of extending indications for total knee arthroplasty." *Arthritis care & research* 67.2 (2015): 203-215.

Izenman, Alan Julian. "Introduction to manifold learning." *Wiley Interdisciplinary Reviews: Computational Statistics* 4.5 (2012): 439-446.

Huo, Xiaoming, Xuelei Sherry Ni, and Andrew K. Smith. "A survey of manifold-based learning methods." *Recent advances in data mining of enterprise data* (2007): 691-745.

Pless, Robert, and Richard Souvenir. "A survey of manifold learning for images." *IPSJ Transactions on Computer Vision and Applications* 1 (2009): 83-94.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention.* Springer, Cham, 2015.

He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Lu, Dengsheng, and Qihao Weng. "A survey of image classification methods and techniques for improving classification performance." *International journal of Remote sensing* 28.5 (2007): 823-870.

Nath, Siddhartha Sankar, et al. "A survey of image classification methods and techniques." *2014 International conference on control, instrumentation, communication and computational technologies (ICCICCT)*. IEEE, 2014.

Chen, Pingjun (2018), “Knee Osteoarthritis Severity Grading Dataset”, Mendeley Data, V1

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).