



### 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원 저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리와 책임은 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)



의학박사 학위논문

Transfer Learning Strategies of Deep  
Convolutional Neural Networks for Osteoporosis  
Screening in Dental Panoramic Radiographs

치과용 파노라마 방사선 사진에서 골다공증 선별을 위한  
심층 합성곱 신경망(deep CNN)의 전이학습 전략

2020년 8월

서울대학교 대학원  
의학과 의공학전공  
이기선

# Abstract

Osteoporosis is a metabolic bone disease characterized by low bone mass and disruption in bone micro-architecture. Clinical diagnostic methods for osteoporosis are expensive and therefore have limited availability in population. Recent studies have shown that Dental Panoramic Radiographs (DPRs) can provide the bone density change clues in bone structure analysis. This study aims to evaluate the discriminating performance of deep convolutional neural networks (CNNs), employed with various transfer learning strategies, on the classification of specific features of osteoporosis in DPRs. For objective labeling, we collected a dataset containing 680 images from different patients who underwent both skeletal bone mineral density and digital panoramic radiographic examinations at the Korea University Ansan Hospital between 2009 and 2018. In order to select the backbone convolutional neural network which is the basis for applying the transfer learning, we conducted preliminary experiments on the three convolutional neural networks, VGG-16, Resnet50, and Xception networks, which were frequently used in image classification. Since VGG-16 showed the best AUC value in the classification experiment conducted without transfer learning, the transfer learning using the fine-tuning technique was tested using VGG-16 as the backbone network. In order to find the optimal fine-tuning degree in the VGG-16 network, a total of six fine-tuning applied transfer learning groups were set according to the number of fine-tuning blocks in the VGG-16 with five blocks as follows: A group that does not perform fine-tuning at all (VGG-16-TF0), a group that fine-tunes the last 1 block (VGG-16-TF1), a group that fine-tunes the last 2 blocks (VGG-16-TF2), a group that fine-tunes the last 3 blocks (VGG-16-TF3), a group that fine-tunes the last 4 blocks (VGG-16-TF4), and a group that performs fine-tuning all 5 blocks (VGG-16-SCR). The best performing model (VGG-16-TF2) achieved an overall area under the receiver operating characteristic of 0.858. In this study, transfer learning and optimal fine-tuning improved the performance of a deep CNN for screening

osteoporosis in DPR images. In addition, using the gradient-weighted class activation mapping technique, a visual interpretation of the best performing deep CNN model indicated that the model relied on image features in the lower left and right border of the mandibular. This result suggests that deep learning-based assessment of DPR images could be useful and reliable in the automated screening of osteoporosis patients.

**Keywords:** osteoporosis screening; artificial intelligence; convolutional neural networks; dental panoramic radiographs; deep learning

**Student Number:** 2017-36469

## Table of Contents

Chapter 1. Introduction .....	1
Chapter 2. Materials and Methods .....	8
2.1 Dataset Collection .....	8
2.2 Image Preprocessing .....	9
2.3 Cross validation .....	11
2.4 Back-bone Convolutional Neural Networks .....	14
2.5 Evaluation .....	16
2.6 Visualizing Model Decisions .....	18
Chapter 3. Results .....	19
3.1 Clinical and Demographic Characteristics .....	19
3.2 Back-bone Convolutional Neural Networks .....	20
3.3 Fine-Tuning of Transferred deep CNN .....	22
3.4 Evaluation .....	27
3.5 Visualizing Model Decisions .....	30
Chapter 4. Discussion .....	33
Chapter 5. Conclusion .....	42
References .....	51
Abstract .....	28

## Tables

[Table 1] .....	19
[Table 2] .....	27

## Figures

[Figure 1] .....	3
[Figure 2] .....	6
[Figure 3] .....	10
[Figure 4] .....	13
[Figure 5] .....	15
[Figure 6] .....	20
[Figure 7] .....	21
[Figure 8] .....	23
[Figure 9] .....	25
[Figure 10] .....	28
[Figure 11] .....	29
[Figure 12] .....	31
[Figure 13] .....	32
[Figure 14] .....	37
[Figure 15] .....	39

# Chapter 1. Introduction

Known as the most common systemic bone disease, osteoporosis is characterized by the low bone mineral density (BMD) and the micro-structural deterioration of bone structure, thereby leading to compromised bone strength and, consequently, an increased risk of fracture [1]. Osteoporosis often lead to disorders caused by hip, spine, and wrist fractures that reduce the quality of life of the patient and, in severe cases, increase the risk of mortality [2,3]. With fast population aging and an increase in life expectancy, osteoporosis is increasingly becoming a global public health issue; it has been estimated that more than 200 million people are suffering from osteoporosis [4]. According to recent statistics from the International Osteoporosis Foundation, approximately one in three women over the age of 50 will experience osteoporotic fractures, as will one in five men over the age of 50 [4-7]. Moreover, it is expected that more people will be affected by osteoporosis in the future and, consequently, the rate of osteoporotic fractures will increase [8]. This is because the disease initially develops without any symptoms,

remains undiagnosed due to scarce symptomatology, and its first manifestation is often a low-energy fracture of long bones or vertebrae [9].

Usually, osteoporosis is diagnosed by bone mineral density (BMD) measurements (expressed as a T-score), and using dual-energy X-ray absorptiometry (DXA) is considered as the gold-standard examination for BMD assessment [10,11]. DXA is A technique for scanning bone and measuring BMD. A DXA scanner is a kind of large X-ray machine that produces 2 X-ray beams, each with different energy levels. One beam is high energy while the other is low energy. The amount of x-rays that pass through the bone is measured for each beam. This will vary depending on the density of the bone. Based on the difference between the 2 beams, the bone density can be measured. Most commonly, your BMD test results are compared to the bone mineral density of a healthy young adult, and you are given a T-score. A score of 0 means your BMD is equal to the norm for a healthy young adult. Differences between your BMD and that of the healthy young adult norm are measured in units called standard deviations (SDs). The more standard deviations below 0, indicated as negative numbers, the lower your BMD and the

higher your risk of fracture. A T-score between +1 and -1 is considered normal or healthy. A T-score between -1 and -2.5 indicates that you have low bone mass, although not low enough to be diagnosed with osteoporosis. A T-score of -2.5 or lower indicates that you have osteoporosis. The greater the negative number, the more severe the osteoporosis (Figure 1).

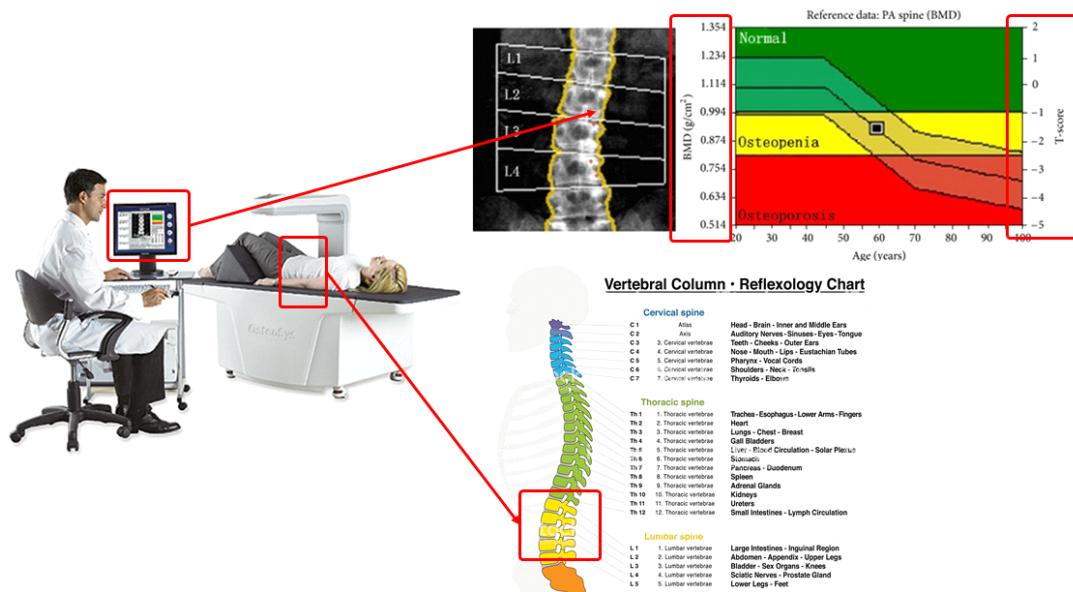


Figure 1. Illustration of an example DXA test and test results screen. DXA test result screen has BMD value and T-score value.

However, this technique is complex, expensive, and the availability is limited for overall population diagnosis [12]. Recently,

digital images of dental panoramic radiographs (DPRs) have been evaluated as cost-effective and important image data for osteoporosis screening. This is because the widespread use of panoramic radiation in dental care for elderly patients with increased life expectancy and a number of studies have demonstrated the feasibility of BMD estimation and osteoporosis screening using panoramic radiography [13–23].

Comparing with DXA, DPRs are a relatively inexpensive and convenient screening method for screening high-risk osteoporosis patients. However, previous approaches primarily relied on manually categorized feature indexes [13–23], such as the Gonion index, mandibular cortical index, mental index, and panoramic mandibular index, and traditional classifier called machine learning (ML) algorithms, such as support vector machine (SVM) [22] and fuzzy classifiers [23], for screening osteoporosis.

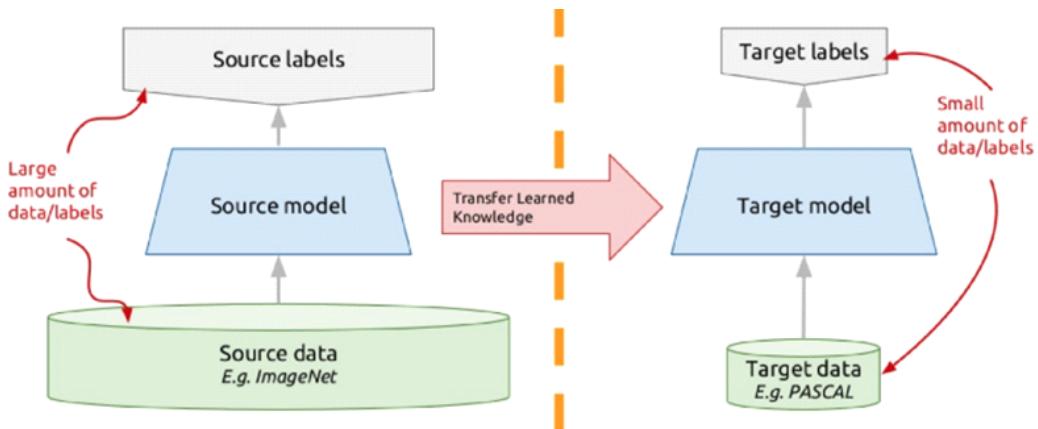
Previously handcrafted feature indices used panoramic radiographs to provide sufficient evidence to support osteoporosis screening, but these methods for distinguishing features are low-order and do not fully characterize heterogeneous patterns in radiographs. In addition, most previous studies require tedious and manual tasks

such as extensive preprocessing, image normalization, and region of interest (ROI) segmentation, which can significantly affect the repeatability of the classification method.

In the last few years, deep learning algorithms, particularly deep convolutional neural networks (CNNs) architecture, have been widely recognized as a reliable approach to learn the classification of the characteristics of features directly from original medical images [24,25]. As opposed to ML approaches that rely on explicitly classified features, deep CNNs are a class of deep neural networks that can learn high dimensional features to maximize the networks ability to discriminate abnormalities among images [26]. There are many different CNN architectures that have been designed to perform image classifications and recognitions. Each of these architectures differ in specific aspects, including the number and size of layers, the connections between these layers, and the overall network depth. Because different network architectures are best suited for different problems, and it is difficult to know in advance which architecture is the right choice for a given task, empirical examination is often recognized as the best way to make these decisions [27].

Although deep CNNs have been recognized as efficient tools

for image classification, they require a large amount of training data, which can be difficult to apply to medical radiographic image data with limited number of images for deep learning training. When the target dataset is significantly smaller than the base dataset, transfer learning is considered a powerful technique for training deep CNNs without overfitting [28,29]. The general process of transfer learning is performed through the use of pretrained models in a two steps as follows: First, copy the first n layers of the pre-trained base network from the regular large data set to the first n layers of the target network. Then, the rest of the layers in the target network are randomly initialized to a small local data set towards the target task to be learned [28] (Figure 2).



**Figure 2.** A Diagram for explaining concept of the transfer learning.

On the basis of the transfer learning techniques, several state-of-the-art results showed outperformance in both general image classification [30–32] and medical image classification [33–36]. However, a few studies have been done to develop and evaluate transfer learning with fine-tuning based deep CNN models for predicting osteoporosis in DPRs.

The aim of this study is to develop and evaluate the deep learning approaches for screening osteoporosis with DPR images. Using the classified panoramic radiograph images based on the BMD value (T-score), this study evaluated several different CNN models based on osteoporosis discriminating accuracy. In addition, we quantitatively evaluated the effect of transfer learning and fine-tuning of a deep CNN model on classifying performance.

## Chapter 2. Materials and Methods

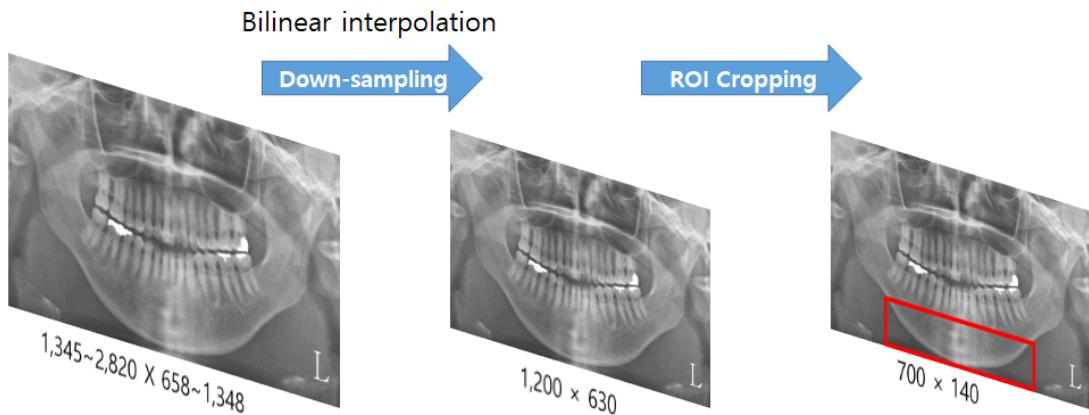
### *2.1 Dataset Collecting*

This study was done on a total of 680 panoramic radiograph images from 680 different patients who visited the Korea University Ansan Hospital. The patients simultaneously underwent skeletal BMD examinations and digital panoramic radiography evaluations within six months, between 2009 and 2018. The subjects were classified into a non-osteoporosis group ( $T\text{-score} \geq 2.5$ ) or osteoporosis group ( $T\text{-score} < 2.5$ ), according to the World Health Organization (WHO) criteria [37]. This criterion has been widely accepted and, in many Member States, provides both a diagnostic and intervention threshold. In this study, collected dataset were divided into which 380 and 300 subjects were assigned, respectively. This study protocol was approved by the institutional review board of the Korea University Ansan Hospital (no. 2019AS0126).

## ***2.2 Image Preprocessing***

The dimensions of the collected dental X-ray images varied from 1,348 to 2,820 pixels in width and 685 to 1,348 pixels in height. For consistency of image preprocessing, the images were down-sampled to a uniform size of  $1,200 \times 630$  pixels, using bilinear interpolation. The final ROI was restricted to the lower part of the mandible, below the teeth-containing alveolar bone, for an image size of  $700 \times 140$  pixels (Figure 3).

This included the most ROI areas of previous studies [13–23] that applied various classification techniques by detailed and specifically indexing the image feature characteristics of the limited small region of mandible. By setting the ROI to include most of the mandible instead of the specific area of it, this study evaluated the area that plays the most distinctive role in osteoporosis classification through explainable deep learning techniques.



**Figure 3.** Image preprocessing for this study. The original DPRs were down-sampled, and the ROI is restricted to the mandibular region below the teeth (region inside the bounding box). DPR, dental panoramic radiograph; ROI, region of interest.

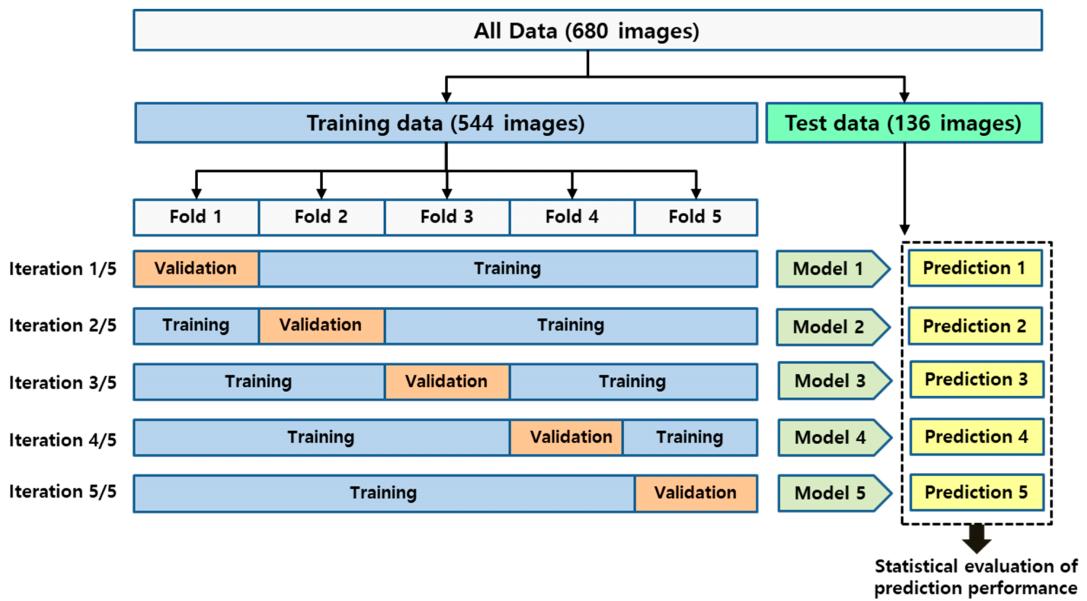
### *2.3. Cross-Validation*

To test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset, this study employed a cross validation technique.

The dataset was divided into training and test sets as follows: The radiographs were selected randomly, and 136 radiographs (20% of the total), 68 each from the osteoporosis and non-osteoporosis groups, were set aside as a test set. This ensured that the testing data set only contained images of novel radiographs that had not been encountered by the model during training. The remaining 544 radiographs were used for the training and validation set.

The 544 images selected as the training dataset were randomly divided into five folds. This was done to perform 5-fold cross validation to evaluate the model training, while avoiding overfitting or bias [39]. Within each fold, the dataset was partitioned into independent training and validation sets, using an 80 to 20 percentage split. The selected validation set was a completely

independent fold from the other training folds and it was used to evaluate the training status during the training. After one model training step was completed, the other independent fold was used as a validation set and the previous validation set was reused, as part of the training set, to evaluate the model training. The 544 images selected as the training dataset were randomly divided into five folds. This was done to perform 5-fold cross validation to evaluate the model training, while avoiding over-fitting or bias [39]. Within each fold, the dataset was partitioned into independent training and validation sets, using an 80 to 20 percentage split. The selected validation set was a completely independent fold from the other training folds and it was used to evaluate the training status during the training. After one model training step was completed, the other independent fold was used as a validation set and the previous validation set was reused, as part of the training set, to evaluate the model training. An overview of the 5-fold cross validation performed in this study is presented in Figure 4.



**Figure 4.** The overview of the performed 5-fold cross validation in this study.

## ***2.4. Back-bone Convolutional Neural Networks***

In order to select the backbone convolutional neural network which is the basis for applying the transfer learning, we conducted preliminary experiments on the three convolutional neural networks, VGG-16, Resnet50, and Xception networks, which were frequently used in image classification (Figure 5). After comparing the classification ability with the AUC value using the three deep CNNs mentioned above, transfer learning with fine-tuning technique applied to the deep CNN model with high AUC value.



**Figure 5.** Schematic diagrams of the three convolutional neural networks (CNN) architectures examined as backbone CNN architecture in this study.

## *2.5. Evaluation*

All process of this study were performed under a 64-bit Windows 10 operating system, with i9-9900K CPU, 64 GB memory and an NVIDIA Quadro RTX8000 GPU. Building, training, validation, and prediction of deep learning models were performed using the Keras (v.2.3.1) [40] library and TensorFlow-GPU (v113.1) [41] backend engine.

The evaluation of the screening performance of the CNN models was performed with the independent test dataset in each cross-validation fold. To comprehensively evaluate the screening performance on the test dataset, the accuracy, sensitivity, specificity, receiver operating characteristic (ROC) curve, and precision recall (PR) curve were calculated. The accuracy, sensitivity, and specificity score can be calculated as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

TP and FP are the number of correctly and incorrectly predicted images, respectively. Similarly, TN and FN represent the number of correctly and incorrectly predicted images, respectively. The area under the ROC curve (AUC) was also calculated.

## *2.6. Visualizing Model Decisions*

Deep learning models have often been referred to as non-interpretable black boxes because it is difficult to know the process by which they make predictions. To know the decision-making process of the model, and which features are most important for the model to screen osteoporosis in DPR images, this study employed the gradient-weighted class activation mapping technique (Grad-CAM) [42] and the most significant regions for screening osteoporosis in DPR images were highlighted.

## Chapter 3. Results

### *3.1. Clinical and Demographic Characteristics of the Subjects*

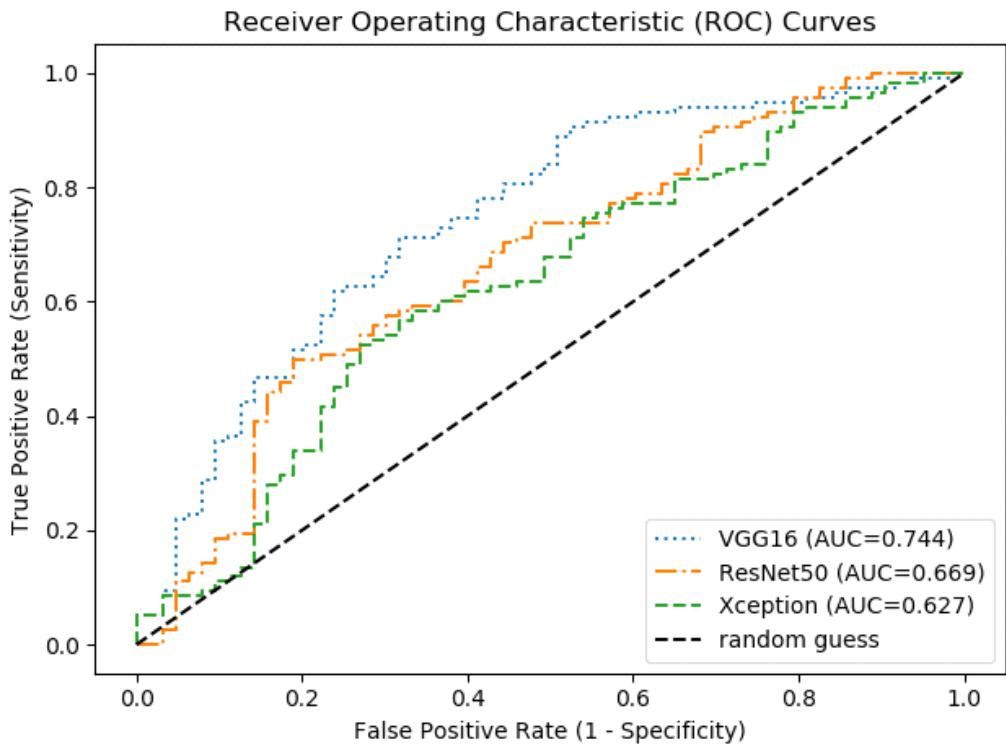
The patients were 565 female and 115 male, with an age range from 27 to 90 years (mean age of 63.0 years). There were 380 patients (mean age 58.5) without osteoporosis ( $T\text{-score} \geq -2.5$ ) and 300 patients (mean age 68.6) with osteoporosis ( $T\text{-score} < -2.5$ ). The clinical characteristics of the DPR dataset used in this study are summarized in Table 1.

**Table 1.** Clinical and demographic characteristics of the dental panorama radiographs (DPRs) dataset in this study.

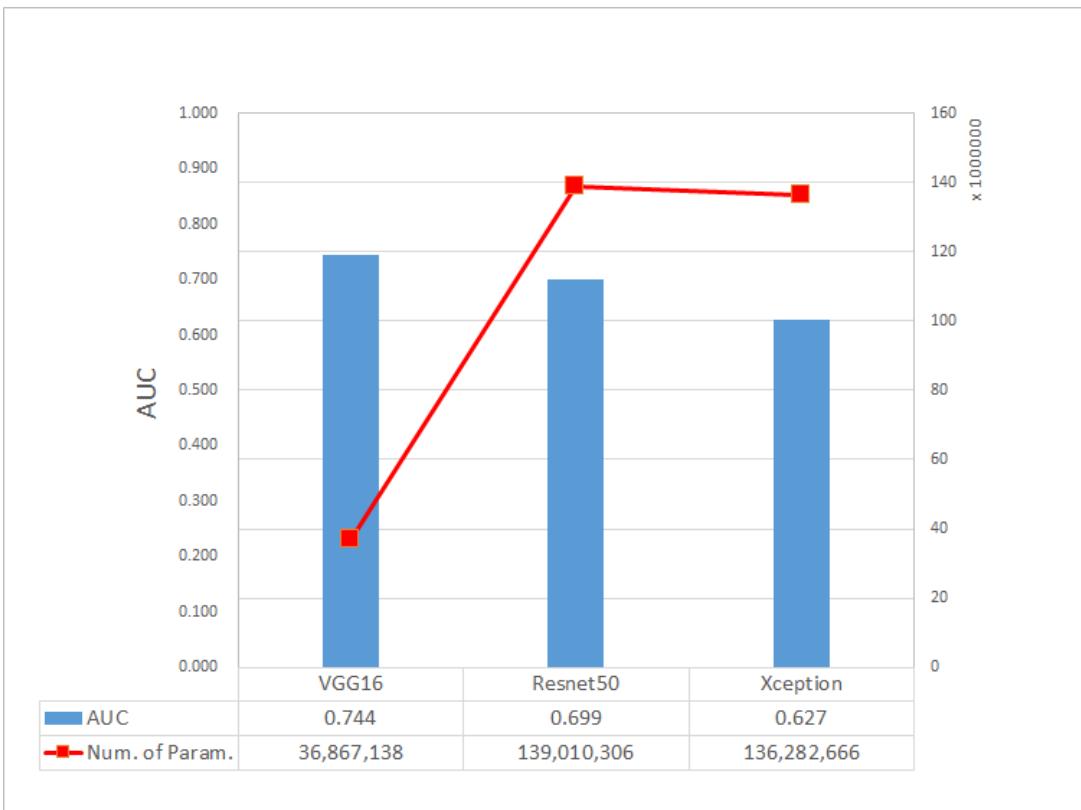
Parameter	Without osteoporosis ( $T\text{-Score} \geq -2.5$ )	With osteoporosis ( $T\text{-Score} < -2.5$ )	Total
Number of patients	380	300	680
Number of female / male	332/48	233/67	565/115
Mean age ( $\pm$ SD)	58.5 ( $\pm 11.8$ )	68.4 ( $\pm 8.4$ )	63.0 ( $\pm 11.6$ )

### 3.2. Back-bone network selection

The predicting results of osteoporosis using three backbone network candidates (VGG-16, Resnet-50, and Xception) without transfer learning were comparatively evaluated by AUC. VGG-16 had the highest AUC value of 0.745, followed by Resnet50 with AUC value of 0.669, and Xception with AUC value of 0.627 (Figure 6 and 7).



**Figure 6.** Mean ROC curves of three deep CNN models selected as candidates for back-bone network of screening osteoporosis on DPR images in this study.



**Figure 7.** Comparison graph of mean ROC values and numbers of trainable parameters of three depp CNN models selected as candidates for backbone network of screening osteoporosis on DPR images in this study.

### ***3.3. Fine-tuning with transferred deep CNN***

Since VGG-16 showed the best AUC value in the classification experiment conducted without transfer learning, the various transfer learning strategies with different fine-tuning degree were employed to VGG-16. In order to find the optimal fine-tuning degree in the VGG-16 network, a total of six fine-tuning applied transfer learning groups were set according to the number of fine-tuning blocks in the VGG-16 having five blocks as follows: A group that does not perform fine-tuning at all (VGG16-TF0), a group that fine-tunes the last 1 block (VGG16-TF1), a group that fine-tuning the last 2 blocks (VGG16-TF2), a group that fine-tuning the last 3 blocks (VGG16-TF3), a group that fine-tuning the last 4 blocks (VGG16-TF4), and a group that performs fine-tuning all 5 blocks (VGG16-TF5). The preceding architectures, along with the six variant deep CNN models (VGG16-TF0, VGG16-TF1, VGG16-TF2, VGG16-TF3, VGG16-TF4, and VGG16-TF5) used in this study, are depicted in the block diagram in Figure 8, and 9.

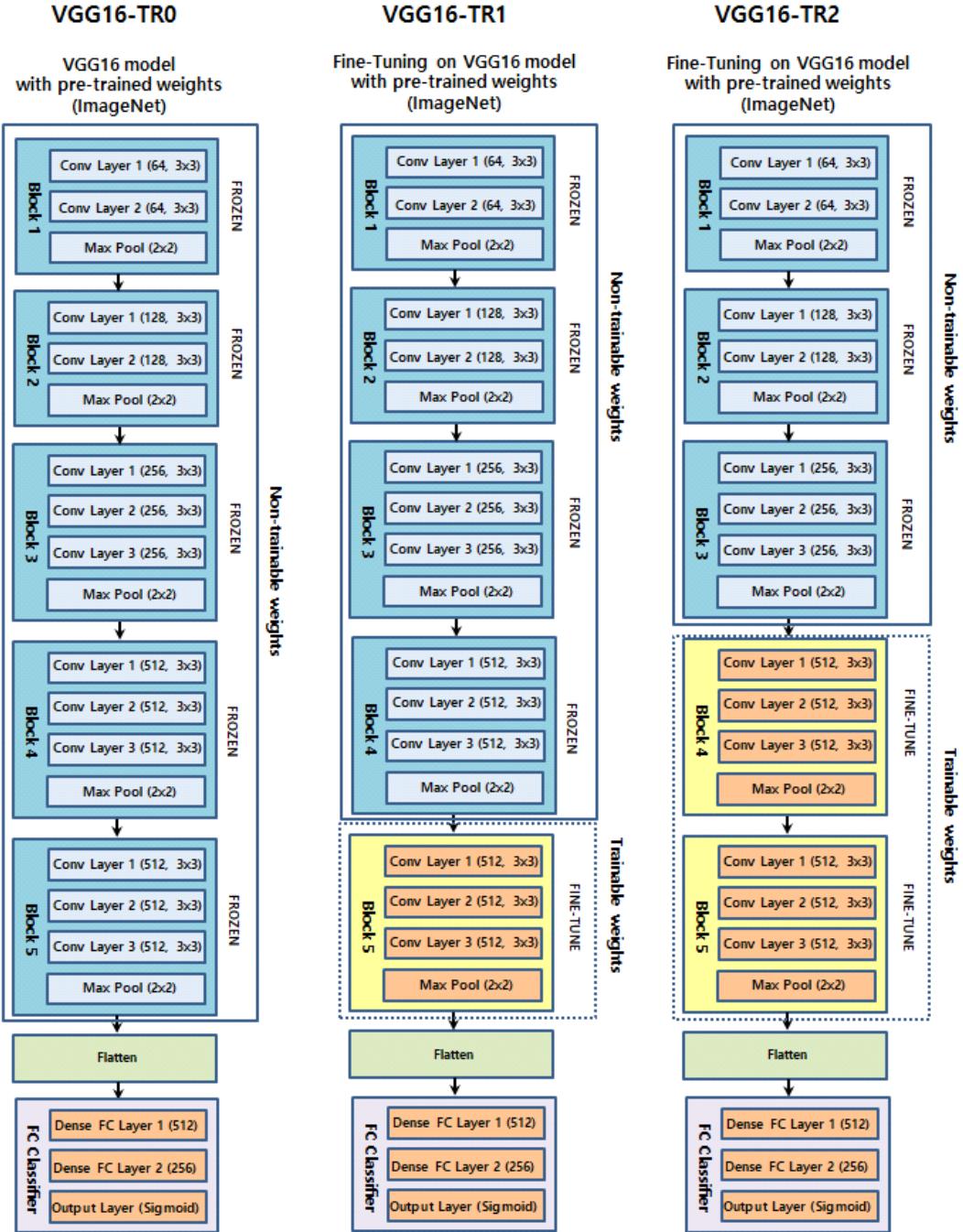


Figure 8(continue).. Six variant deep CNN models (VGG-16-TF0, VGG-16-TF1, VGG-16-TF2, VGG-16-TF3, VGG-16-TF4, and VGG-16-TF5) used in this study (continue).

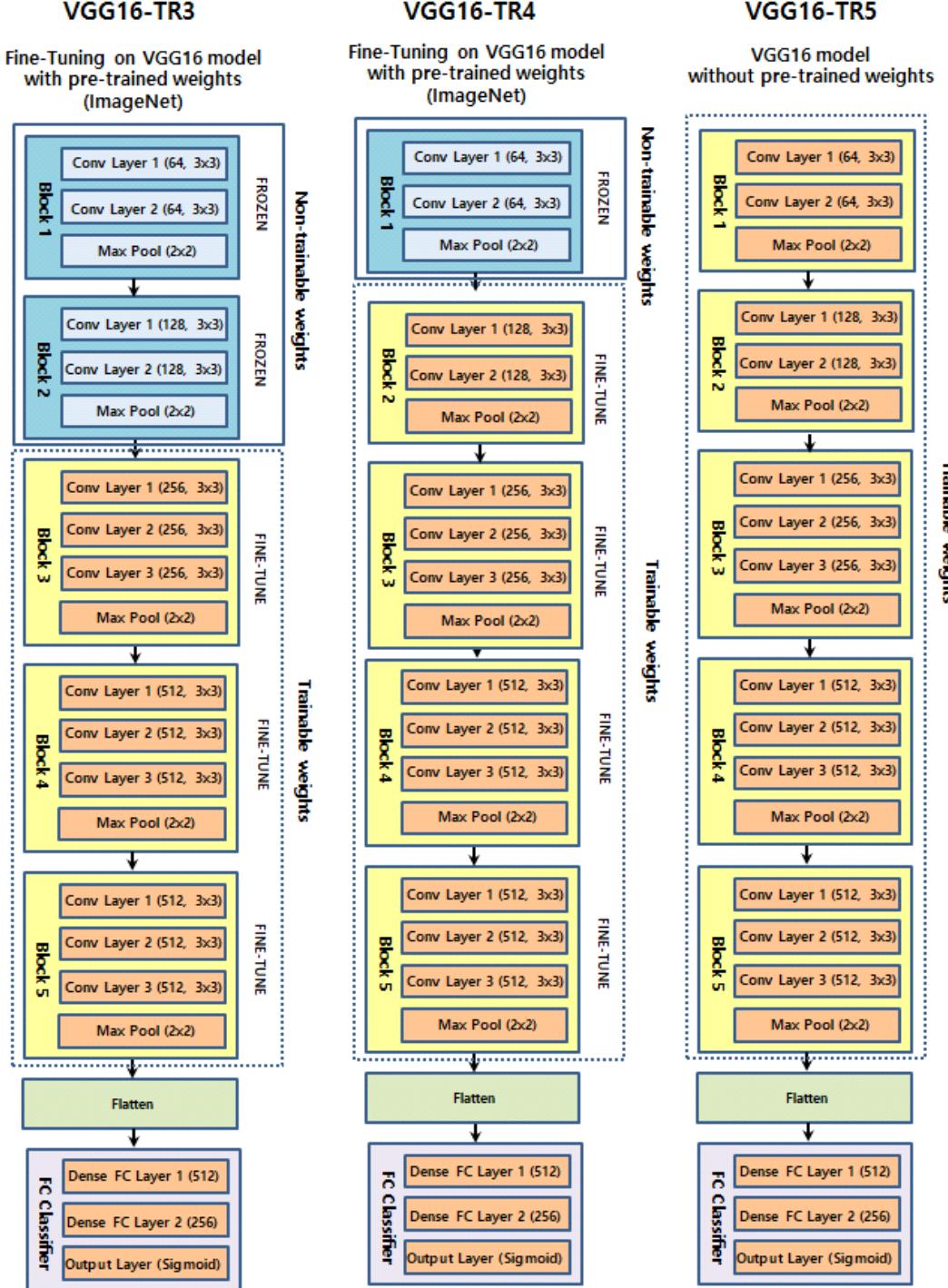


Figure 8. Six variant deep CNN models (VGG-16-TF0, VGG-16-TF1, VGG-16-TF2, VGG-16-TF3, VGG-16-TF4, and VGG-16-TF5) used in this study.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 140, 700, 3)	0
block1_conv1 (Conv2D)	(None, 140, 700, 64)	1792
block1_conv2 (Conv2D)	(None, 140, 700, 64)	36928
block1_pool (MaxPooling2D)	(None, 70, 350, 64)	0
block2_conv1 (Conv2D)	(None, 70, 350, 128)	73856
block2_conv2 (Conv2D)	(None, 70, 350, 128)	147584
block2_pool (MaxPooling2D)	(None, 35, 175, 128)	0
block3_conv1 (Conv2D)	(None, 35, 175, 256)	295168
block3_conv2 (Conv2D)	(None, 35, 175, 256)	590080
block3_conv3 (Conv2D)	(None, 35, 175, 256)	590080
block3_pool (MaxPooling2D)	(None, 17, 87, 256)	0
block4_conv1 (Conv2D)	(None, 17, 87, 512)	1180160
block4_conv2 (Conv2D)	(None, 17, 87, 512)	2359808
block4_conv3 (Conv2D)	(None, 17, 87, 512)	2359808
block4_pool (MaxPooling2D)	(None, 8, 43, 512)	0
block5_conv1 (Conv2D)	(None, 8, 43, 512)	2359808
block5_conv2 (Conv2D)	(None, 8, 43, 512)	2359808
block5_conv3 (Conv2D)	(None, 8, 43, 512)	2359808
block5_pool (MaxPooling2D)	(None, 4, 21, 512)	0
flatten_1 (Flatten)	(None, 43008)	0
dense_1 (Dense)	(None, 512)	22020608
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 256)	131328
dropout_2 (Dropout)	(None, 256)	0
visualized_layer (Dense)	(None, 2)	514
<hr/>		
Total params: 36,867,138		
Trainable params: 35,131,650		
Non-trainable params: 1,735,488		

Figure 9. Output result of `model.summary()` of the transferred VGG16-TR2 network where the parameters of the last two blocks are set to trainable.

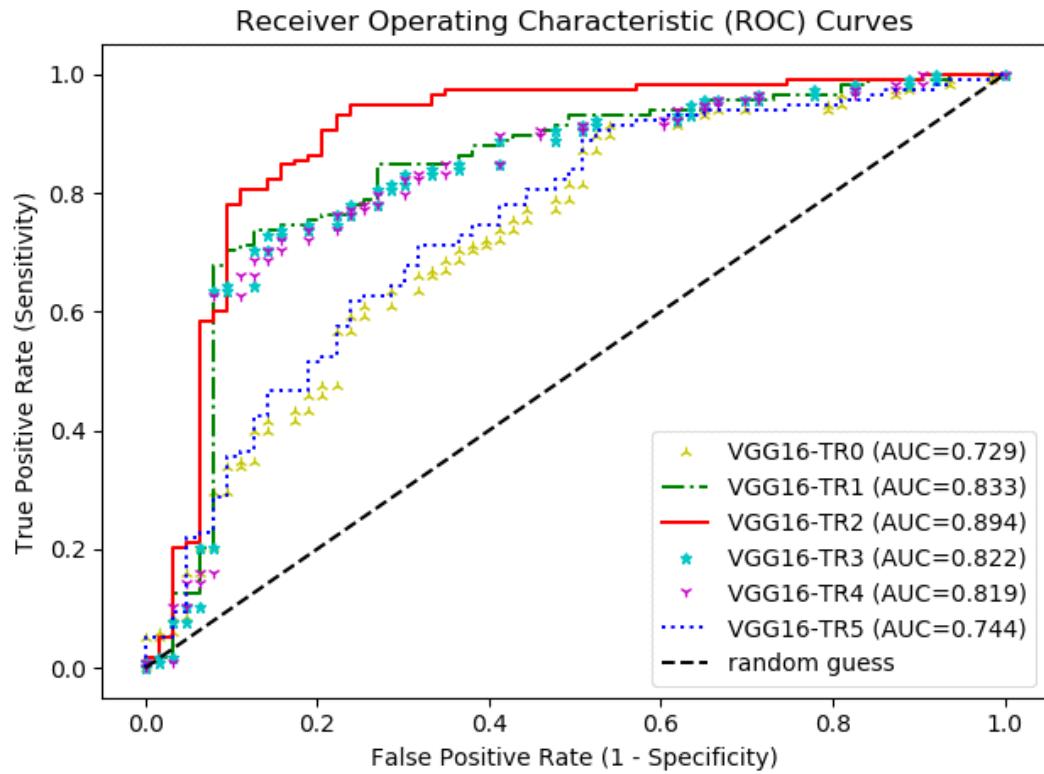
In the transfer learning with fine-tuning version, model weights were initialized based on pre-training on a general image dataset, except that some of the last blocks were unfrozen so that their weights were updated in each training step. In this study, all the transfer learning version models employed pre-trained weights using the ImageNet database [38]. ImageNet is an image dataset containing thousands of different objects used to train and evaluate image classification models.

### **3.4. Evaluation**

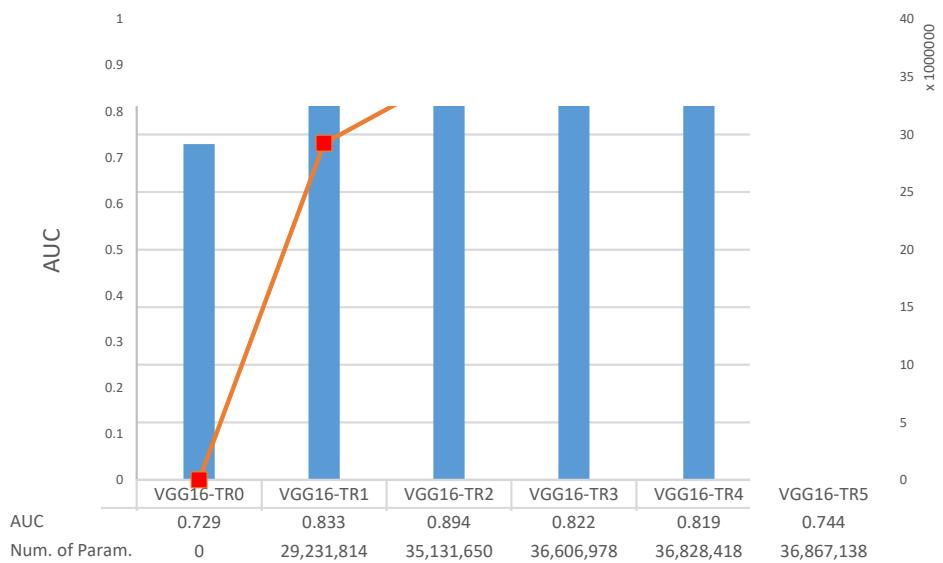
The deep CNN models of this study were trained using a cross-entropy loss function on the selected training image dataset. The screening performances of the six deep CNN models tested in this study are displayed in Table 2. It was observed that the transfer learning and fine tuning with last two blocks of VGG-16 model with pre-trained weights (VGG16-TR2) achieved the top performance, with the highest AUC of 0.894, sensitivity of 0.898, specificity of 0.778, and accuracy of 0.856. However, fine-tuning of different number of VGG-16 blocks showed lower screening performances than that of last two blocks of VGG-16 trainable (Figure 10 and 11).

**Table 2.** Osteoporosis screening accuracy of convolutional neural network models in this research.

<b>Model</b>	<b>AUC</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>Accuracy</b>
VGG16-TR0	0.729	0.686	0.650	0.674
VGG16-TR1	0.833	0.797	0.809	0.801
VGG16-TR2	0.894	0.898	0.778	0.856
VGG16-TR3	0.822	0.797	0.793	0.796
VGG16-TR4	0.819	0.788	0.809	0.796
VGG16-TR5	0.744	0.669	0.698	0.679



**Figure 10.** Mean ROC curves of each CNN models for screening osteoporosis on DPR images in this study.



**Figure 11.** Mean ROC curves of each CNN models for screening osteoporosis on DPR images in this study.

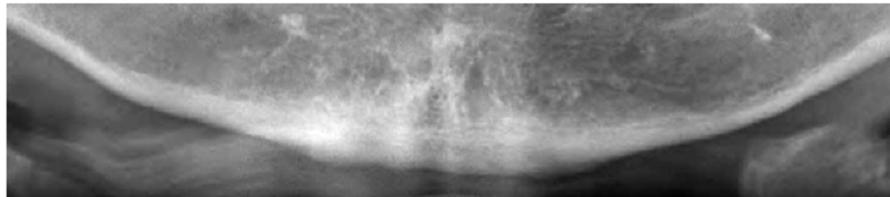
### ***3.5. Visualizing Model Decisions***

Figure 11 and Figure 12 illustrate the case examples of predictions using the best predictive VGG16-TR2 model. Each case example employed a Grad-CAM technique to perform a visual interpretation to determine which areas affected the deep CNN’s class classification. In the case of screening correctly for osteoporosis (Figure 11A), the region showing the weak lower border of the mandibular cortical bone and the less dense, spongy bone texture at its periphery was extracted as the main image feature of the classification. In correctly screened cases of no osteoporosis (Figure 5B), the region showing the strong lower boundary of the mandible cortical bone and the dense texture around its periphery was extracted as the main image feature of the classification. However, in the case of incorrectly screened cases, i.e., the non-osteoporosis case predicted as osteoporosis (Figure 12A) or the osteoporosis case predicted as non-osteoporosis (Figure 12B), the central region of the mandible or the ghost images of the hyoid bone was extracted as the main image feature.

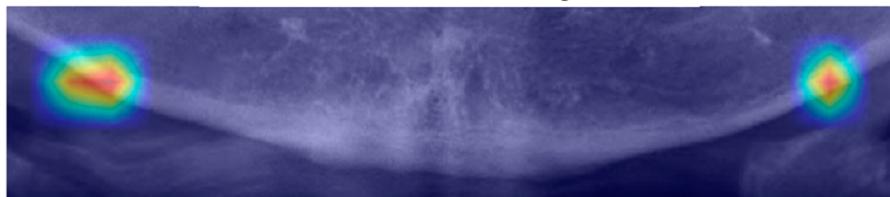
**(A) True Positive Case**

(True label = Osteoporosis / Predicted Label = Osteoporosis)

Original Image



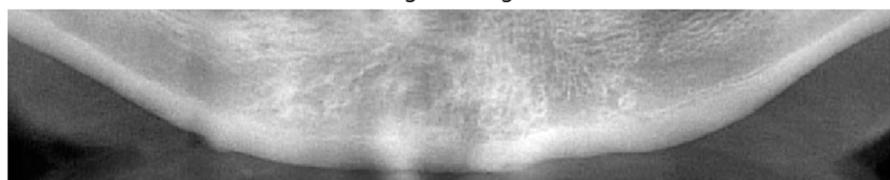
Grad-CAM overlaid Image



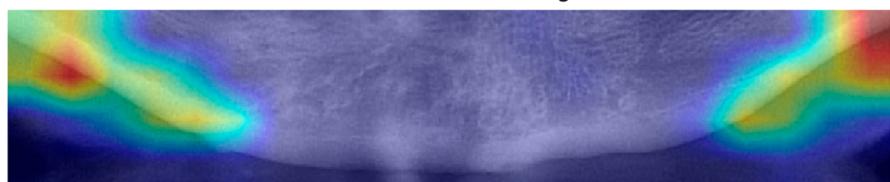
**(B) True Negative Case**

(True label = Non-osteoporosis / Predicted Label = Non-osteoporosis)

Original Image



Grad-CAM overlaid Image

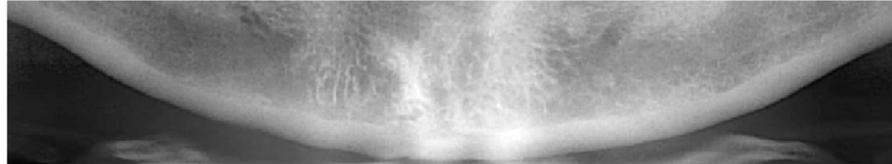


**Figure 12.** Original and Grad-CAM sample images of correctly predicted by the best-performing deep CNN model (VGG16-TR2) for DPR image-based osteoporosis screening are illustrated. Below each original sample images, a Grad-CAM image is superimposed over the original image. The bright red in each Grad-CAM image indicate the region that has the greatest impact on screening osteoporosis patients.

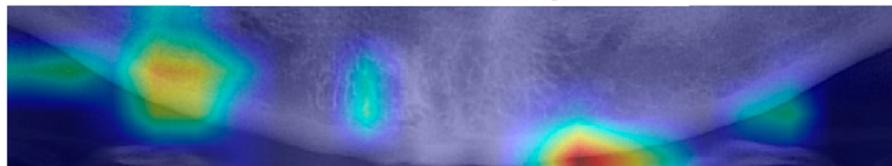
### (A) False Positive Case

(True label = Non-osteoporosis / Predicted Label = Osteoporosis)

Original Image



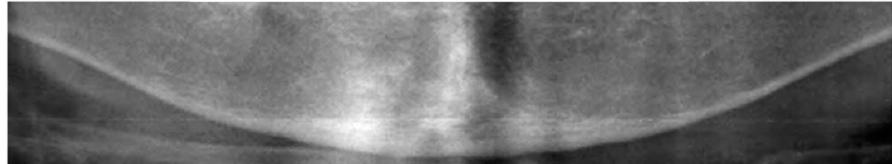
Grad-CAM overlaid Image



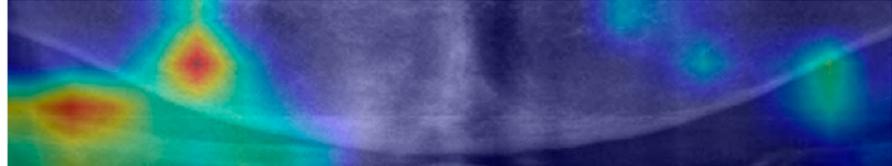
### (B) False Negative Case

(True label = Osteoporosis / Predicted Label = Non-osteoporosis)

Original Image



Grad-CAM overlaid Image



**Figure 13.** Original and Grad-CAM sample images of incorrectly predicted by the best-performing deep CNN model (VGG16-TR-TF) for DPR image-based osteoporosis screening are illustrated. Below each original sample images, a Grad-CAM image is superimposed over the original image. The bright red in each Grad-CAM image indicate the region that has the greatest impact on screening osteoporosis patients.

## Chapter 4. Discussion

Although DPRs are commonly performed for the evaluation of dentition and adjacent structures of the jaw, some clinical assistant diagnosis (CAD) systems based on DPRs have been suggested for screening systemic diseases, such as osteoporosis and carotid artery calcification [13–23,43]. However, the approaches of most previous studies are only valid when image features are accurately extracted, using sophisticated and manual image preprocessing algorithms or techniques. If a DPR image is imported from an unfamiliar environment or unexpected noise is added to the image, the prediction can easily be distorted. The neural network algorithm can resolve this problem. All the knowledge necessary for diagnosis is established only with the given training image data, without complicated or sophisticated image preprocessing. In recent years, a cutting-edge neural network technology, called deep learning, has been applied to medical imaging analysis and has shown a level of performance that is equal to or better than a clinician. As mentioned above, most previous CAD system studies, which used manual or sophisticated

image preprocessing and machine learning algorithms for the screening of osteoporosis based on DPRs, presented variable diagnostic performances, in terms of sensitivity and specificity [13–23]. Recently, a deep learning-based osteoporosis prescreening study, which resulted in a very high AUC score (0.9763 to 0.9991) and accuracy (92.5% to 98.5%), was published [44]. However, in that study, osteoporosis labeling was subjectively performed by dental specialists, rather than BMD score (T-score) which is the gold standard for diagnosing osteoporosis. In addition, the study did not visually interpret the decision of the trained CNN model, and using five arbitrarily established convolutional layers, there is a limitation to the reproducibility of the deep CNN model.

According to Table 2 and Figure 9 and 10, the first major findings of the present study showed that applying appropriate transfer learning and fine-tuning techniques on pre-trained deep CNN architectures had an equivalent DPR-based osteoporosis screening level of previous studies, even with small image datasets without complex image preprocessing and image ROI settings. The VGG16-TF0, having no trainable layers and the lowest number of trainable parameters, showed the lowest true-positive screening

performance and accuracy among the experimental groups. On the basis of these results, it can be estimated that a deep CNN model with a small number of trainable layers can have limitation in learning the true data distribution from a small number of dataset.

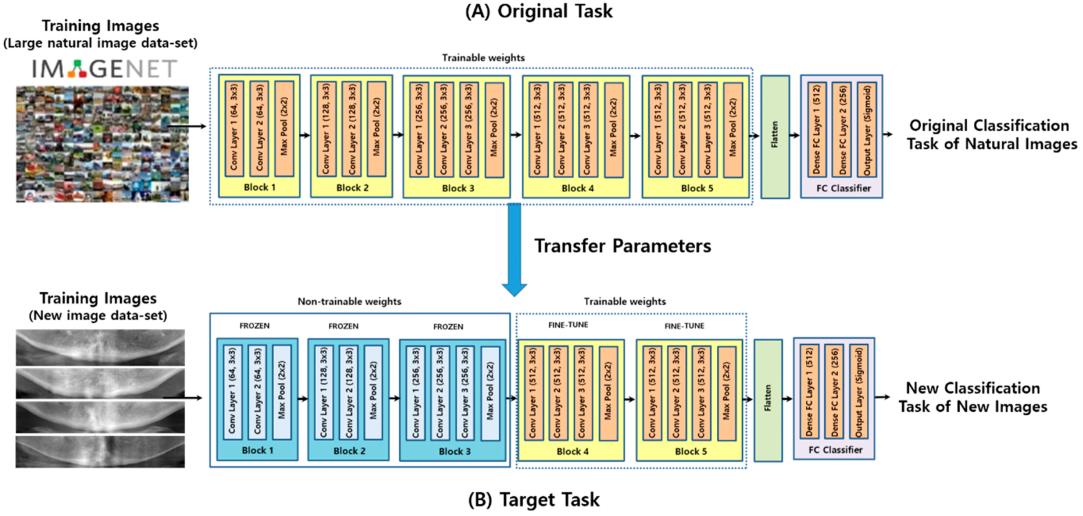
However, comparing other models having higher number of trainable parameters (VGG16-TR3, VGG16-TR4 and VGG16-TR5) with VGG16-TR2, deep CNNs use of excessive trainable parameters in the networks again degrades its classification performance.

In general, the deep CNN model learned from pre-trained deep neural networks on a large natural image dataset could be used to classify common images but cannot be well utilized for specific classifying tasks of medical images (Figure 13A). However, according to a previous study that described the effects and mechanisms of fine tuning on deep CNNs [45], when certain convolutional blocks of a deep CNN model were fine-tuned, the deep CNN model could be further specialized for specific classifying tasks (Figure 13B). More specifically, earlier layers of a deep CNN contain generic features that should be useful to many classification tasks, but later layers progressively contain more specialized features to the details of the classes contained in the original dataset (i.e., the large natural image

dataset on which the deep CNN was originally trained). Using this property, when the parameters of the early layers are preserved and the parameters in later layers are updated during training new datasets, the deep CNN model can be effectively used in new classification tasks. In conclusion, fine-tuning uses the parameters learned from a previous training of the network on a large dataset and, then, adjusts the parameters in later layers from the new dataset, improving the performance and accuracy in the new classification task. As with the previous study, the fine-tuning technique, which freezes the weight parameters of some initial convolutional blocks in the deep CNN model called VGG-16, and, then, updates the weight parameters of the later convolutional blocks (Figure 8B), show higher performance than other experimental groups. The conceptual diagram of the fine-tuning technique mentioned above can be seen in Figure 13.

Thus, in the case of having a small-scale image dataset like classification for medical images with a small number of data as in this study, this study also suggests that the use of optimal fine-tuning with transfer learning on a optimal deep CNN models with pre-trained weights can be an efficient solution for the

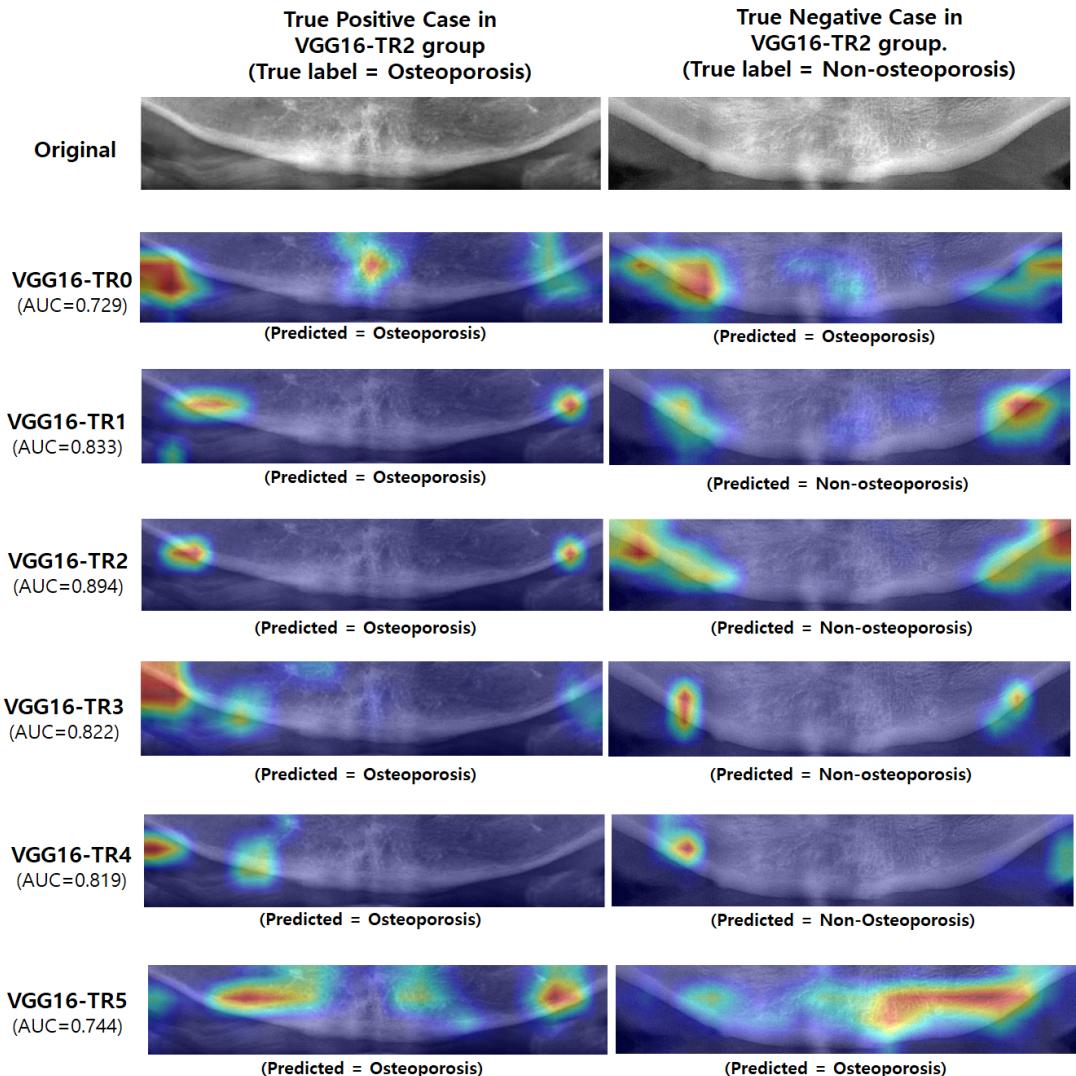
classification of medical images, instead of learning a deep neural network from scratch.



**Figure 14.** The conceptual diagram of the fine-tuning technique in the transfer learning of a deep CNN in this study.

The second major result of this study was to identify areas where image feature differences occurred when screening osteoporosis in DPR images using the Grad-CAM technique. To understand and visualize the decision of deep CNN models, some samples of the correctly and incorrectly screened examples were reviewed (Figure 11, 12, and 14). For additional insight to model decisions, a Grad-CAM technique was performed in this study. This technique identified the areas of input images that had the greatest impact on model

classification. According to this additional review, the model does seem to identify the feature characteristics of osteoporosis in DPR images (e.g., cortical bone thinning). According to the Grad-CAM evaluation of this study, DPR-based screening performances of osteoporosis were high when the image features were specified in the middle region of the left and right side of the mandibular lower border. This region is also consistent with the regions used to discriminate osteoporosis using DPR images, in most previous studies [13–23], although the measurement algorithm was different. This indicates that most osteoporosis patients have image feature characteristics, on DPR images, at the lower border of the cortical bone in the mandible. However, image quality issues, such as blurring, low contrast, and ghost images of adjacent objects can cause incorrect predictions. When the image features were specified in the center region of the mandible, or when the ghost images of the hyoid bone were in the ROI region, the accuracy was reduced. Therefore, to improve the deep CNN-based screening performance of osteoporosis in DPR images, it is suggested that the ROI setting be limited to the area around the middle of the left and right side of the lower border of the mandible.



**Figure 15.** Comparison of grad-CAM images from other groups against some original images showing true positive and true negative in the best performing VGG16-TR2 group.

## Chapter 5. Conclusion

This study showed an appropriate learning strategy in applying the transfer learning method to deep CNN for the screening of osteoporosis in DPR images when there is limited number of dataset. Before proceeding with the transfer learning, the back-bone network was selected, and then the transfer learning was conducted by applying an appropriate fine-tuning degree to the selected network to make it an efficient transfer learning method. The experimental results showed that transfer learning with pre-trained weights and fine-tuning techniques achieved the highest overall accuracy of 84%. The presented results suggest that the combination of the appropriate deep CNN architectures and transfer learning techniques has effectively resolved the issue of a small training set of images and that DPR images have the potential for osteoporosis pre-screening. In addition, using the Grad-CAM technique, this study performed a deep learning-based visual explanation for the area where the image feature difference occurred. Therefore, this study confirmed the previous osteoporosis screening studies using DPR images that set

the ROI at the middle of the left and right side of the lower border of the mandible. Given the increasing burden of osteoporosis on the global healthcare system, as our population ages, and the proliferation of dental panoramic image devices, the results presented in this study suggest that deep learning-based image analysis of DPRs could serve an important role in cost-effective prescreening for patients unaware of osteoporosis. To further improve screening performance, future research is needed, using different deep CNN architectures and deep learning techniques, more validated and qualified labeled image dataset, the appropriate number of datasets, and automated configuration techniques for more limited range of ROI.

## References

1. NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy, March 7 - 29, 2000: Highlights of the conference. *South. Med. J.* 2001, 94, 569 - 573.
2. Cauley, J.A. Public health impact of osteoporosis. *J. Gerontol. A Biol. Sci. Med. Sci.* 2013, 68, 1243 - 1251.
3. Bliuc, D.; Nguyen, N.D.; Nguyen, T.V.; Eisman, J.A.; Center, J.R. Compound risk of high mortality following osteoporotic fracture and refracture in elderly women and men. *J. Bone Miner. Res.* 2013, 28, 2317 - 2324.
4. Sozen, T.; Ozisik, L.; Basaran, N.C. An overview and management of osteoporosis. *Eur. J. Rheumatol.* 2017, 4, 46 - 56.
5. Melton, L.J., 3rd; Chrischilles, E.A.; Cooper, C.; Lane, A.W.; Riggs, B.L. Perspective. How many women have osteoporosis? *J. Bone Miner. Res.* 1992, 7, 1005 - 1010.
6. Melton, L.J., 3rd; Atkinson, E.J.; O'Connor, M.K.; O'Fallon, W.M.; Riggs, B.L. Bone density and fracture risk in men. *J. Bone Miner. Res.* 1998, 13, 1915 - 1923.

7. Kanis, J.A.; Johnell, O.; Oden, A.; Sembo, I.; Redlund-Johnell, I.; Dawson, A.; De Laet, C.; Jonsson, B. Long-term risk of osteoporotic fracture in Malmo. *Osteoporos. Int.* 2000, 11, 669 - 674.
8. Kalinowski, P.; Rozylo-Kalinowska, I.; Piskorz, M.; Bojakowska-Komsta, U. Correlations between periodontal disease, mandibular inferior cortex index and the osteoporotic fracture probability assessed by means of the fracture risk assessment body mass index tool. *BMC Med. Imaging* 2019, 19, 41.
9. Marcucci, G.; Brandi, M.L. Rare causes of osteoporosis. *Clin. Cases Miner. Bone Metab.* 2015, 12, 151 - 156.
10. Kanis, J.A.; Johnell, O. Requirements for DXA for the management of osteoporosis in Europe. *Osteoporos. Int.* 2005, 16, 229 - 238.
11. Kanis, J.A. Diagnosis of osteoporosis and assessment of fracture risk. *Lancet* 2002, 359, 1929 - 1936.
12. Mithal, A.; Bansal, B.; Kyer, C.S.; Ebeling, P. The Asia-Pacific Regional Audit-Epidemiology, Costs, and Burden of Osteoporosis in India 2013: A report of International Osteoporosis Foundation. *Indian J. Endocrinol. Metab.* 2014, 18, 449 - 454.
13. Taguchi, A.; Suei, Y.; Ohtsuka, M.; Otani, K.; Tanimoto, K.;

- Ohtaki, M. Usefulness of panoramic radiography in the diagnosis of postmenopausal osteoporosis in women. Width and morphology of inferior cortex of the mandible. *Dentomaxillofac. Radiol.* 1996, 25, 263 - 267.
14. Ledgerton, D.; Horner, K.; Devlin, H.; Worthington, H. Radiomorphometric indices of the mandible in a British female population. *Dentomaxillofac. Radiol.* 1999, 28, 173 - 181.
15. White, S.C.; Taguchi, A.; Kao, D.; Wu, S.; Service, S.K.; Yoon, D.; Suei, Y.; Nakamoto, T.; Tanimoto, K. Clinical and panoramic predictors of femur bone mineral density. *Osteoporos. Int.* 2005, 16, 339 - 346.
16. Yasar, F.; Akgunlu, F. The differences in panoramic mandibular indices and fractal dimension between patients with and without spinal osteoporosis. *Dentomaxillofac. Radiol.* 2006, 35, 1 - 9.
17. Taguchi, A.; Ohtsuka, M.; Tsuda, M.; Nakamoto, T.; Kodama, I.; Inagaki, K.; Noguchi, T.; Kudo, Y.; Suei, Y.; Tanimoto, K. Risk of vertebral osteoporosis in post-menopausal women with alterations of the mandible. *Dentomaxillofac. Radiol.* 2007, 36, 143 - 148.
18. Devlin, H.; Karayianni, K.; Mitsea, A.; Jacobs, R.; Lindh, C.; van der Stelt, P.; Marjanovic, E.; Adams, J.; Pavitt, S.; Horner, K.

Diagnosing osteoporosis by using dental panoramic radiographs:  
The OSTEODENT project. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* 2007, 104, 821 - 828.

19. Okabe, S.; Morimoto, Y.; Ansai, T.; Yoshioka, I.; Tanaka, T.; Taguchi, A.; Kito, S.; Wakasugi-Sato, N.; Oda, M.; Kuroiwa, H.; et al. Assessment of the relationship between the mandibular cortex on panoramic radiographs and the risk of bone fracture and vascular disease in 80-year-olds. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* 2008, 106, 433 - 442.
20. Taguchi, A. Triage screening for osteoporosis in dental clinics using panoramic radiographs. *Oral Dis.* 2010, 16, 316 - 327.
21. Al-Dam, A.; Blake, F.; Atac, A.; Amling, M.; Blessmann, M.; Assaf, A.; Hanken, H.; Smeets, R.; Heiland, M. Mandibular cortical shape index in non-standardised panoramic radiographs for identifying patients with osteoporosis as defined by the German Osteology Organization. *J. Craniomaxillofac. Surg.* 2013, 41, e165 - e169.
22. Kavitha, M.S.; Asano, A.; Taguchi, A.; Kurita, T.; Sanada, M. Diagnosis of osteoporosis from dental panoramic radiographs using the support vector machine method in a computer-aided system.

BMC Med. Imaging 2012, 12, 1.

23. Kavitha, M.S.; Ganesh Kumar, P.; Park, S.Y.; Huh, K.H.; Heo, M.S.; Kurita, T.; Asano, A.; An, S.Y.; Chien, S.I. Automatic detection of osteoporosis based on hybrid genetic swarm fuzzy classifier approaches. *Dentomaxillofac. Radiol.* 2016, 45, 20160076.
24. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.; van Ginneken, B.; Sanchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* 2017, 42, 60 - 88.
25. Park, C.; Took, C.C.; Seong, J.K. Machine learning in biomedical engineering. *Biomed. Eng. Lett.* 2018, 8, 1 - 3.
26. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436 - 444.
27. Baker, B.; Gupta, O.; Naik, N.; Raskar, R. Designing neural network architectures using reinforcement learning. *arXiv* 2016, arXiv:1611.02167. [Google Scholar]
28. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *arXiv* 2014, arXiv:1411.1792. [Google Scholar]
29. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans.*

Knowl. Data Eng. 2009, 22, 1345 - 1359.

30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. arXiv 2015, arXiv:1512.03385. [Google Scholar]
31. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556. [Google Scholar]
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 3 - 8 December 2012; Curran Associates Inc.: Red Hook, NY, USA, 2012; Volume 1, pp. 1097 - 1105. [Google Scholar]
33. Han, Z.; Wei, B.; Zheng, Y.; Yin, Y.; Li, K.; Li, S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci. Rep. 2017, 7, 4172.
34. Christopher, M.; Belghith, A.; Bowd, C.; Proudfoot, J.A.; Goldbaum, M.H.; Weinreb, R.N.; Girkin, C.A.; Liebmann, J.M.; Zangwill, L.M. Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. Sci. Rep. 2018, 8, 16685.

35. Shin, H.-C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* 2016, 35, 1285 - 1298.
36. Ravishankar, H.; Sudhakar, P.; Venkataramani, R.; Thiruvenkadam, S.; Annangi, P.; Babu, N.; Vaidya, V. Understanding the mechanisms of deep transfer learning for medical images. *arXiv* 2017, arXiv:1704.06040. [Google Scholar]
37. Kanis, J.A. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: Synopsis of a WHO report. WHO Study Group. *Osteoporos. Int.* 1994, 4, 368 - 381.
38. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Compute. Vis.* 2015, 115, 211 - 252.
39. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Methodol.* 1974, 36, 111 - 133.
40. Chollet, F. Keras: Deep Learning Library for Theano and Tensorflow. 2015, 7, p. T1. Available online: <https://keras.io>

(accessed on 30 January 2020).

41. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv 2016, arXiv:1603.04467. [Google Scholar]
42. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. arXiv 2016, arXiv:1610.02391. [Google Scholar]
43. Sawagashira, T.; Hayashi, T.; Hara, T.; Katsumata, A.; Muramatsu, C.; Zhou, X.; Iida, Y.; Katagi, K.; Fujita, H. An automatic detection method for carotid artery calcifications using top-hat filter on dental panoramic radiographs. IEICE Trans. Inf. Syst. 2013, 96, 1878 - 1881.
44. Lee, J.-S.; Adhikari, S.; Liu, L.; Jeong, H.-G.; Kim, H.; Yoon, S.-J. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: A preliminary study. Dentomaxillofac. Radiol. 2019, 48, 20170344.
45. Nogueira, K.; Penatti, O.A.; Dos Santos, J.A. Towards better

exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognit.* 2017, **61**, 539–556.

# 국문초록

## 치과용 파노라마 방사선 사진에서 골다공증 선별을 위한 심층 합성곱 신경망(deep CNN)의 전이학습효과 비교분석

이기선

의학과 의공학전공  
서울대학교 대학원

골다공증은 골밀도가 낮고 골 미세 구조의 붕괴가 특징인 대사성 골 질환입니다. 그러나, 골다공증에 대한 임상 진단 방법중에 하나인 DXA 검사는 대형의 검사용 엑스레이 장비가 별도로 필요하고 검사비용이 높아, 해당 검사의 이용성에 제한성이 있습니다. 최근 연구에 따르면 치과 파노라마 방사선 사진 (DPR) 또한 골 밀도 변화를 예측 할 수 있다고 연구되었습니다. 이에 본 연구는 DPR에서 골다공증에 의한 골 밀도 변화에 따른 엑스레이 영상 특이성 분류에 다양한 전이 학습전략을 적용한 심층 합성곱 신경망 (CNN)의 분류 성능을 평가하는 것에 목표로 두었습니다. 합습 및 검증용 데이터의 객관적인 라벨링을 위해 2009년부터 2018년까지 고려 대학교 안산 병원에서 골밀도 검사와 디지털 파노라마 방사선 촬영을 6개월 이내에 동시에 시행한 환자들로부터 680개의 데이터 세트를 수집했습니다. 전이 학습 전 기본이 되는 합성곱 신경망을 선택하기 위해 이미지 분류에 자주 사용되는 3개의 합성곱 신경망인 VGG-16, Resnet-50 및 Xception 네트워크에 대해 전이학습이 없는 상태로 사전 분류성능 평가를 수행했습니다. VGG-16은 전이 학습 없이 수행 된 분류 성능 평가에서 다른 2개의 네트워크에 비해 높은 AUC 값을 보여 주었기에, 해당 네트워크를 백본(back-bone) 네트워크로 사용하여 전이학습 효과를 비교 분석하였습니다. 백본 네트워크에서 최적의 fine-tuning 정도를 찾기 위해 VGG-16에 fine-tuning이 적용 가능한 블록 수에 따라 총 6 개의 fine-tuning 적용 전이 학습 그룹이 다음과 같이 설정 하였습니다. fine-tuning을 전혀 하지 않는 그룹 (VGG16-TR0), 마지막 1 블록을 fine-tuning 하는 그룹 (VGG-16-TF1), 마지막 2 블록을 fine-tuning 하는 그룹 (VGG-16-TF2), 마지막 3 개 블록을 fine-tuning하는 그룹 (VGG-16-TF3), 마지막 4 개 블록을 fine-tuning하는 그

룹 (VGG-16-TF4) 및 5 개 블록 모두를 fine-tuning하는 그룹 (VGG16-TR5). 실험 결과 최고 성능 모델은 VGG-16-TF2 였으며, 분류 성능 값의 하나인 AUC 값이 0.858를 달성했습니다. 본 연구를 통하여 학습용 데이터 수에 제한이 있더라도, 전이 학습 및 fine-tuning을 통하여 DPR 이미지를 이용한 골다공증 스크리닝 성능의 개선이 가능함을 보여주었습니다. 또한 gradiant-CAM 기법을 이용하여 성능이 가장 우수한 CNN 모델의 시각적 해석을 통하여, DPR 이미지 상에서 적절한 골다공증의 분류성능은 하악골의 왼쪽 및 오른쪽 하연 경계에 있는 이미지에 의존한다는 것을 확인 할 수 있었습니다. 본 결과는 DPR 이미지의 딥러닝 기반 평가가 골다공증 환자의 자동 선별에 유용하고 신뢰할 수 있음을 시사 하였습니다.

주요어 : 골다공증 선별, 딥러닝, 합성곱 신경망, 파노라마 엑스레이

학 번 : 2017-36469