공학박사 학위 논문

딥러닝을 이용한 흉부 단순 촬영에서 생물학적 연령 예측과 골다공증 선별 모델 개발

Development of an osteoporosis screening model with biological age prediction on chest radiographs using deep learning

울 산 대 학 교 대 학 원

의 과 학 과

장 미 소

딥러닝을 이용한 흉부 단순 촬영에서 생물학적 연령 예측과 골다공증 선별 모델 개발

지도교수 김 남 국

이 논문을 공학박사 학위 논문으로 제출함

2022년 8월

울 산 대 학 교 대 학 원
의 과 학 과
장 미 소

장미소의 공학박사 학위 논문을 인준함.

심사위원장 서 준 범 (인)

심사위원 고 정 민 (인)

심사위원 김 남 국 (인)

심사위원 안 경 식 (인)

심사위원 유 재 준 (인)

울산대학교 대학원

2022년 8월

Abstract

Recently, deep learning algorithms, especially convolutional neural network (CNN) architectures, have been widely recognized as an outperforming and reliable approach to identify clinically useful features directly from the medical images. The majority of studies that use deep learning for medical image analysis aim to replicate a task already performed by humans. In addition to emulating humans, deep learning also offers the potential of identifying significant imaging features that are beyond a radiologist's visual search pattern, and potentially improving the diagnostic or screening value of medical images.

Chest radiography is the most widely available medical imaging modality, providing a wealth of information regarding the cardiovascular and musculoskeletal systems. Research using chest radiography-derived biological age estimates were showed successful predictions of longevity, long-term mortality, and cardiovascular risk. Prediction of sex on chest radiographs using deep learning model has been shown high accurate performance and it could be a potential contributor in imbalanced sex representation datasets. Opportunistic screening using deep learning also has been applied various medical imaging for early detecting osteoporosis, compression fractures, cardiovascular risk, and so on.

For the first phase of this thesis, deep learning models that can identify demographic information from a single chest radiograph were developed. Chest radiographs are the first-line tool for screening patients with nonspecific symptoms of the thorax in routine clinical practice and demographic information is the most basic and important factor in predicting the prognosis of a disease. The benefits of medical imaging for predicting age and sex may stem from the fact that age and sex are not simply one single concept. In fact, the overall consequences of temporal, biological, and/or pathological alterations may be reflected in imaging findings. We aimed to assess the accuracy of a deep learning model for age and sex estimation based on chest radiographs reported within normal limits of adult from a real-world clinical dataset. The performance of deep learning models for predicting age and sex across architectures were evaluated and stress tests on limited number of data were conducted.

For the second phase of this thesis, deep learning models for screening osteoporosis with chest radiographs were developed. Osteoporosis is often not detected until fracture presentation and is hence considered a "silent epidemic" with a need for early diagnosis. Some studies have conducted osteoporosis screening in bones other than the vertebrae and femur on

hand radiographs and panoramic radiographs. When using chest radiographs, readers often mention the presence of osteoporosis, but it is known that even if made by an experienced radiologist, the diagnosis of osteoporosis using chest radiographs is still unreliable. However, in this thesis, the deep learning models for opportunistic screening osteoporosis using chest radiographs were developed and verified in the external dataset.

In this thesis, the stress tests in the osteoporosis dataset were also conducted on limited number of datasets and reconstructed dataset by sampling the same number of men and women. Osteoporosis is a prevalent age-related condition that is more common in women than in men and it also found in our osteoporosis dataset. In order to become an unbiased and generally applicable model, it must be able to utilize or exclude gender and age information well. To this end, various experiments using gender and age information were conducted. In addition, to be applied and used in clinical practice, it is necessary to have the explanatory power of the model's decision as well as the performance of the model. The gradient-weighted class activation mapping technique (Grad-CAM) was used to identify the discriminative regions on chest radiographs contributed to the model's predictions.

In conclusion, deep learning model can extract more information on chest radiographs than human perceptions. Although this thesis has only shown the potential, chest radiographs-derived biologic age may help to evaluate prognosis of individual or diagnosis of disease. In addition, osteoporosis screening deep learning model using chest radiographs was developed. Experiments were conducted to explain this result and to make the model robust in various situations. Examples of usage scenarios are presented in this model so that it can be applied to clinical practice. This model is expected to provide practical utility that chest radiographs can be used for the opportunistic screening of osteoporosis, without additional exposure and substantial costs.

Key words: Deep learning, Opportunistic screening, Osteoporosis, convolutional neural network (CNN), chest radiograph, chest radiographs-derived biologic age

Contents

List of Tables

List of Figures

1. Introduction

1.1 Motivations

The rapid development of artificial intelligence (AI) in medicine has been due to advances in algorithms, the computing power of graphics processing units (GPUs), and the generation of healthcare bigdata. [1] Pattern recognition and machine learning techniques are widely used and form the basis of useful medical image analysis systems. In particular, deep learning is one of the most important technologies in machine learning. It is actively being applied and developed in the medical field and the number of studies [2] has increased significantly in recent years. Deep learning algorithms have been widely recognized as an outperforming and reliable approach to identify clinically useful features directly from the medical images. The applications include image segmentation, image reconstruction, classification, and estimation biological age from images. [3-6]

There are several important issues in the medical fields, such as diagnosis of diseases, determination of treatment, prediction of prognosis and prevention of diseases. Prevention activities are generally divided into three levels with the primary prevention goal of preventing disease or injury before it occurs. Secondary prevention aims to reduce the effects of a disease or injury that has already occurred, while tertiary prevention aims to mitigate the effects of an ongoing disease or injury that has a lasting effect. In particular, primary prevention consists of measures aimed at susceptible populations or individuals. The target population is healthy individuals. It usually initiates action to limit risk exposure or increase immunity in at-risk individuals to prevent the disease from progressing to asymptomatic disease in vulnerable individuals. [7] For primary prevention, selecting individual high risk of diseases is very important.

Chest radiographs are the most common and non-invasive images in medical fields [8] and there are the largest number of public datasets consisted of them. [9, 10] Chest radiography using a very small dose of ionizing radiation is primary screening tool for detecting lesion of tuberculosis, pneumonia, and lung cancer. Most results of chest radiographs are reported as normal, in that they rule out a specific diagnosis. However, even normal radiographs manifest additional minor abnormalities, such as aortic calcification or an enlarged heart. [11] Some research [6, 11, 12] showed Advanced deep learning model could extract prognostic information from chest radiographs. These images have age and sex information, and deep learning-based model can learn demographic information from these images [13-16]. One recent research suggested that chest radiographs-derived biologic age could be used as a prognostic imaging biomarker of treatment decisions in non-small cell lung cancer patients. [17]

On the other hands, osteoporosis is a systemic disease characterized by low bone mineral density (BMD) and microstructural deterioration of the bone structure, leading to a consequent increase in fracture risk. [18] Osteoporosis and osteoporotic fractures have become global health issues of major concern owing to their association with age-related fractures in the aging countries including South Korea. [19] Hip, spine, and wrist fractures caused by osteoporosis often lead to disorders that reduce the patient's quality of life and, in severe cases, increase the mortality risk. [20] According to recent statistics from the International Osteoporosis Foundation, approximately one-third of women and one-fifth of men aged ≥ 50 years will experience an osteoporotic fracture. [21-23] Osteoporosis is often not detected until fracture presentation and is hence considered a “silent epidemic” with a need for early diagnosis. [24, 25]

Selecting high risk group of osteoporosis assessment with chest radiographs is worth applying with deep learning. In general, osteoporosis is diagnosed using BMD measured on

central dual energy X-ray absorptiometry (DXA), which is considered a reference standard test. [26] In this study, we aimed to develop and evaluate the deep learning approaches for screening osteoporosis using the classified chest radiographs based on the DXA based diagnosis of osteoporosis. We trained deep learning models in various ways and investigated how age and sex information affect it. In addition, we evaluated trained models for classifying performance and validated the performances in an external validation set.

1.2 Contributions

The main contributions of this thesis are summarized as follows. First, the age and sex information that can be extracted by deep learning from chest radiographs was studied. This will help develop a deep learning model for prognosis prediction in the clinical field in the future. In addition, an ablation study was conducted to limit the number of data, and the amount of data in which age and sex information could be predicted was confirmed. Second, age assessment model was applied to clinical dataset and the results were analyzed. Chest radiographs-derived biologic age showed potential of correlation of clinical information. Third, a deep learning model was developed for screening high risk of osteoporosis using chest radiographs. This model evaluated in the external validation set and showed usable results with AUC 0.88. These chest radiographs can be used for the opportunistic screening of osteoporosis, without additional exposure and substantial costs in real clinical setting. Fourth, in the deep learning model for screening osteoporosis, an ablation study that limits the number of data and various learning methods that add gender and age information were applied. This will be of great help in future multi-nation and multi-institution verification studies.

2. Background

2.1. X-ray(radiography) and modalities

X-rays discovered in 1895 are a form of electromagnetic radiation. [27] Prior to 1912, some x-rays of metal were produced, but x-rays were rarely used outside the medical and dental fields. However, this was changed in 1913 when Coolidge designed a high vacuum X-ray tube. In 1922, industrial radiography was able to produce radiographs of thick steel parts one step further by the advent of a 200000-volt X-ray tube at a reasonable time. In 1931, General Electric Company developed a 1000000-volt X-ray generator to provide an effective tool for industrial radiography. X ray has been developed for use in various fields in the 20th century and has been used as a major diagnostic tool, particularly in the medical field. Conventional X ray imaging [28] is usually analog technology unlike other biomedical images such as computed tomography (CT), ultrasound, nuclear medicine, and magnetic resonance imaging (MRI). Digital X-ray s or digital radiation is beneficial because it allows image processing to improve aspects of image quality such as image contrast. Digital X-rays are often used to compare with other imaging modalities in the hospital for remote access and archiving or detect various diseases as an initial diagnosis with a computer aided diagnosis. In addition, a large number of X-ray imaging can be more easily acquired than other imaging modalities. Therefore, chest radiographs are the first diagnostic tool for patients with nonspecific symptoms of the chest in routine clinical practice. They can be performed efficiently at low cost and is readily available to most institutions. Recently, the development of deep learning has resulted in meaningful results with detecting pulmonary nodules, supporting diagnosis pneumonia and tuberculosis, pneumothorax, and so on. [15, 29-31]

Dual-energy X-ray absorptiometry (DXA) is indispensable for clinical practice in osteoporosis. DXA is the current ‘gold standard’ for measuring bone mineral density (BMD)

in central skeleton (lumbar spine and proximal femur). [32] The fundamental physical principle behind DXA is the measurement of the transmission through the body of x-ray with high- and low-photon energies. Because of the dependence of the x-ray attenuation coefficient on atomic number and photon energy, measurement of the transmission factors at two different energies enables the areal densities (i.e., mass per unit projected area) of two different types of tissue to be inferred. In DXA scans these are taken to be bone mineral (hydroxyapatite) and soft tissue respectively. [33] BMD is ratio of the measured bone mineral content (BMC) in grams divided by the measured two-dimensional projected area in cm^2 of the bone(s) being measured; thus, the units of BMD are g/cm^2. However, most clinical decisions are based on the T-score, which is calculated by comparing the patient's BMD with the mean value for young normal individuals and expressing the difference as a standard deviation score. The T-score is calculated using the formula: (patient's BMD−young normal mean)/SD of young normal. [34]

2.2. Convolutional neural networks

AI technology has been growing fast and wide application for a government agency, commercial, medical and consumer use in business, school, sports, electronics, and medical, etc. For example, AI systems are used to recognize and detect various objects in natural images, automatically transcribe speech into a large number of text s teach a foreign language without a human tutor and select relevant results in search engines. These applications have been gradually developed using deep learning, one of the artificial neural networks (NNs). NNs are a subfield of machine learning, a small subset of AI. NN have already been introduced from the past. These studies [35-37] showed that NNs consist of many connected processors called neurons in the human neuronal synapse system, each producing a sequence of real valued

activations. Deep learning first appeared in 2006 as a new research area in machine learning. Deep learnings have made an advance in various studies to solve the difficult problems that cannot be solved by existing machine learning.

The main point of deep learning is that it could learn representation s of big data through multiple levels of abstraction using a computational model consisting of multiple processing layers and a backpropagation algorithm [38] to indicate how a machine should change its internal parameters used to compute the feature in each layer from the feature in the previous layer. [39, 40] Its learning methods can also be defined as representation learning [40] which involving a hierarchy of features from high level to low level. Bengio [41] examined various challenges of deep learning research and proposed a few forward-looking research directions overcoming these More to the point, deep learning mainly considers nonlinear processing in multiple layers and supervised or unsupervised learning. [42] Nonlinear processing in multiple layers refers to an algorithm where the current layer takes the output of the previous layer as an input. Hierarch y is established among layers to organize the importance of the data to be considered as useful or not. On the other hand, supervised and unsupervised learning is related to the class target label, the availability with a target label means a supervised system, whereas the availability without a target label means an unsupervised system.

Currently, deep learning has been used in various fields. [43-55] The important aspects of using deep learning were shown in the fields of image recognition [43-45] and speech research. [47, 48] In addition, it has been superior to other machine learning at inferencing the activity of potential drug molecules potential drug molecules [50] and analyzing particle accelerator data analyzing particle accelerator data. [51] Along with the progress of these studies, deep learning has received much attention from both the academic and industrial communities. Deep learning will have many more successes soon because of increases in the amount of available computation and data.

In deep learning, convolutional neural network (CNN) is certainly one of the main methods to classify, detect, and segment such a s image or video datasets. CNN comprises convolution layers, pooling layers, activation layers, and fully connected layers to infer outputs. [42, 56] The convolution layers extract features from an input image and retain the connection between image pixels through continuously learning a majority of images with a small kernel of input data. Pooling layers provide specific features while down sampling the output of convolution layer. Max pooling layer and average pooling layer are usually used in deep learning. [57, 58] The activation function is a nonlinear transformation that processes the input signal. The output is then entered into the next convolution layer as input. This function involves sigmoid, rectified linear unit (ReLu [59]), and Leaky ReLu [60].

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \quad (1)$$

$$ReLu\ (x) = \begin{cases} 0\ for\ x\ \leq 0 \\ x\ for\ x\ > 0 \end{cases} \quad (2)$$

$$Leaky\ ReLu\ (x) = \begin{cases} 0.01x\ for\ x\ < 0 \\ x\quad\ for\ x\ \geq 0 \end{cases} \quad (3)$$

Dropout layers [61] are used to prevent overfitting while training on CNN. For training of deep learning with a large amount of dataset, batch normalization [62] has used to mitigate gradient vanishing problems or gradient exploding In the training step, we could see the internal covariance shift [63] (i.e., the difference between input distributions at each network layers and network activations). To solve this problem, ReLu, careful initialization, [64] or small learning rates were used. In addition, the method to make the input distribution with zero mean and unit variance of each layer's inputs was performed in the training step. However, this solution normally requires high computational cost for normalization to calculate mean and variance. So, batch normalization was proposed to address the limitations of the existing approaches.

The advantage of this method is that greater learning rates can be used. It has the effect of training a more general model and has less dependence on drop out, l2 regularization, [65] etc. While learning, it proceeds to find the optimal feature map with weights sharing. [66] Additional filtering is performed in the next layer, etc. These processes converge into a feature map that best reflects the features of input images CNN uses local connection pattern s between units in adjacent layers to prevent errors in the spatial units that do not affect the others in back propagation step CNN exploits optimizer function to reduce errors in back propagation step. So, optimal weights can be selected with a minimum error rate.

As these CNN continue to evolve, it has shown top results in a variety of image related challenge contests to date. In addition, CNN adopts weight s haring, which assigns the same parametrization to each hidden unit in the same feature map. Applying weight sharing allows CNN to capture the re lative rather than the absolute distance between features. This specific characteristic of CNN protects robustly it against distortions or shifts. [56] As these CNN continue to evolve; it has shown top results in a variety of image related challenge contests to date. CNN was proposed in the early 1990s and noticed at ILSVRC 2012 with AlexNet [43] competing with other algorithms. This model has shown tremendous results compared to other algorithms with a multi-layer based on CNN Afterward, various models were introduced in ILSVRC like ResNet [67], and DenseNet [68], inception [69], EfficientNet [70], and etc. were introduced in deep learning fields.

Medical image classification is performed with two methods. One way is to analyze finding the texture of the images using image processing based on texture properties [71] in different regions. Second, it is using deep learning with a big dataset. Image classification based on deep learning is applied to determine cancer of the lung in chest radiographs and to classify lung image patches using a customized CNN. [72] Simple deep learning architecture called principal component analysis network (PCANet [73] was proposed to achieve more

accurate classification with the spatial distribution information of color images [74] in various datasets. CNNs trained by ImageNet to identify different types of pathologies in chest radiographs was employed. [29]

2.3. Transfer learning

Transfer learning is a popular approach in deep learning where a model trained on one task is repurposed on a second related task. [75] This is typically considered as a method to overcome the lack of training data set for a specific task or to train the target data set different from the input data set. Specifically, transfer learning is widely exploited to develop various models based on deep learning that need a large amount of data set in the training step. [76] One of first ideas to use transfer learning was to adopt pretrained models of the ImageNet [77] dataset of training from scratch. [78] The number of medical images including gold standard or references is less than that of general natural images to analyze for medical images, so, transfer learning is one of important methods to solve insufficient new data for training. [79]

Medical images of a specific field frequently have standardized views, such as the features of relevant tasks tending to have limited texture variants or small patches rather than high-level semantic features. A high-resolution is commonly significant, and images are often grayscale, X-ray images. [80] Raghu et al. [81] performed empirical experiments using two large medical datasets, retinal fundus [82] and a chest radiography (CheXpert [9]), to improve understanding of the ImageNet tradeoffs, including TL. These two experiments demonstrated that the domain variance between medical and natural images restricts TL. Recent works proposed another highly effective technique, known as in-domain pretraining, alongside the previously discussed transfer learning. [83, 84] For instance, Heker and Greenspan [84] introduced the utilization of trained weights using an in-domain dataset for liver-segmentation

purposes rather than the utilization of initialized weights from a model trained using ImageNet. Transfer learning method uses the weights of the trained on the first domain dataset as initialization weights for training in a new domain dataset, similar to the features of the first domain dataset. In addition, Training with a small dataset is possible with the pre trained model being only fine-tuned to in-domain dataset.

2.4. Interpretable tools for deep learning

Interpretable tools are considered to be the most important in order to explain the predictions of the CNN models toward the challenge of classifying medical images. Class activation map (CAM) was generated from a specific CNN architecture where global average pooled convolutional feature maps were fed into the fully connected final output layer. [85] Because one of the disadvantages of CAM is that it requires feature maps to directly precede the softmax layers, it is only relevant to CNN architectures that perform global average pooling over convolutional maps immediately before prediction. (i.e., conv feature maps → global average pooling →softmax layer).

Generalization of CAM, known as Gradient-weighted class-activation maps (Grad-CAM [86]), used the gradient information of a target class flowing back into the last convolutional layer to generate visual explanations from any CNN models. Grad-CAM for the class c was also defined as a weighted sum of all feature maps resulting from the last convolution layer in the CNN. In addition, ReLU function was applied to remove a potential influence from negative weights on the class of interest, considering that the spatial elements in the feature maps associated with the negative weights were likely to belong to other categories in the image.

$$Grad\_M_c(x,y) = ReLU(\sum_k \alpha_k^c \, f_k(x,y)) \qquad (4)$$

Here, α_k^c is the weight obtained by computing the gradient of a prediction score, s_c with respect to the k-th feature map:

$$\alpha_k^c = \sum_{x,y} \frac{\partial s_c}{\partial f_k(x,y)} \qquad (5)$$

Feature maps from the last block of convolutional layers of the models were used to compute the Grad-CAM heatmap. In many deep learning research using medical images, Grad-CAM heatmaps have been presented and explained. [12, 87-91] Heatmaps were superimposed on the original images for visualization.

2.5 Osteoporosis

Osteoporosis and osteoporotic fractures have become global health issues of major concern owing to their association with age-related fractures in the aging countries including South Korea. [19] By 2020, approximately 12.3 million people aged ≥50 years in the United States are expected to develop osteoporosis. [92] Prevalence of osteoporosis in South Korea is 7.3% in males and 38.0% in females aged ≥50 years by the data Fourth Korea National Health and Nutrition Examination Survey (KNHANES IV) 2008–2011. [93] Osteoporosis is a systemic disease characterized by low bone mineral density (BMD) and microstructural deterioration of the bone structure, leading to a consequent increase in fracture risk. [18] The T-score threshold value for osteoporosis is -2.5, according to the World Health Organization (WHO). For every standard deviation below the mean in a young adult, the chance of fracture increases by around threefold. As a result, poor bone mineral density continues to be a powerful indicator of future fracture risk. [94] Hip, spine, and wrist fractures caused by osteoporosis often lead to disorders that reduce the patient's quality of life and, in severe cases,

increase the mortality risk. [20] According to recent statistics from the International Osteoporosis Foundation, approximately one-third of women and one-fifth of men aged ≥ 50 years will experience an osteoporotic fracture. [21-23] Osteoporosis is often not detected until fracture presentation and is hence considered a “silent epidemic” with a need for early diagnosis. [24, 25]

Raising awareness of osteoporosis increased treatment and treatment compliance rates, [95] so awareness of osteoporosis may be the most effective strategy for prevention of osteoporotic fracture. [96] In general, osteoporosis is diagnosed using BMD measured on central dual energy X-ray absorptiometry (DXA), which is considered a reference standard test. [26] However, the application of DXA is complex and expensive, and has limited availability for diagnosing the entire population. [97] Another barriers to DXA screening is the absence of symptoms, which leads to low demand and reduced financial incentives for screening. As a result, nearly half of women covered by Medicare in the United States do not undergo DXA; even certain high-risk populations have low screening rates of ≤ 10%. [98, 99] Likewise, despite the reasonable prices and accessibility of DXA in South Korea, KNHANES IV 2008–2009 found that only 37.5% among the Korean general female population aged ≥ 50 years with osteoporosis were aware of their diagnosis and only 23.5% were under treatment for osteoporosis. [100]

3. Exploring extraction of demographic information and interpretation

In this chapter, age assessment models and sex classification models developed. These models were trained and tested in the clinical dataset of radiology department on more than

90000 radiographs. In addition, the training ability of these models in a limited number was confirmed. The best age assessment model was applied to the clinical dataset and the results were analyzed. Chest radiography may provide information about longevity and prognosis. [11] Single chest radiograph can give basis of medical decision-making by predicting age. [87] Chronologic age is an imperfect measure of longevity, which defined the number of years since birth but training model for predicting it will foundation to tuning other models via transfer learning.

3.1. Dataset

A large number of chest radiographs were collected in the department of radiology of Asan Medical Center (AMC) between January 2011 and December 2018. The original dataset was cleaned as illustrated in Figure 3.1. Normal chest x-ray images were classified using diagnostic codes. This study was conducted on chest x-rays of adults aged 19 years and older, and the images were included solely from fixed radiography systems of GE Healthcare. This ensured the control of domain shift due to various types of x-ray equipment. Furthermore, chest radiographs posteroanterior (PA) were acquired by removing other chest images because the original dataset contained various types of chest images, such as chest lateral images and chest decubitus images, which can be only differentiated by using Digital Imaging and Communication in Medicine (DICOM) fields. This was further confirmed by an expert radiologist. Finally, we selected them between the ages of 20 and 80 and the number of chest images was 90785. The reason for excluding videos over the age of 80 is because the number of data was too small to reflect that age group. The dataset age distribution is shown in Figure 3.2. The labels which were age and gender, were extracted from header information of DICOM files such as study date, patient's birth date, and patient's sex.

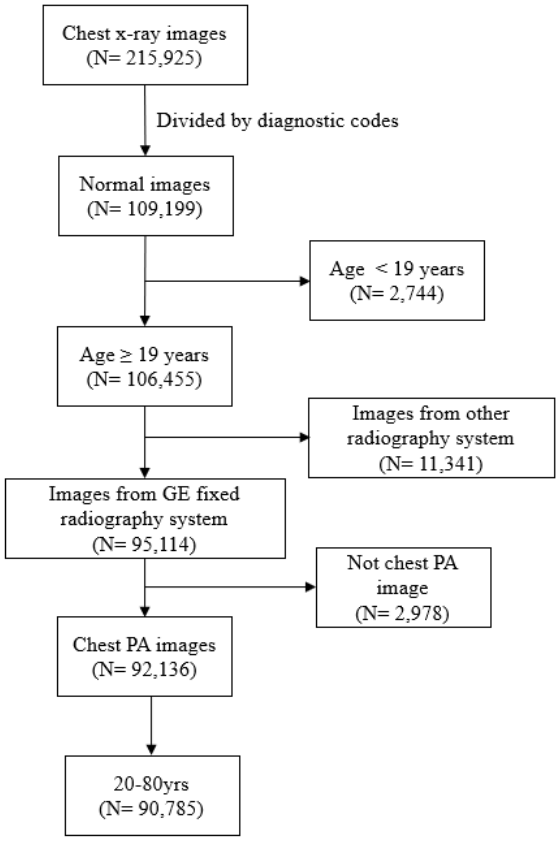


Figure 3.1. Flowchart of normal chest radiographs acquisition and cleansing.

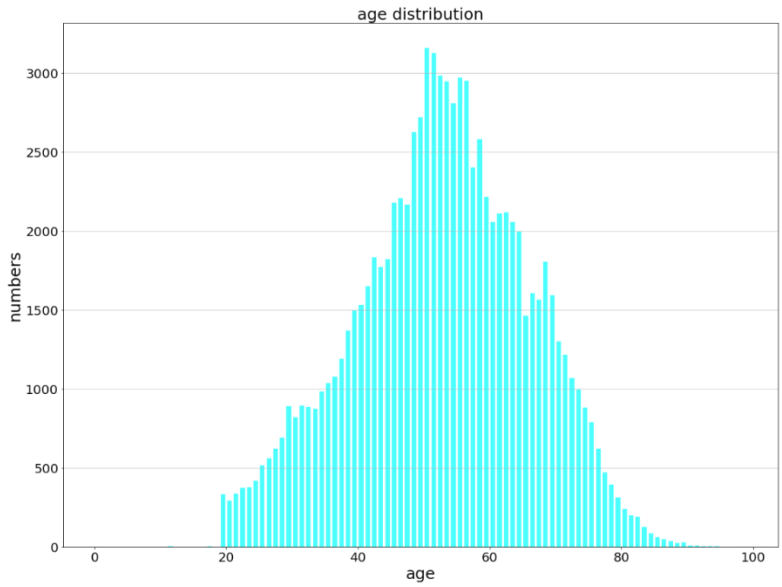


Figure 3.2. Age distribution of the dataset.

3.2. preprocessing

For each image, two simple pre-processing methods were applied. First, min-max normalization with 0.5% clipping of upper and lower bounds was performed to suppress the effect of L/R mark in radiographs and remove the outlier pixel values. A set of pixel values of original and scaled images is represented by X, Z respectively; the formula of min-max normalization is

$$Z = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (6)$$

Second, due to the limitation of GPU resources and experiments of image sizes, all images were resized down into 224 × 224, 512 × 512, 1024 × 1024 by bi-cubic interpolation and converting from DICOM to 8-bit Portable Network Graphics (PNG) format files. We randomly selected 63549, 9078, and 18158 chest radiographs so the proportions of the training, validation, and test datasets were split by 70%, 10%, and 20%. Ubuntu 18.04 with a 24GB TitanRTX, CUDA 10.0 (NVidia Corporation), TensorFlow 1.15.0, and Keras 2.3.0 were used as the experimental environments.

3.3. Age and sex prediction on chest radiographs

3.3.1. Development of age assessment model and evaluation

We trained DenseNet-169 [68] for age assessment. [101] DenseNet in Figure 3.3. is configured in a dense block involving four BN-ReLU-Conv modules. Colored squares are feature maps generated at different stages. The convolution layer was the most common convolution.

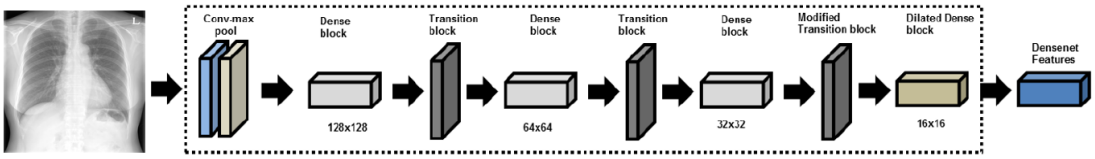


Figure 3.3. Architecture for DenseNet: DenseNet has three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature map sizes via convolution and pooling.

The model was implemented in Keras with a TensorFlow backbone and adaptive moment estimation (Adam)[102] optimizer with an initial learning rate of 0.001, which optimizes learning rates efficiently during training. The regression task loss is defined as square error loss (MSE loss), and given as follows:

$$Mean\ Square\ Error\ Loss(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{y}_i]^2 \qquad (7)$$

, where $\hat{y}$ is the probability of model output and y is the ground truth. Evaluation matric used R2 score, ± 4 years age error range, and ± 9 years age error range. All statistical analysis was performed using scikit-learn. Performance of 5 CNN models trained with various image sizes and batch sizes are showed in Table 3.1. All models have R^2 score of 89 or higher. The scatter plots show the results well in Figure 3.4.

Table 3.1 Performances of 5 DenseNet-169 models trained with various image sizes and percentage of cases predicted correctly within the age error ranges.

Input image size	Batch size	R^2	± 4 years (%)	± 9 years (%)
224	100	89.20	67.83	96.66
224	20	90.03	69.61	97.27
512	20	91.79	73.38	98.31
512	4	90.40	69.93	97.39
1024	4	89.20	67.60	96.79

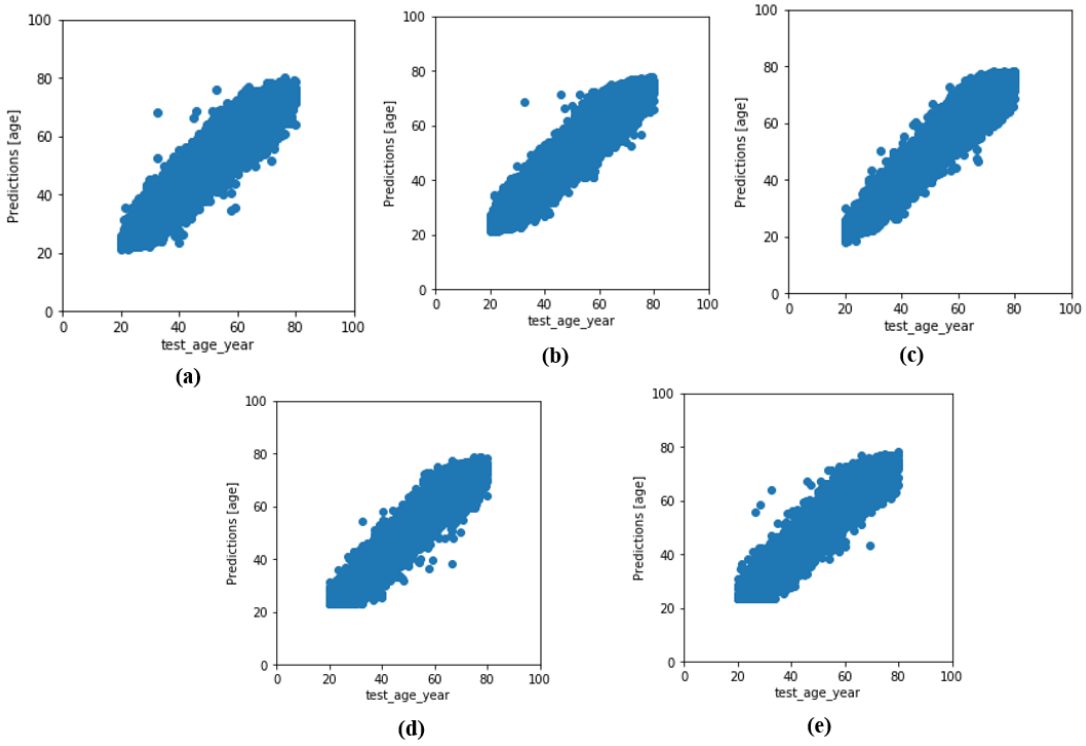


Figure 3.4. Scatter plots of 5 DenseNet-159 models trained with various image sizes; (a) 224 × 224 size and batch size 100, (b) 224 × 224 size and batch size 20, (c) 512 × 512 size and batch size 20, (d) 512 × 512 size and batch size 4, (e) 1024 × 1024 size and batch size 4.

The number of training data was very large, more than 70000, and chest radiographs are normal. All 5 models were trained at 30 epochs and showed better results than previously known results trained by the NIH ChestX-ray8. [101] From our experimental results, it was found that the image size suitable for predicting age was 512 × 512 and the batch size was 20. Additional experiment was conducted using the inceptionV3 [103]. Figure 3.5. shows the Inception network structure. Over-fitting can be avoided and this network's expression and adaptability to multiple scales can be improved in this approach.

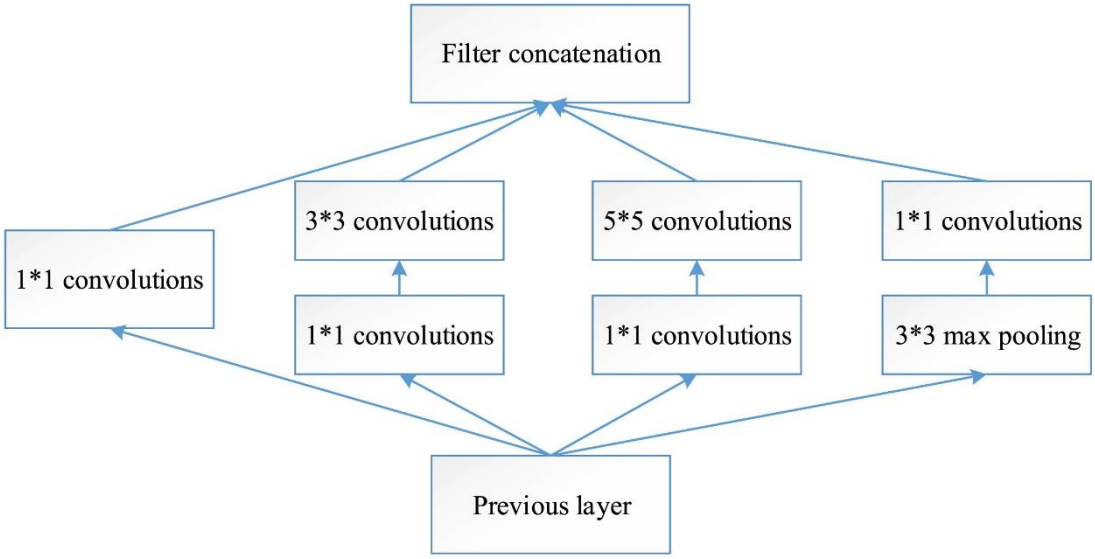


Figure 3.5. Inception network structure. Three different sizes of convolution and one maximum pooling are typically contained in the Inception module. The channel is aggregated following the convolution process for the previous layer's network output, and then nonlinear fusion is conducted.

Age predicting inceptionV3 was trained image size of 512 × 512 and batch size 20. This model presented best results with 92.04 of R^2 score, 74.18% accuracy ±4 years of error range and 98.38% accuracy ± 9 years error range. The scatter plots show the results in Figure 3.6. When the ResNet-50 model was trained as an additional experiment, it was not trained to an appropriate level.

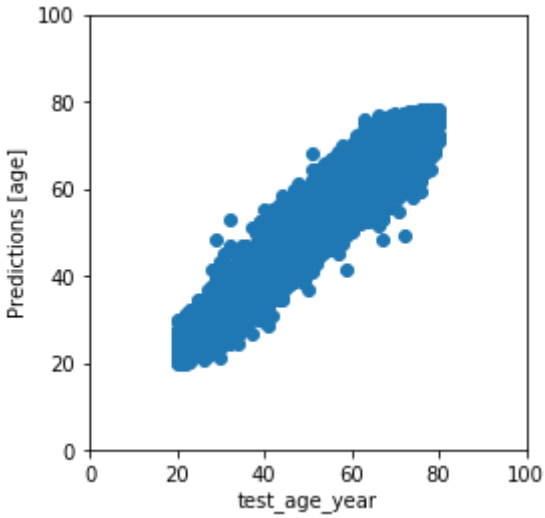


Figure 3.6. Scatter plots of the best result by the InceptionV3.

3.3.2. Development of sex classification model and evaluation

We trained ResNet-50 [67] for sex classification. ResNet-50 is a 50-layer residual network in short form and has an additional identity mapping capability in Figure 3.7.

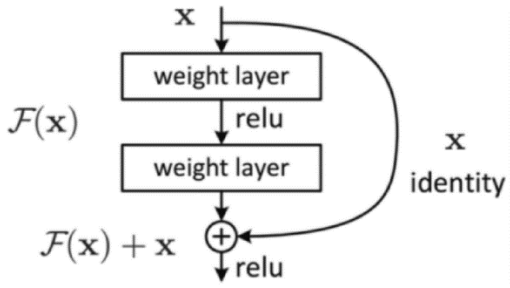


Figure 3.7. ResNet predicts the delta that is required to get the final prediction from one layer to the next. ResNet solves the vanishing gradient problem by allowing this alternate shortcut path for gradient to flow through. The identity mapping used in ResNet allows the model to bypass a CNN weight layer if the current layer is not necessary. This helps in avoiding the over fitting problem to the training set.

The model was implemented in Keras with a TensorFlow backbone and adaptive moment estimation (Adam) [102] optimizer with an initial learning rate of 0.001, which optimizes learning rates efficiently during training. The binary classification task loss is defined as the binary cross-entropy loss (BCE loss) as follows:

$$Binary\ Cross\ Entropy\ Loss(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})] \quad (8)$$

, where $\hat{y}$ is the probability of model output and y is the ground truth. We selected 512 × 512 sized images according to results of age assessment. Model performance was evaluated using area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} = Recall \quad (8)$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (9)$$

$$Accuracy = \frac{True\ Postivie + True\ Negative}{True\ Postive + True\ Nagative + False\ Posive + False\ Negative} \quad (10)$$

Performance of the CNN model trained are showed the receiver operating characteristic (ROC) curve and confusion matrix in Figure 3.8. The model achieved AUC 0.9989, sensitivity 99.55%, specificity 99.77%, and accuracy 99.14%. Because chest radiographs show the differences between men and women, radiologists know the difference. Therefore, the model completed training in 15 epochs on such a large amount of data and showed high performance.

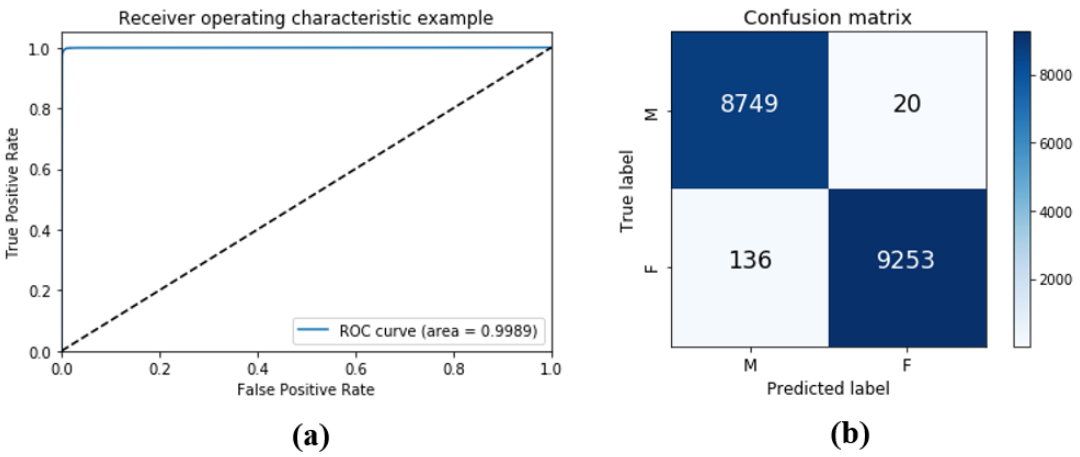


Figure 3.8. Performance of the sex classification model;(a) Roc curve, (b) confusion matrix

3.3.3. Stress study

We exploit the effect of a stress test on the number of training dataset for age assessment and sex classification. For a stress test of age assessment, number of datasets were subsampled that training and tuning datasets were selected 9000 and 1000 chest radiographs, respectively. Test dataset was maintained as it is. As presented Figure 3.9, the age distributions of each dataset are almost identical to each other. By training from a small dataset of about 10% of the original dataset using InceptionV3, performances showed decrease in performance with 84.05 of R^2 score, 57.11% accuracy ±4 years of error range, 92.55% accuracy ± 9 years error range. In particular, percent of accuracy ±4 years of error range was decreased by more than 15% compared with result of training original dataset.

For stress tests of sex classification, numbers of datasets were subsampled that training and tuning datasets were selected 800 and 100 chest radiographs, respectively. Recognizing sex in chest radiographs is possible to some extent even with human cognition, and it was judged that it is a relatively easy task to train by deep learning. We experimented to train

smaller dataset. Test dataset was maintained as it is. ResNet-50 and Inceptionv3 model trained subsampled datasets and two models also presented over 0.95 of AUC in Figure 3.10. However, performance of InceptionV3 model was better than that of ResNet50. Details of performance of two models are presented in Table 3.2.

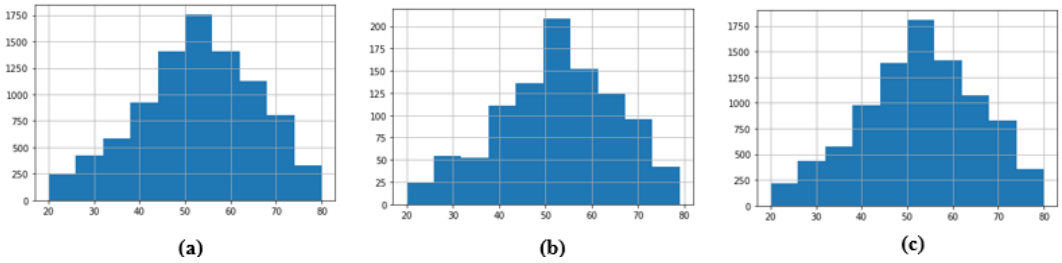


Figure 3.9. age distributions; (a) training dataset, (b) tuning dataset, (c) test dataset.

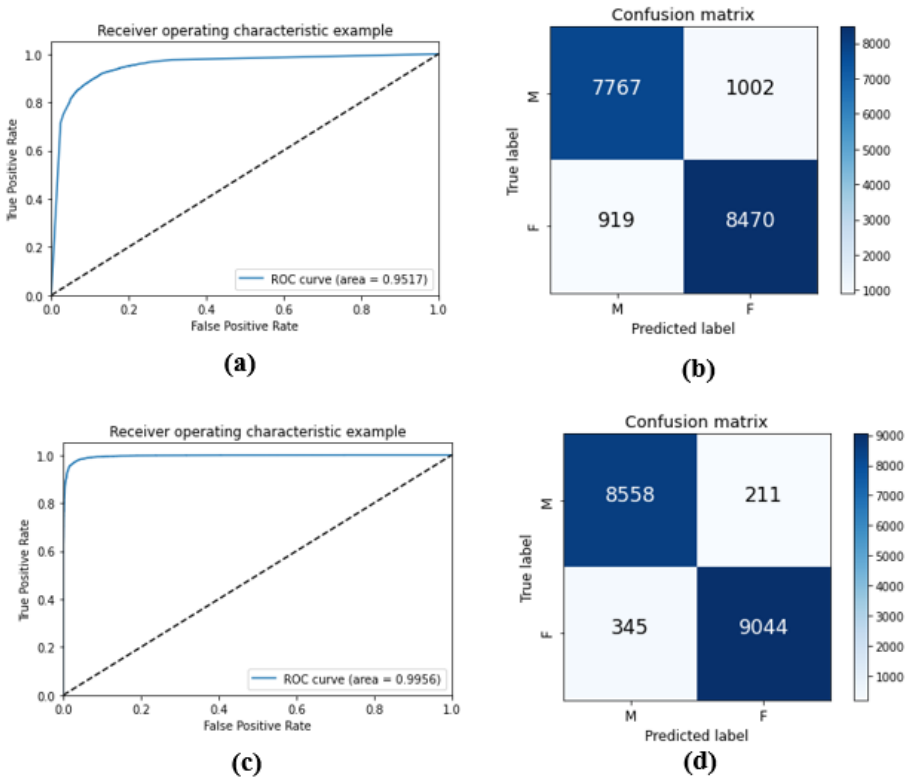


Figure 3.10. Performance of the sex classification model in 800 images training and 100 images tuning dataset;(a) Roc curve of ResNet-50 model, (b) confusion matrix of ResNet-50

model, (c) Roc curve of InceptionV3 model, (d) confusion matrix of InceptionV3 model.

Table 3.2. Performance of the sex classification models in 800 images training and 100 images tuning dataset.

	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)
ResNet-50	0.9517	89.42	90.21	89.42
InceptionV3	0.9956	96.94	96.33	96.12

As a result of checking the performance on a smaller training dataset using InceptionV3, the results were as follows in Figure 3.11. and Table 3.3.

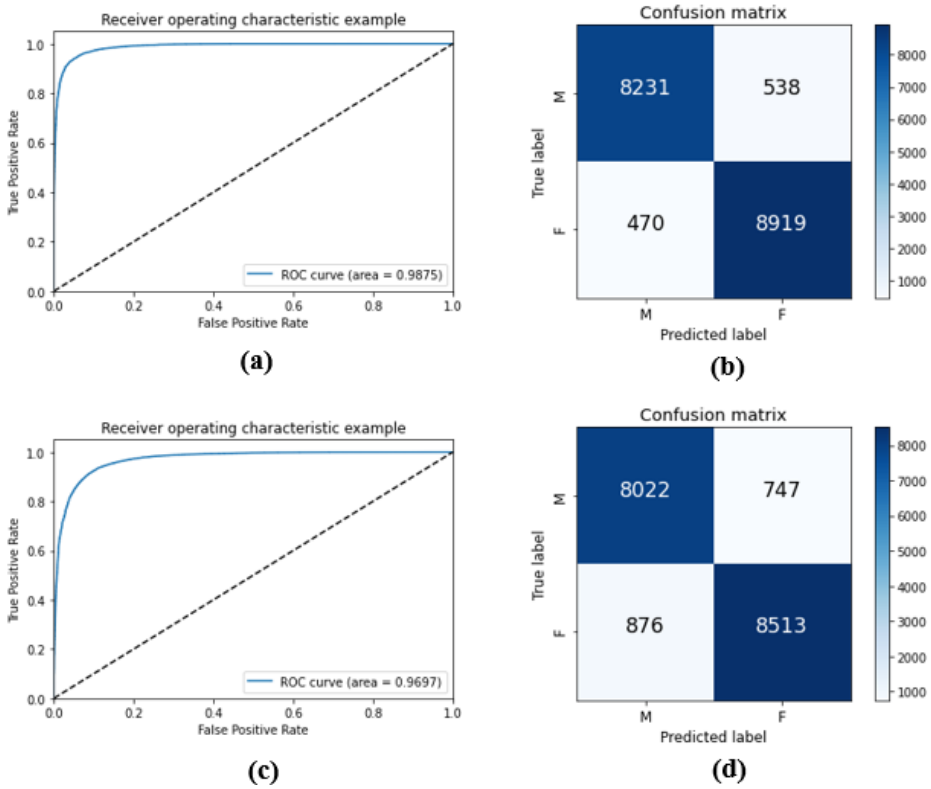


Figure 3.11. Performance of the sex classification model using InceptionV3 model in a smaller dataset; (a) Roc curve of 450 images training and 50 images tuning datasets, (b) confusion

matrix of 450 images training and 50 images tuning datasets, (c) Roc curve of 180 images training and 20 images tuning datasets, (d) confusion matrix of 180 images training and 20 images tuning datasets.

Table 3.3. Performance of the sex classification models using InceptionV3 in smaller datasets.

	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)
500 images training and tuning datasets	0.9875	94.49	94.99	94.60
200 images training and tuning datasets	0.9697	91.16	90.67	90.16

3.3.4 Application of age assessment model in clinical data

For application of the age assessment model and interpretation of the results, the model tested in the external test dataset using data from osteoporosis screening model development dataset. The dataset consisted of 13026 images and included heights, weights, blood test results, DXA results, medication history, and fracture history from health check-up of AMC health promotion center. This model presented results with 78.11 of R^2 score, 73.71% accuracy ±4 years of error range and 98.31% accuracy ± 9 years error range. Two histograms of age and predicted age are presented in Figure 3.12. and scatter plots according to the drug history are showed in Figure 3.13.

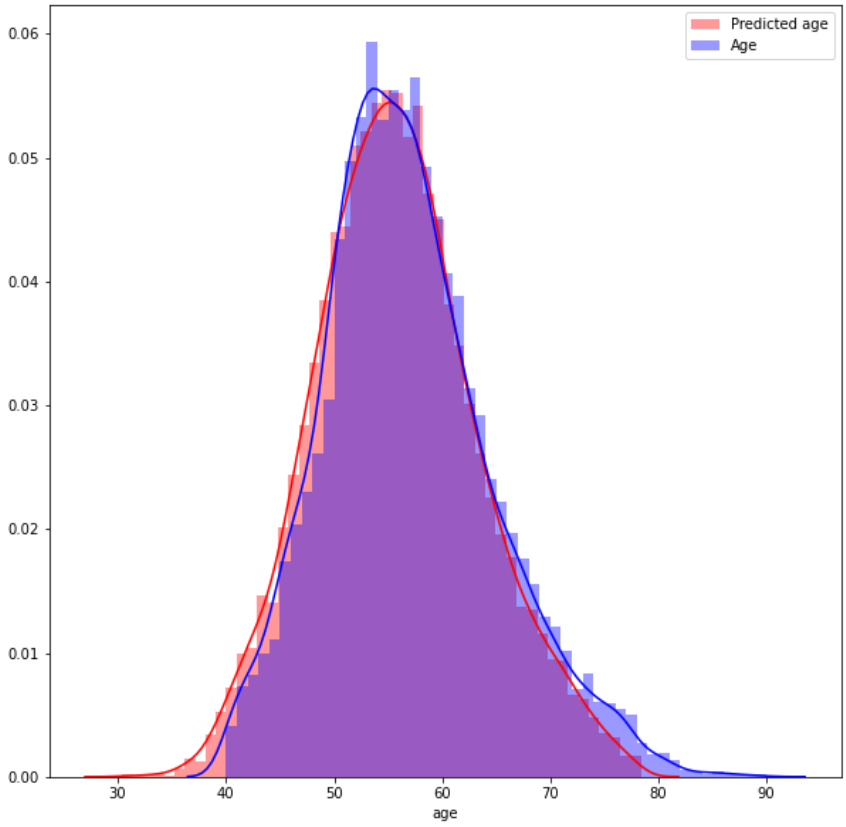


Figure 3.12. Histograms of chronological age and predicted age in external test set.

As a result of the test, even if the actual age was over 80, it was predicted to be under 80. Because this model was trained on data under 80 years of age, it is thought that the predictive power of the model is lowered for those older than 80 years of age. Therefore, the analysis due to the difference between the actual age and the predicted age was conducted only for those under 80 years of age. The total data used for the analysis was 12972, and the baseline characteristics are presented in Table 3.4. The result value was classified into the old-predicted group when the predicted age was 5 years or more older than the actual age, the young-predicted group when the predicted age was 5 years younger, and the well-predicted group in between.

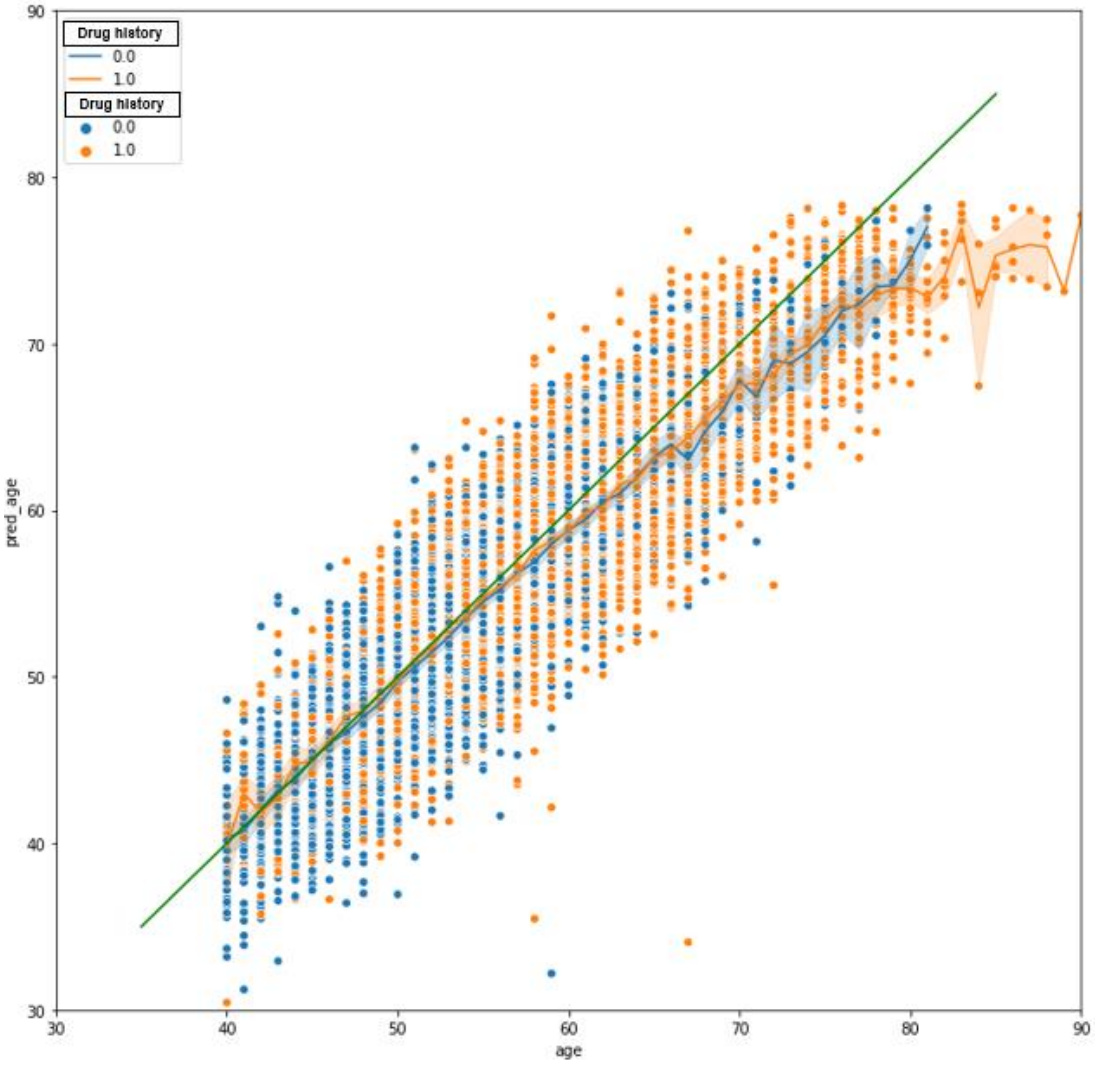


Figure 3.13. Scatter plots of chronological age and predicted age according to the medication history.

Table 3.4. Baseline characteristics.

Characteristics		N (%)
Sex	Female	10794 (83.21)
BMD groups	osteoporosis	4301 (33.16)
	osteopenia	4328 (33.36)
Mediation history	yes	6975 (53.77)
Fracture history	yes	5519 (42.55)
Hypertension drug user	yes	2606 (20.09)
Dyslipidemia drug user	yes	2216 (17.08)
Anti-PLT drugs user	yes	924 (7.12)
Diabetes Miletus	yes	1234 (9.51)
		Mean ± standard deviation
BMI		22.66 ± 2.95
Heights (cm)		159.83 ± 6.79
Weights (kg)		58.00 ± 9.32
Age (year)		56.71 ± 7.65
Hb		13.52 ± 1.29
HbA1c		5.64 ± 0.63
Calcium		9.29 ± 0.43
Phosphorus		3.55 ± 0.53
Creatinine		0.73 ± 0.19
BUN		12.95 ± 3.76
AST		26.65 ± 13.73
ALT		22.54 ± 15.6
Total protein		7.32 ± 0.44
Albumin		4.09 ± 0.27
ALP		66.17 ± 22.55
K		4.12 ± 0.37
Na		142.31 ± 2.25
S-mass		1.01 ± 1.14
S-Tscore		-1.21 ± 1.58
F-mass		0.85 ± 1.34
F-Tscore		-0.83 ± 1.04

BMI (body mass index), Hb (hemoglobin), FBS (fasting blood sugar), HbA1c (glycated hemoglobin), BUN (blood urea nitrogen), AST (aspartate aminotransferase), ALT (alanine

aminotransferase), ALP (alkaline phosphatase), S-mass (bone mass from lumbar spine), S-Tscore (T-score from lumbar spine), F-mass (bone mass from total femur), F-Tscore (T-score from total femur)

A total of 12972, 499 of old-predicted, 10907 of well-predicted, and 1566 of young-predicted groups were classified. As for the sex distribution, the proportion of males was higher in the old-predicted and young-predicted groups based on the well-predicted group in Table 3.5. The mean of chronological age was the lowest in old-predicted and the most in young-predicted groups in Table 3.6 and the density distribution plot of chronological age is showed Figure 3.14. The proportions of people taking any drugs and people with any fractures were highest in the order of young-predicted, old-predicted, and well-predicted groups. People with diabetes and those taking dyslipidemia drugs were the highest in the young-predicted group. However, the proportion of hypertension drug users was higher in the order of old-predicted, young-predicted, and well-predicted groups. The proportion of osteoporosis patients was higher in the order of old-predicted, well-predicted, and young-predicted groups.

The mean of blood tests results, such as hemoglobin (Hb), fasting blood sugar (FBS), glycated hemoglobin (HbA1c), blood urea nitrogen (BUN), aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase (ALP), serum sodium (Na), serum potassium (K), etc. were within normal limits. The mean of body mass index (BMI) in 3 groups was less than 23 kg/m^2. T-scores from lumbar spine and total femur were lower in the order of old-predicted, well-predicted, and young-predicted groups like proportion of osteoporosis patients. Density distribution plot of BMI, T-score from lumbar spine, and T-score from total femur are showed Figure 3.15, Figure 3.16, and Figure 3.17, respectively.

Table 3.5. Number of subjects and percentage of 3 groups in categorical variables.

	Sex	BMD groups			medication history	Fracture history
N (%)	Female	normal	osteopenia	osteoporosis	yes	yes
Old-predicted	397 (79.56)	124 (24.85)	155 (31.06)	220 (44.09)	280 (56.11)	223 (44.69)
Well-predicted	9142 (83.82)	3650 (33.46)	3643 (33.40)	3614 (33.13)	5732 (52.55)	4584 (42.03)
Young-predicted	1255 (80.14)	559 (35.70)	530 (33.84)	467 (29.82)	963 (61.49)	712 (45.47)
total	10794(83.21)	4333(33.40)	4328 (33.36)	4301(33.16)	6975 (53.77)	5519 (42.54)
p-value*	<0.001		<0.001		<0.001	0.013

	Hypertension drug user	Dyslipidemia drug user	Anti-platelet drug user	Diabetes Miletus	
N (%)	yes	yes	yes	yes	total
Old-predicted	132 (26.45)	76 (15.23)	49 (9.87)	43 (8.62)	499
Well-predicted	2123 (19.46)	1744 (15.99)	721 (6.61)	975 (8.94)	10907
Young-predicted	351 (22.41)	396 (25.29)	154 (9.83)	216 (13.79)	1566
total	2606 (20.89)	2216(17.08)	924 (7.12)	1234 (9.51)	12972
p-value*	<0.001	<0.001	<0.001	<0.001	

p-value* calculated by chi-square test

Table 3.6. Mean and standard deviation (SD) of 3 groups in continuous variables.

Mean ± SD	Age (year)	Heights (cm)	Weights (kg)	BMI (kg/m^2)	Hb (g/dL)	FBS (mg/dL)	HbA1c (%)	Calcium (mg/dL)	Phosphorus (mg/dL)	Creatinine (mg/dL)	BUN (mg/dL)
Old-predicted	53.95 ± 6.19	159.71 ± 7.27	57.48 ± 9.58	22.86 ± 2.94	13.60 ± 1.31	97.21 ± 17.12	5.60 ± 0.67	9.32 ± 0.44	3.58 ± 0.50	0.72 ± 0.15	12.65 ± 3.39
Well-predicted	56.16 ± 7.28	159.82 ± 6.72	57.91 ± 9.20	22.62 ± 2.90	13.52 ± 1.29	98.35 ± 18.32	5.63 ± 0.62	9.29 ± 0.44	3.56 ± 0.53	0.73 ± 0.18	12.87 ± 3.73
Young-predicted	61.45 ± 8.75	159.92 ± 7.09	58.5 ± 10.03	22.83 ± 3.29	13.53 ± 1.25	100.86 ± 19.39	5.72 ± 0.68	9.32 ± 0.41	3.54 ± 0.53	0.77 ± 0.26	13.56 ± 4.06
p-value†	<0.001	0.698	0.077	0.022	0.281	<0.001	<0.001	0.011	0.383	<0.001	<0.001

Mean ± SD	AST (U/L)	ALT (U/L)	total protein (g/dL)	Albumin (g/dL)	ALP (U/L)	K (mEq/L)	Na (mEq/L)	S-mass (g/cm^2)	S-Tscore	F-mass (g/cm^2)	F-Tscore
Old-predicted	28.30 ± 17.09	24.40 ± 26.96	7.33 ± 0.42	4.12 ± 0.26	70.21 ± 31.57	4.09 ± 0.39	142.44 ± 2.27	0.95 ± 0.20	-1.58 ± 1.50	0.81 ± 0.14	-1.06 ± 1.02
Well-predicted	26.44 ± 13.02	22.29 ± 14.84	7.32 ± 0.44	4.09 ± 0.27	66.22 ± 22.55	4.11 ± 0.37	142.30 ± 2.26	0.99 ± 0.21	-1.21 ± 1.59	0.83 ± 0.16	-0.82 ± 1.04
Young-predicted	27.60 ± 16.90	23.69 ± 15.71	7.31 ± 0.43	4.06 ± 0.28	64.55 ± 18.58	4.18 ± 0.39	142.29 ± 2.21	1.01 ± 0.22	-1.09 ± 1.52	0.84 ± 0.20	-0.80 ± 1.06
p-value†	<0.001	<0.001	0.300	<0.001	<0.001	<0.001	0.179	<0.001	<0.001	<0.001	<0.001

p-value† calculated by Kruskal-Wallis test. BMI (body mass index), Hb (hemoglobin), FBS (fasting blood sugar), HbA1c (glycated hemoglobin), BUN (blood urea nitrogen), AST (asparate aminotransferase), ALT (alanine aminotransferase), ALP (alkaline phosphatase), S-mass (bone mass from lumbar spine), S-Tscore (T-score from lumbar spine), F-mass (bone mass from total femur), F-Tscore (T-score from total femur)

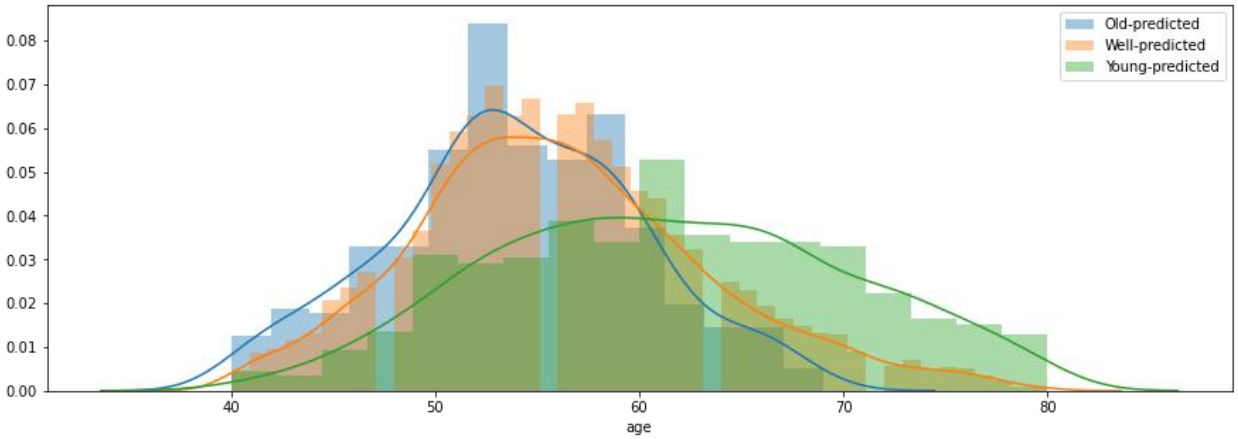


Figure 3.14. Density distributions of chronological age in 3 groups.

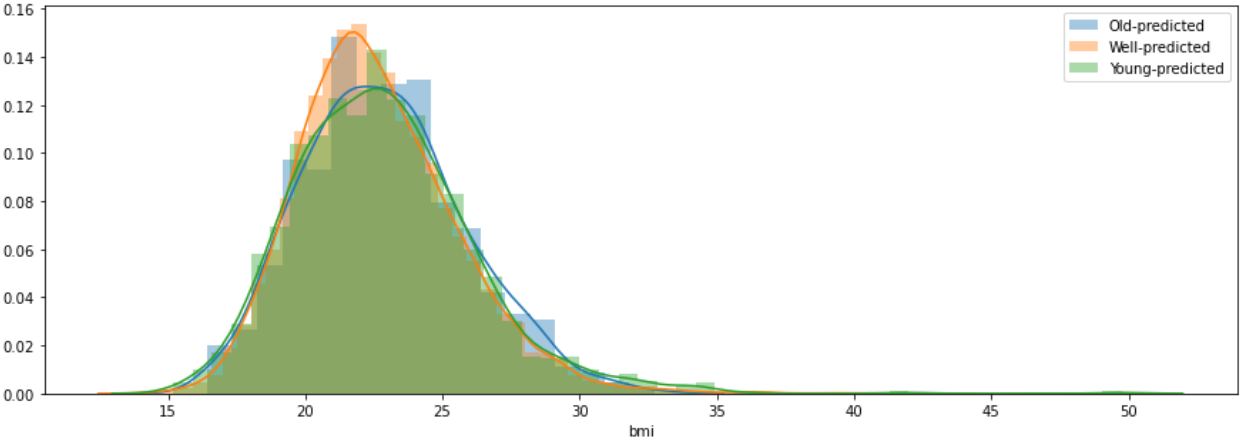


Figure 3.15. Density distributions of BMI in 3 groups.

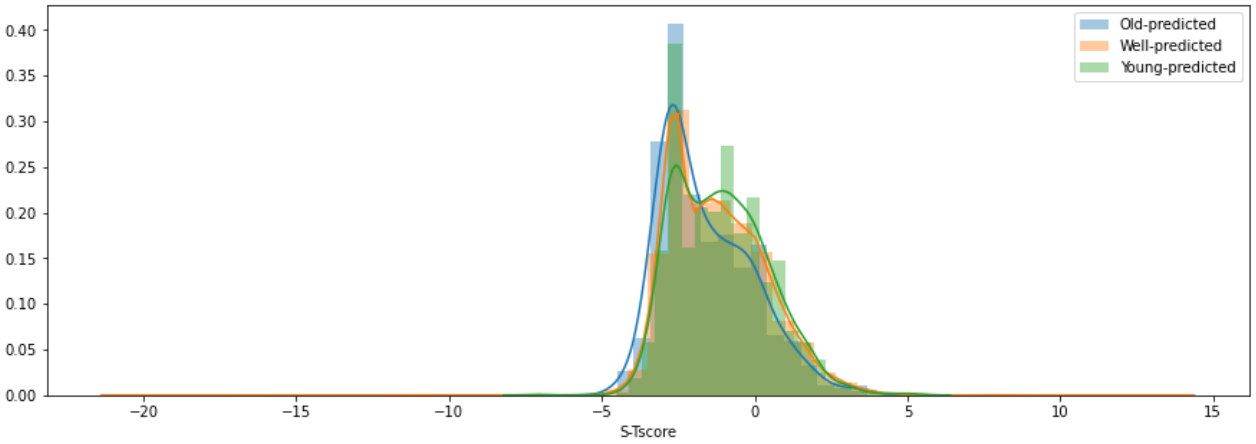


Figure 3.16. Density distributions of T-score from lumbar spine in 3 groups

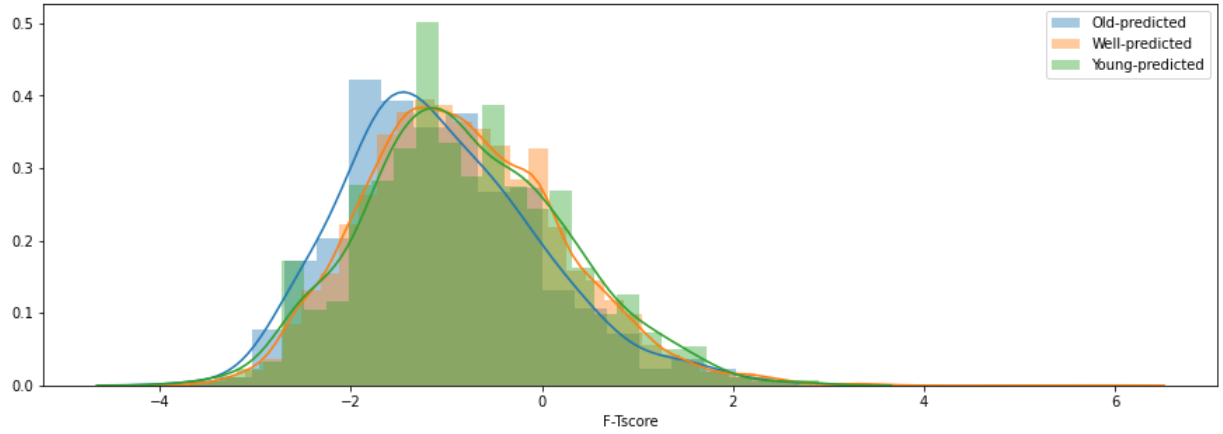


Figure 3.17. Density distributions of T-score from total femur in 3 groups.

3.4. Discussion

For decades, numerous medical imaging techniques have been used to predict age and sex. Neuroimaging-derived age prediction has been found to correspond with influences from other disorders, such as cognitive impairment and Alzheimer disease, in addition to age and sex per se. [104-106] T-scores acquired from DXA scans of bone density in both sexes can be used to predict osteoporosis and fracture risk. T-scores decline with age and can be used to predict osteoporosis and fracture risk. [107] These previous investigations demonstrate that certain medical imaging modalities have characteristics that are related to chronological age and gender. Based on chest radiographs classified normal ranges of adults in a hospital-based real-world clinical dataset, this study developed a deep learning model for age and sex prediction. And the best age assessment model was applied in the clinical dataset. The best prediction models were AUC of 0.9989 for sex classification and R^2 score of 92.04 for age assessment.

In previous studies, CNN can predict sex on chest radiographs with a very high degree of accuracy. [14, 108-110] these studies used public datasets such as CheXpert and NIH chest X-ray14 or about 60000 chest images dataset. On the stress tests, InceptionV3 presented better results than ResNet-50 and these results were obtained consistent with the results of previous studies. InceptionV3 with 90% accuracy was developed by training only 200 chest radiographs for sex classification. Recent work [109, 110] has found that even when protected attributes, such as sex and race are not explicitly input to the model, CNN can be biased. [14] In general, it is well known that CNNs tend to learn representations useful to solve the task they are being trained for. If CNNs can properly detect sex, they may be able to make judgments based on sex-specific features. Li et al. [14] previously showed that CAM heatmaps were qualitatively similar across CNN architectures and two large public chest radiograph-datasets. CNN performance on disease diagnosis has already been shown to be a reflection of their capacity

to exploit confounding information, such as the hospital or unit where the image was obtained [111], which could theoretically apply to demographic variables, such as sex. As a result of these findings, researchers will need to take demographic representations into account in datasets in order to construct fair deep learning models.

Larson et al. [112] developed a hand-age assessment model with a mean absolute error (MAE) of 0.5 years and root mean absolute error (RMSE) of 0.63 years, with a similar method. Instead of chronological age, the evaluation target of the hand age study was skeletal maturity, defined by image findings, and the ground truth was also determined by human reviewers based on images. Therefore, compared to the prediction of chronological age, a model predicting skeletal maturity may be easier to comprehend because the evaluation target and ground truth are both image-based. In another brain age prediction study [113], the models were trained using a brain MRI dataset (N = 2001) with an MAE > 4 years and RMSE > 5 years. Although speculative, it seems possible that our model, demonstrating greater accuracy, may extract more relevant features for age prediction from the chest radiographs than those extracted in brain MR images.

Chronological age has been widely used in various studies to predict disease and treatment prognosis. Chest radiographs findings may reflect the overall effects of chronological, biological, and/or pathological changes. Chronological changes reflect natural development over time, e.g., endochondral ossification in skeletal evolution, changes in body shape, or breast growth related to hormones. Biological changes indicate cumulative effects resulting from interactions between the body and environment or lifestyle, e.g., obesity-related to an unhealthy diet and osteoporosis-related to diet and a lack of exercise. Beyond the physiological changes, including chronological and biological ones, pathological changes can be related to a certain disease, cohort, or sex-specific issue, e.g., cancer growth and chronic tuberculosis sequela usually seen in elderly persons and breast cancer mostly seen in females.

Neuroimaging-derived age predictions have been explored in a variety of brain illnesses, and the disparities between predicted brain and chronological ages may be attributed to the accumulation of age-related alterations in pathological circumstances [113-115] or protective factors on brain aging [116, 117]. Chest radiograph-derived age obtained, meantime, might be used as an imaging biomarker to indicate the thorax's state or something of bone metabolism. In fact, studies using Chest radiograph-derived age to successfully predict lifespan, mortality, cardiovascular risk, and heart failure prognosis have recently emerged [11, 87, 118, 119], providing a solid foundation for the imaging biomarker concept.

This clinical dataset collected from AMC health promotion center could consist of relatively healthy people. In subjects over the age of 40, the average BMI was less than 23, well managed, and there were diabetic patients, but the average HbA1c was about 5.7. The chest radiograph-derived age for these subjects was mostly young. Taking drugs can be inferred as having a disease to be managed. Nevertheless, the high distribution of diabetic patients and dyslipidemia drug users in the young-predicted group for which chest radiograph-derived age was predicted indicates that management has a greater correlation with prognosis. In old-predicted group, the proportion of osteoporosis patients was high even though the chronological age was young, indicating the possibility that chest radiograph-derived age was also related to bone density.

This study has several limitations. First, we used a dataset comprising with classified diagnostic codes from a single medical center, which might introduce some potential biases in data collection. Additional multi-center and multi-nation studies should be conducted to test the generalizability of our findings. To reduce such biases, more representative image data with reasonable variation should be examined for training. Second, while best age assessment model showed 92.04 of R^2 score in this study, clinical settings require a more exact and precise prediction for each person. Yang et.al showed the model trained 59979 chest radiographs of

heathy adults which had a 2.1-year MAE and a 2.8-year RMSE. [108] Additional efforts with optimization of data manipulation or deep learning algorithms are needed to improve the performance of the model. Third, in order to know exactly what chest radiograph-derived age implies, it is necessary to learn chronological age from a big data set composed only of healthy people and then obtain the results of age assessment for a specific disease group. Currently, the age assessment model in this study was more than 90000 data collected from the radiology department, and the reports of chest radiographs were within normal limits, but information about other underlying diseases was unknown. Nevertheless, as chest radiograph-derived age has sufficiently shown the potential to be related to several diseases, further study is needed through securing additional datasets in the future.

4. Development osteoporosis screening model on chest radiographs

In this chapter, osteoporosis screening models were developed using various deep learning methods with health check-up dataset of health promotion center. In Experiment 1, the specific area of chest radiographs and effective matrix size were measured in order to develop screening model for osteoporosis. In Experiment 2, we developed a deep learning model that just uses chest radiographs, a deep learning model that uses both chest radiographs and demographic data, and a deep learning model that uses both chest radiographs and chest radiograph-derived age. With weights of the previously trained sub-group classification model, age assessment model, and sex classification model, we constructed an osteoporosis screening model using transfer learning. In Experiment 3, we compared performance of deep learning models with only using chest radiographs and using both images and demographic information

in stress tests using a small dataset. Finally, interpretation of the results of the best-performing model using Grad-CAM and its application to clinical sites are described.

4.1. Dataset

This retrospective study was conducted in accordance with the principles of the Declaration of Helsinki and was performed in accordance with current scientific guidelines. The study protocol was approved by the institutional review board of AMC, University of Ulsan College of Medicine, Seoul, Korea (IRB No 2019-1226), which waived the requirement for informed patient consent.

The data of participants aged > 40 years from medical health check-ups at the Health Screening and Promotion Center of AMC between January 2012 and February 2019, were used for this study. The medical health check-ups included surveys, physical examinations, laboratory testing, and medical imaging. This study was conducted using paired chest radiographs with normal results and DXA examinations on the same day. Areal BMD (g/cm2) was measured at the lumbar spine (LS), femoral neck (FN), and total hip (TH) using DXA (Lunar system running software version 9.30.044; Prodigy, Madison, WI). The participants were classified into normal, osteopenia, and osteoporosis groups by the lowest T-score at LS, FN, or TH according to the WHO criteria. [26] Osteoporosis was defined by the lowest T-score of ≤ -2.5, osteopenia by -1.0 < the lowest T-score < -2.5, and normal as the lowest T-score ≥ -1.0. Osteoporosis, osteopenia, and normal diagnoses were regarded as ground truth and used as weak labels for supervised learning.

The dataset was selected and divided into training, tuning and validation sets, including several chest radiographs and DXA results from periodic check-ups. The proportions of the training, tuning, and validation datasets were split by 70%, 10%, and 20%, by avoiding data

from the same individuals but different examination dates were not split into different datasets. We randomly selected the three labels (i.e., osteoporosis, osteopenia, and normal) with 9825 and 1212 chest radiographs in the training and tuning datasets balanced with the number of data and sex ratio, respectively. In the internal validation dataset, we randomly selected 1989 chest radiographs of 50-year-old individuals balanced with the number of data and sex ratio.

For the external validation set chest radiographs were obtained from the ordered outpatient, inpatient, emergency departments and Health Screening and Promotion Center, from June 2006 to July 2019 in the Asan osteoporosis cohort, which consisted of consecutive ambulatory men and postmenopausal women who visited the AMC osteoporosis clinic between January 2010 and October 2017. [120] The inclusion criteria for the external validation dataset were as follows: (1) 50-year-old individuals who underwent both chest radiography and DXA within a 3-month period, (2) posteroanterior chest radiographs of results within the normal ranges, and (3) individuals with chest radiographs only from the Health Screening and Promotion Center before 2012. The exclusion criteria were as follows: (1) patients with medical devices such as electrocardiographic lines, a pacemaker, or an implantable defibrillator, (2) patients who underwent operations, such as internal fixation and bone cement filling, (3) patients with abnormal chest imaging results, and (4) patients with low-quality images. The final inclusion configuration of the Asan osteoporosis cohort data is shown in Table 4.1. In the external validation dataset, 340 healthy individuals from the primary validation dataset who had independently obtained DXAs were added in external validation dataset, because patients with normal DXA findings were rare in the Asan cohort dataset. Figure 4.1. demonstrates the flowchart of the configurations of the four datasets, which ensure that the validation dataset only contained images of novel chest radiographs that had not been encountered by the model during training.

Table 4.1. Data configuration of the Asan osteoporosis cohort.

Sources of CXR images	N	Collection period
Patients from the Health promotion center	460	2008.04.10 ~ 2011.12.21
Outpatients	251	2006.07.06 ~ 2019.06.25
Inpatients	19	2008.09.04 ~ 2018.03.16
Patients from the emergency department	19	2009.01.13 ~ 2019.06.20
Total	749	

The total images were chest radiographs of 11811 female and 2304 male individuals, with ages ranging from 40 to 90 years (mean age, 57.04 years). Of the radiographs, 4729 showed normal results, 4726 indicated osteopenia, and 4660 indicated osteoporosis. Table 4.2. presents the baseline characteristics, ages, heights, and weights in the training, tuning, internal, and external validation datasets. Figure 4.2 shows the male and female age distribution by BMD class in 4 datasets.

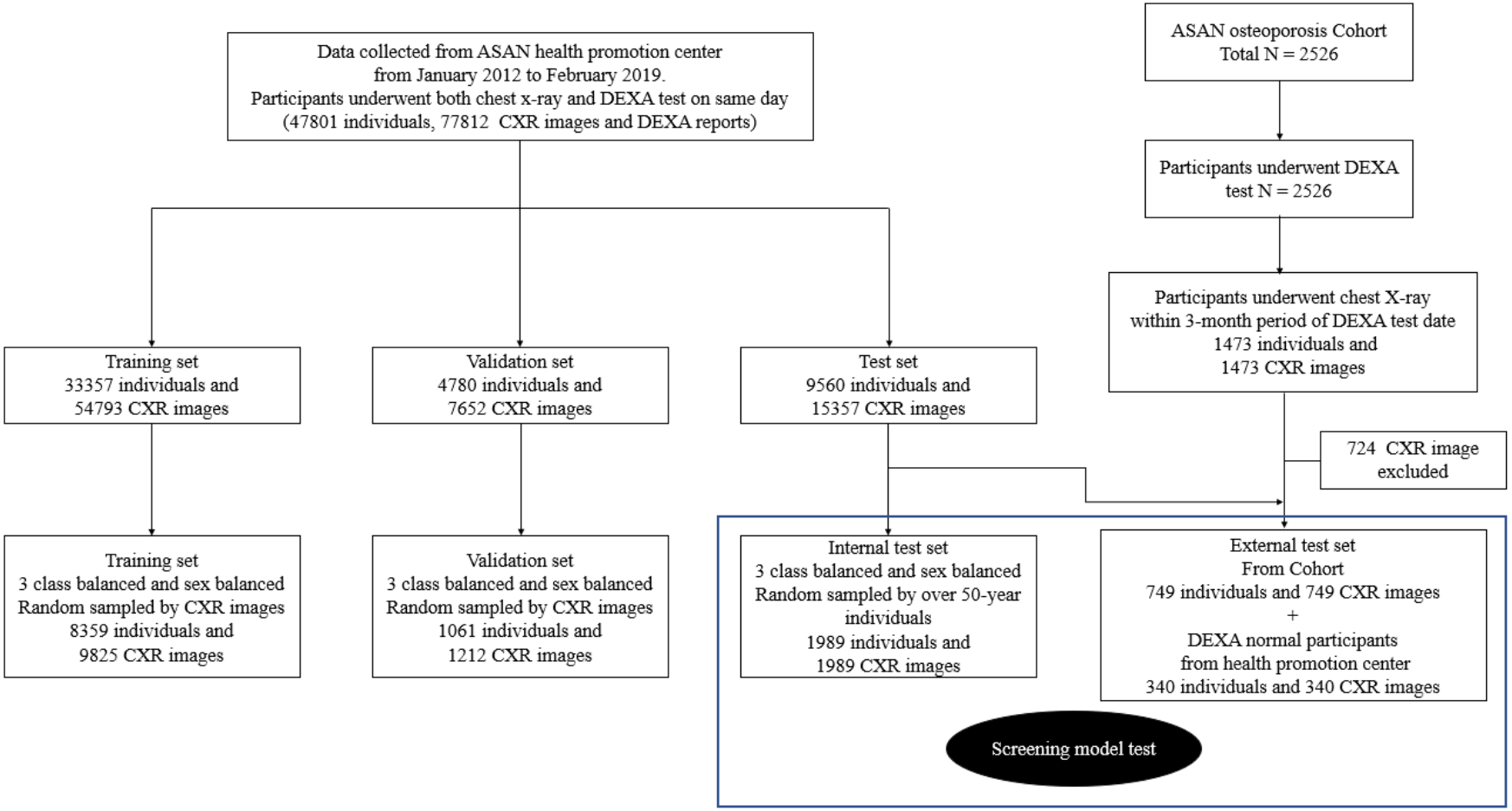


Figure 4.1. Dataset configuration of osteoporosis screening model.

Table 4.2. Demographic characteristics of the datasets.

Characteristics	Training dataset		Tuning dataset		Internal validation dataset		External validation dataset	
Sex	female	male	female	male	female	male	female	male
Participants	8148	1677	1026	186	1668	321	969	120
Age, yr., mean (SD)	56.14 ± 7.92	58.15 ± 7.86	56.08 ± 7.96	57.95 ± 7.22	58.78 ± 6.78	58.29 ± 6.66	58.59 ±6.57	59.42 ± 6.73
Height, cm, mean (SD)	157.86 ± 7.23	169.32 ± 5.59	157.77 ± 8.82	168.55 ± 5.77	157.45 ± 5.37	168.52 ± 5.83	156.56 ± 5.31	168.38 ± 5.62
Weight, kg, mean (SD)	55.68 ± 7.76	68.56 ± 9.79	56.00 ± 8.11	67.87 ± 11.09	56.58 ± 8.03	67.62 ± 9.34	56.98 ± 7.08	67.80 ± 8.32
T-Score, mean								
L1-L4	-1.23 ± 1.53	-1.07 ± 1.81	-1.16 ± 1.58	-1.08 ± 1.73	-1.33 ± 1.52	-1.10 ± 1.76	-1.41 ± 1.50	-0.98 ± 2.11
Total femur	-0.9. ± 1.03	-0.49 ± 1.00	-0.88 ± 1.05	-0.56 ± 0.97	-0.94 ± 1.52	-0.53 ± 1.00	-0.83 ± 0.99	-0.11 ± 1.03
BMD, mean (SD)								
L1-L4	1.00 ± 1.01	1.04 ± 0.40	1.00 ± 0.20	1.55 ± 6.54	0.98 ± 0.31	1.04 ± 0.23	0.98 ± 0.18	1.06 ± 0.26
Total femur	0.85 ± 1.42	0.87 ± 0.16	0.83 ± 0.14	0.86 ± 0.17	0.82 ± 2.04	0.88 ± 0.16	0.86 ± 0.12	0.93 ± 0.14
BMD categories, n (%)								
Normal	2716 (33.33)	559 (33.33)	342 (33.33)	62 (33.33)	556 (33.33)	107(33.33)	342 (35.29)	45 (37.50)
Osteopenia	2716 (33.33)	559 (33.33)	342 (33.33)	62 (33.33)	556 (33.33)	107 (33.33)	336 (34.67)	48 (40.00)
Osteoporosis	2716 (33.33)	559 (33.33)	342 (33.33)	62 (33.33)	556 (33.33)	107 (33.33)	291 (30.03)	27 (22.50)

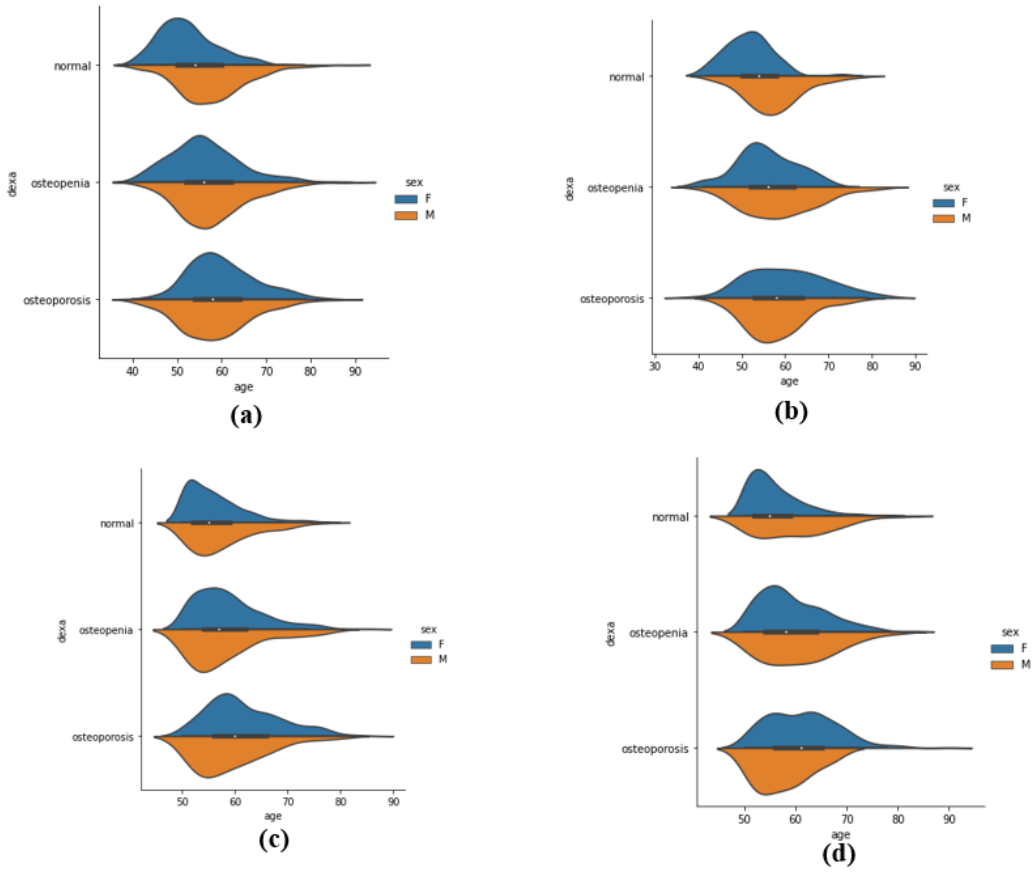


Figure 4.2. Male and female age distributions by BMD class; (a) training dataset, (b) tuning dataset, (c) internal validation dataset, (d) external validation dataset.

4.2. preprocessing

As the chest radiographs had various images matrix sizes of approximately 2000 × 2000 pixels, all the images were resized to 512 × 512 pixels with bicubic interpolation and were normalized with by z-scores. A set of pixel values of original and scaled images is represented by X, Z respectively; the formula of z-scores normalization is

$$Z = \frac{X - \text{mean}\,(X)}{Standard\ deviation(X)} \qquad (11)$$

In general, the deep learning-based training required careful preprocessing, because of the absence of normalized physical meaning of the pixel values of the chest radiographs. The signal to noise, edge patterns, and textures on chest radiographs may depend on the imaging protocol, vendor, and physical characteristics of patients in multi-centers. Therefore, sharpening and blurring processes were randomly applied to the images during the training in terms of data augmentation, making the model to be robust to variations from various imaging protocols, multi-vendors, and physical characteristics of patients. In addition, we augmented the data by using additional augmentation techniques, such as rotation (±5°), shifting (±3%), and zoom (<15%) for more robust training.

4.3. Statistical analyses

To comprehensively evaluate the screening performance for the two validation datasets, the accuracy (ACC), sensitivity (recall, SEN), specificity (SPE), positive predictive value (precision, PPV), negative predictive value (NPV), and area under the ROC curve (AUC) were calculated. The two validation datasets were imbalanced data with osteoporosis and non-osteoporosis. We also calculated F1-score. The confusion matrix in this study was given as a 2 × 2 contingency table displaying numbers of true positives, true negatives, false positives, and false negatives.

$$Precision = PPV = \frac{True\ Positive}{True\ Postive+False\ Positive} \quad (12)$$

$$NPV = \frac{True\ Negative}{True\ Nagative+False\ Negative} \quad (13)$$

$$F1 - score = 2 * \frac{Precsion*Recall}{Precsion+Recall} \quad (14)$$

4.4. Experiment setting

This study used the InceptionV3 for binary classification as an osteoporosis screening tool. The InceptionV3 was known as computationally efficient model [103] which showed good performance in the 2014 ILSVRC. [46] Considering that our task was to develop a binary classifier, we modified the InceptionV3 by replacing the last layer with a global average pooling layer, three dense layers and two dropout layers as shown Figure 4.3. Ubuntu 18.04 with a V100 GPU, CUDA 10.0 (NVidia Corporation), TensorFlow 1.15.0, and Keras 2.3.0 were used as the experimental environments.

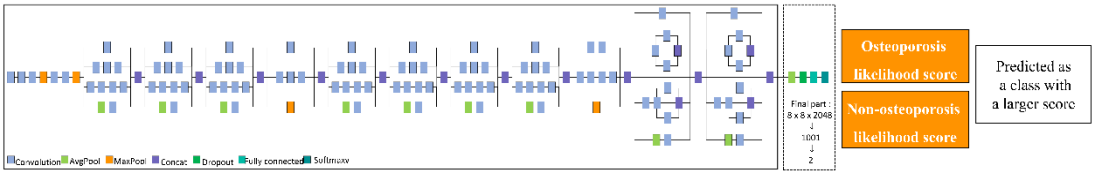


Figure 4.3. The CNN model architecture.

Sub-studies were conducted for various regions of interest (ROI) and images sizes in sub-group including only the normal and osteoporosis datasets. As the osteoporosis, osteopenia, and normal groups were categorized from continuous BMD values, the osteopenia group was difficult to distinguish from the other groups. First, it was necessary to check whether CNN trained using chest radiographs distinguishes between those with normal BMD and those with osteoporosis. Second, the image size suitable for training and ROI on chest radiographs were confirmed. Third, various attempts in total dataset were made to develop a model that best screening osteoporosis patients from the entire dataset using transfer learning, input demographic variates and so on. Finally, experiments were conducted on the effect of sex and age information in a dataset with the same sex and age distribution for each BMD class. Sex imbalance in medical imaging datasets may produce biased classifiers based on CNNs, with

lower performance in underrepresented groups. The scheme of this study is presented in Figure 4.4.

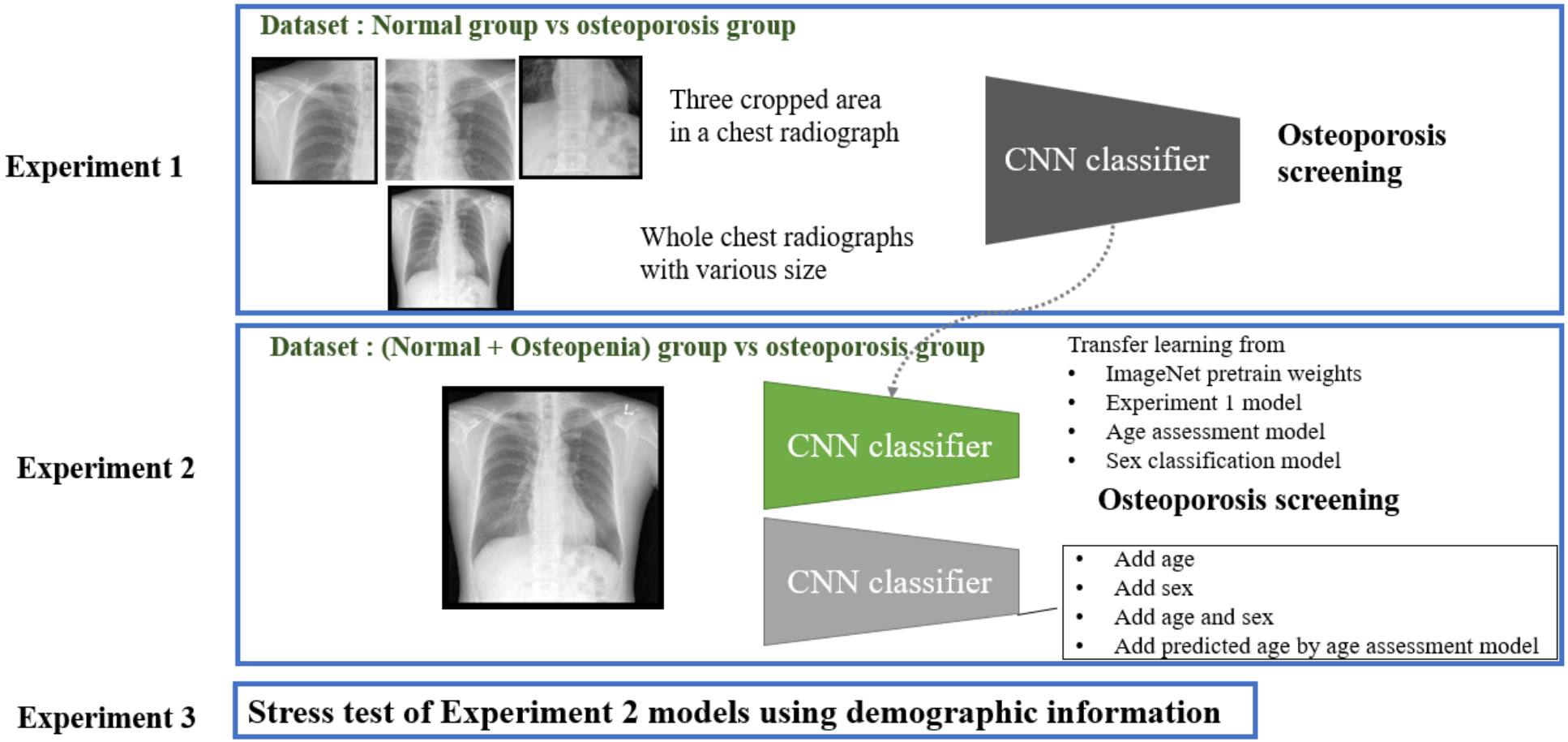


Figure 4.4. The scheme of this study.

4.5. Experiment 1

4.5.1. Studies of various regions on chest radiographs

Various ROIs at the right shoulder area, cervical and thoracic spine, and thoracic and lumbar spine areas on chest radiographs with 1024 × 1024 pixels in size were cropped and trained separately. The classification performances of the three CNN models of the various input ROIs tested in only the normal and osteoporosis datasets are displayed in Table 4.3. The AUCs of the models trained with the right shoulder, cervical and thoracic, and thoracic and lumbar areas showed equivalent performance. Confusion matrixes of the three CNN models of the various input ROIs are shown in Figure 4.5.

Table 4.3. Performances of the models trained with various input ROI.

Input image area	AUC	ACC (%)	SEN (%)	SPE (%)
Right shoulder area	0.98	92.72	94.01	91.44
Cervical and thoracic area	0.98	92.01	92.58	91.44
Thoracic and lumbar area	0.97	91.37	91.16	91.58

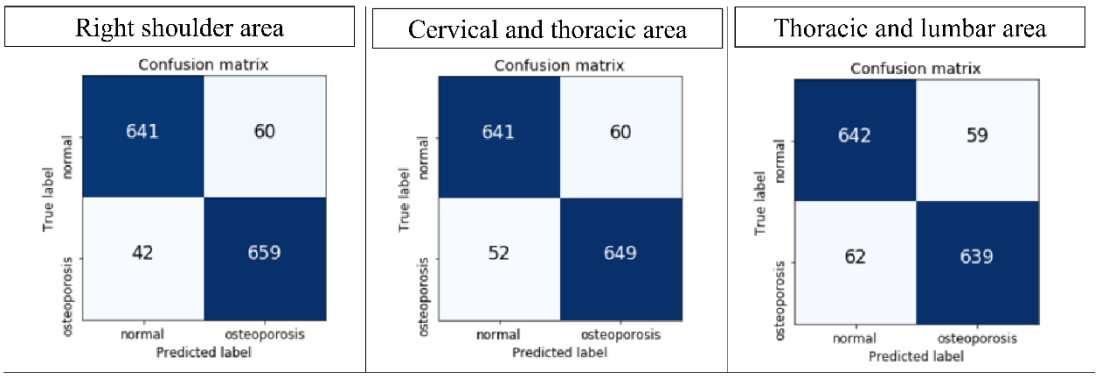


Figure 4.5. Confusion matrixes of the models trained with various inputs ROIs.

4.5.2. Studies of image sizes

The performances of the five CNN models of various image sizes on whole-chest radiographs are shown in the Table 4.4. and confusion matrixes of them are presented in the Figure 4.6. The model trained with 512 × 512 pixel-sized chest radiographs showed the best performance regardless of batch size.

Table 4.4. Performances of the models trained with various image sizes on whole-chest images.

Image size	Batch size	AUC	ACC (%)	SEN (%)	SPE (%)
128	100	0.92	84.24	84.02	84.45
256	100	0.97	90.37	92.15	88.59
512	60	0.99	94.01	95.15	92.87
512	15	0.99	94.44	95.86	93.01
1024	15	0.98	94.01	96.86	91.16

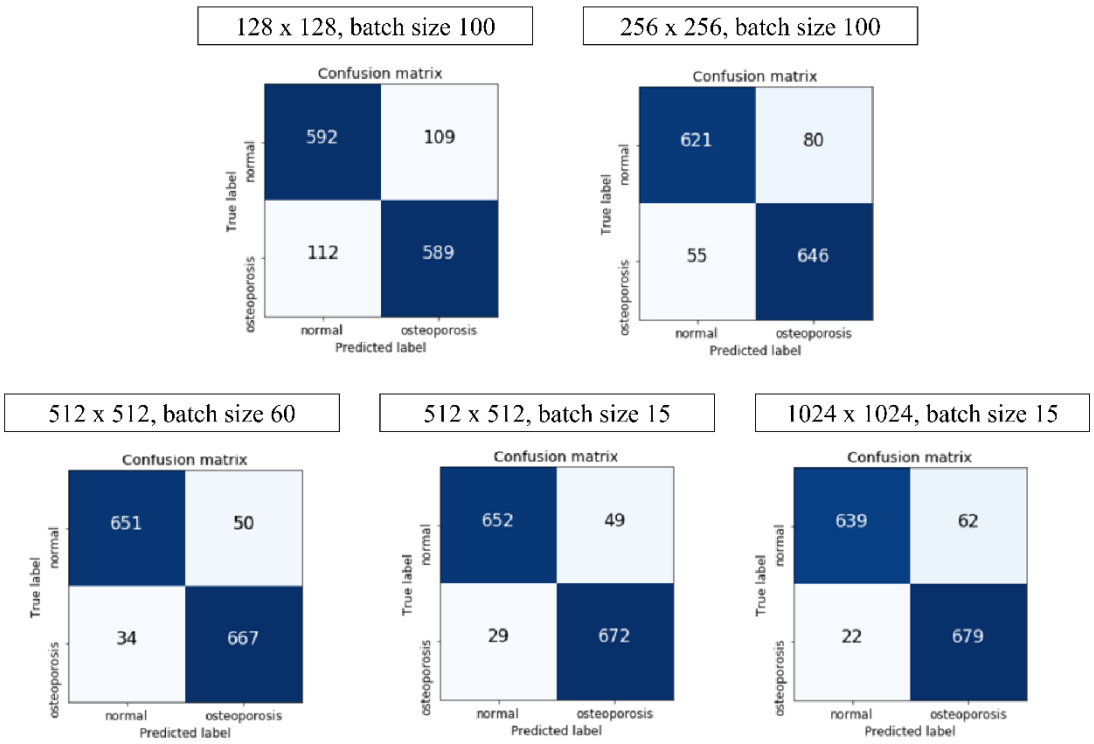


Figure 4.6. Confusion matrixes of the models trained with various inputs image sizes.

4.6. Experiment 2

4.6.1. Development osteoporosis screening model on chest radiographs

In previous sub-studies, the result of the CNN trained 512 × 512 sized whole-chest images was the best. According to this, development of osteoporosis screening model was decided to train with 512 × 512 sized whole-chest images in entire dataset. BMD is graded into three stages, normal, osteopenia, and osteoporosis, but in the current Korean medical system, medical insurance is applied only to osteoporosis patients. Therefore, in clinical situations, a model that selects osteoporosis patients well is the most useful. A binary classifier was developed to classify osteoporosis by labeling normal and osteopenia as non-osteoporosis.

4.6.2. Development of baseline model

The prevalence of osteoporosis in Korea is 7.3% in males and 38.0% in females 50 years and older. [121] Our dataset showed sex difference of osteoporosis prevalence obviously. There are 8148 and 1677 of training dataset, 1026 and 186 of tuning dataset, 1668 and 321 of internal, and 969 and 120 external validation datasets in female and male, respectively.

Table 4.5 demonstrates the screening performances of female model in dataset consisted of female and Table 4.6 shows the performances of male model in dataset consisted of male. In the internal validation set, the performance of female and male models achieved an AUC of 0.91 and 0.82 with a sensitivity of 76.98% and 75.39, and F1-score of 75.29 and 64.38, respectively. In the external validation set, the female and male models yielded an AUC of 0.86 and 0.82 with a sensitivity of 74.23% and 85.19%, and F1-score of 64.38 and 53.49, respectively. The confusion matrixes of female and male models in the internal and external validation datasets are shown in Figure 4.7.

Table 4.5. Performance of female model in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.91	83.15	76.98	86.24	73.67	88.22	75.29
external	0.86	75.34	74.23	75.81	56.84	87.27	64.38

Table 4.6. Performance of male model in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.82	75.39	68.24	78.97	61.86	83.25	64.89
external	0.82	66.67	85.19	61.29	39.98	93.44	53.49

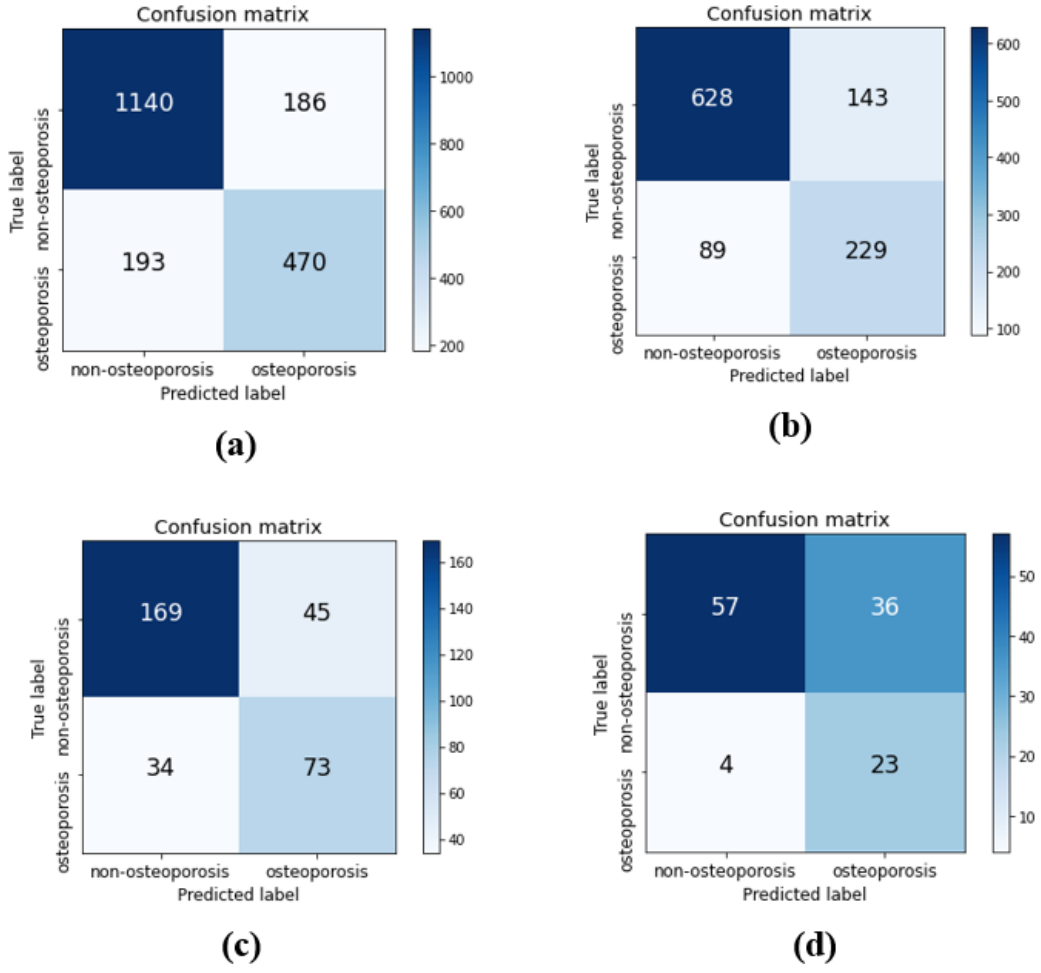


Figure 4.7. The confusion matrixes of sex separate training models: (a) female model in the internal validation dataset, (b) female model in the external validation dataset, (c) male model in the internal validation set, (d) male model in the external validation set.

Table 4.7 demonstrates the screening performances of the baseline model in screening osteoporosis based on chest radiographs. In the internal validation set, the osteoporosis screening performance of the baseline models achieved an AUC of 0.91 with a sensitivity of 76.77% and F1-score of 74.85. In the external validation set, the baseline model screening osteoporosis yielded an AUC of 0.87 with a sensitivity of 78.93% and F1-score of 68.77. The confusion matrixes of the baseline model in the internal and external validation datasets are shown in Figure 4.8.

Table 4.7. Performance of the baseline model in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.91	82.81	76.77	85.82	73.03	88.08	74.85
external	0.87	79.06	78.93	79.12	60.92	90.10	68.77

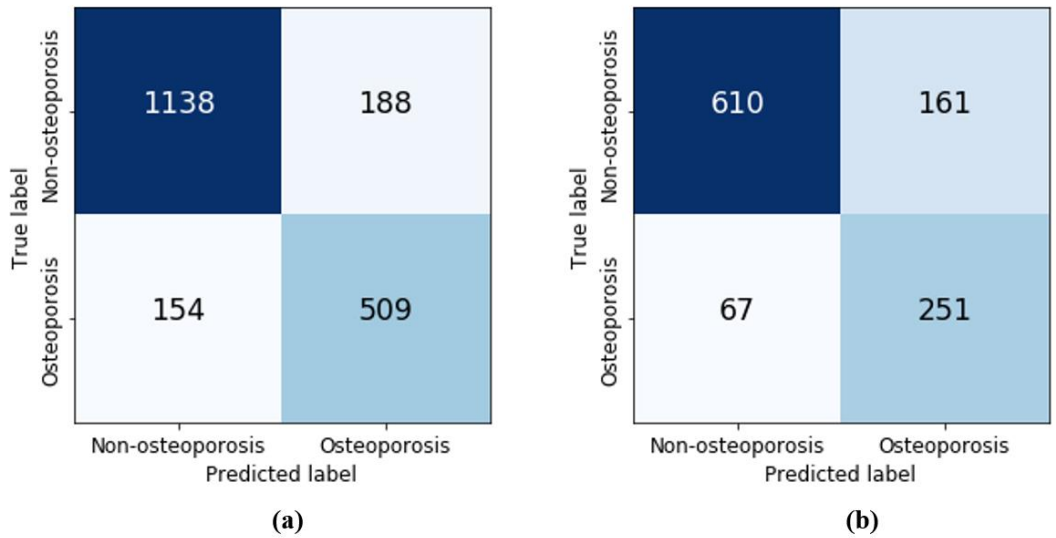


Figure 4.8. The confusion matrixes of the baseline model; (a) the internal validation dataset, (b) the external validation dataset**.**

4.6.3. Development of models using transfer learning methods

CNN layers settings to be trained during transfer learning were set differently according to the content of pre-training. When the age assessment, sex classification and ImageNet models were a pre-trained model, the pre-trained content was not directly related to the osteoporosis screening model, so it was used only for the initial weights setting and all layers were re-trained. However, in case of transfer learning using subgroup classification model, only the layer of the last two blocks was tuned to be trained because only osteopenia group needs to be further trained and classified.

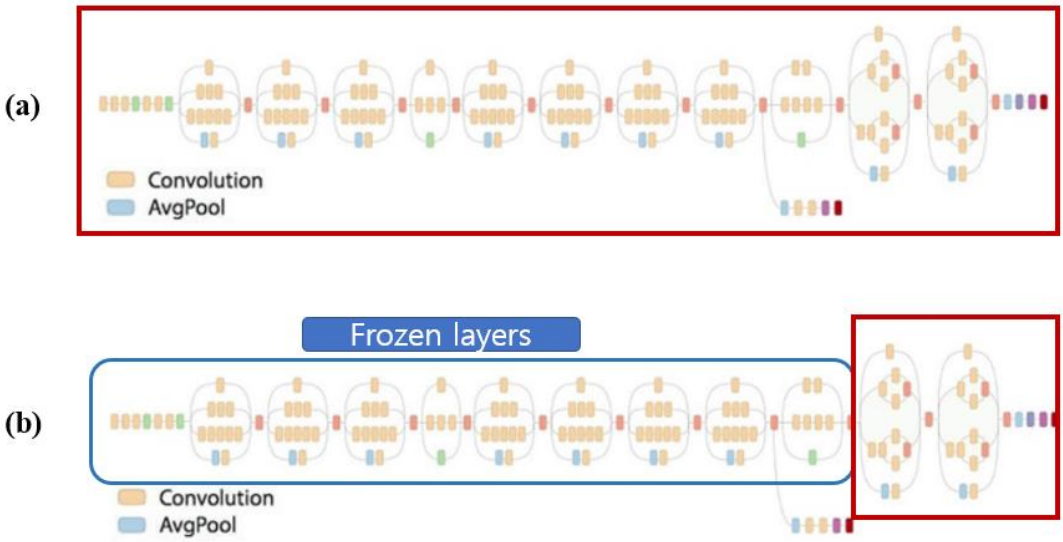


Figure 4.9. Trainable layer setting of transfer learning; (a) set all layers trainable from pretrain weights of ImageNet classification model, age assessment model, and sex classification model, (b) set last block trainable from pretrain weights of subgroup classification model.

4.6.3.1. Development of models using transfer learning methods of ImageNet pretrain weights and sub-group study model

ImageNet exist prevalent transfer learning mainly uses the representation of pre-trained models using the natural image dataset on RGB color space. [43] Chest radiographs are a single-channel image on grayscale space, and we converted 1-channel image to 3-channel image by stacking same image. Table 4.8 demonstrates the performances of the ImageNet transfer model. The osteoporosis screening performance of the ImageNet transfer model achieved an AUC of 0.91 and 0.80 in the internal and external validation dataset, respectively.

Table 4.8. Performance of the ImageNet transfer model in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.92	84.77	81.45	86.43	75.00	90.31	78.09
external	0.80	76.31	51.57	86.51	61.19	81.24	55.97

The pre-trained model was the best classifier of whole-chest images with 512 × 512 pixels in size in the sub-group studies. Pre-trained weights were developed by training the binary classifier for differentiating normal findings and osteoporosis, except osteopenia. Table 4.9 demonstrates the screening performances of the transfer model from sub-group classification model. In the internal validation set, the osteoporosis screening performance of the transfer models achieved an AUC of 0.91 with a sensitivity of 84.31%. In the external validation set, the transfer model screening osteoporosis yielded an AUC of 0.88 with a sensitivity of 86.16%. The confusion matrixes of the transfer model in the internal and external validation datasets are shown in Figure 4.10. The ROC curves in male and female of internal and external validation dataset are shown in Figure 4.11. This model achieved an AUC of 0.92 and 0.88 in female and male of internal test set. It also yielded an AUC of 0.88 and 0.83 in female and male of external test set.

Table 4.9. Performance of the transfer model from sub-group classification in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.91	82.40	84.31	81.45	69.44	91.22	76.16
external	0.88	77.69	86.16	74.19	57.93	92.86	69.28

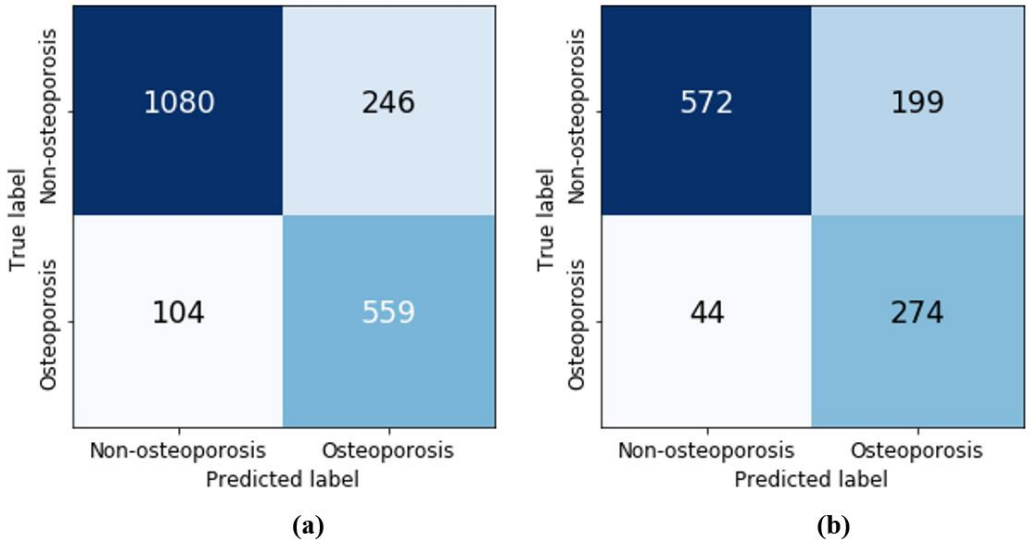
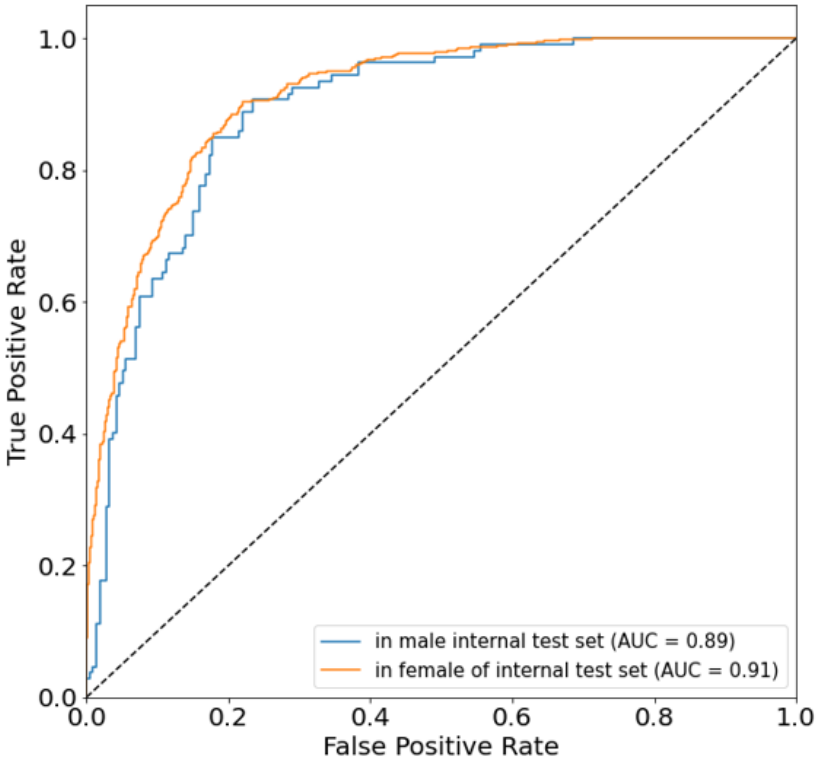
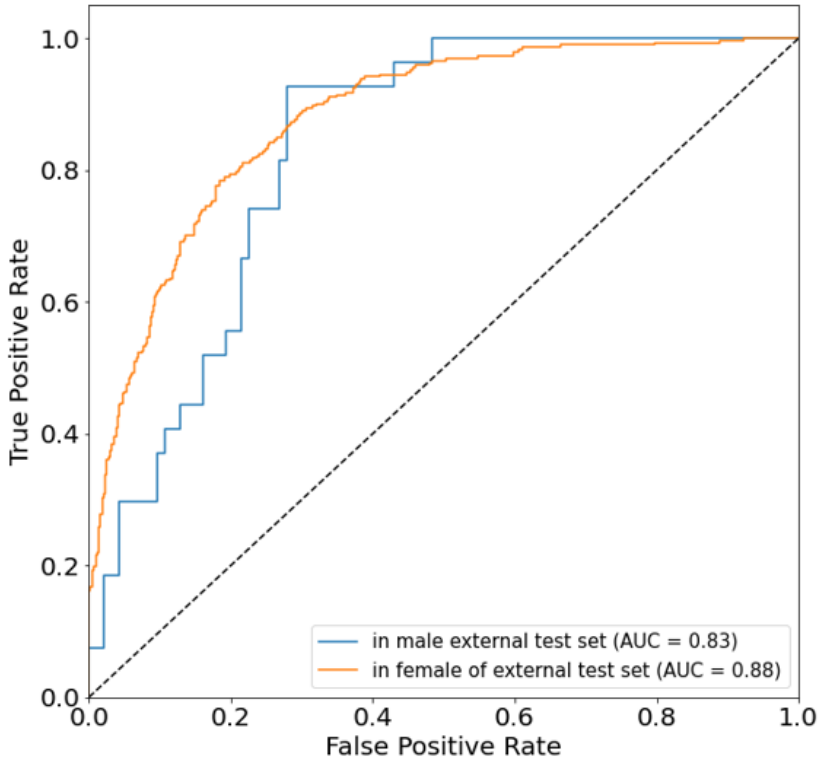


Figure 4.10. The confusion matrixes of the transfer model from sub-group classification; (a) in the internal validation dataset, (b) in the external validation dataset.


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
in male internal test set (AUC = 0.89)
in female of internal test set (AUC = 0.91)

(a)


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
in male external test set (AUC = 0.83)
in female of external test set (AUC = 0.88)

(b)

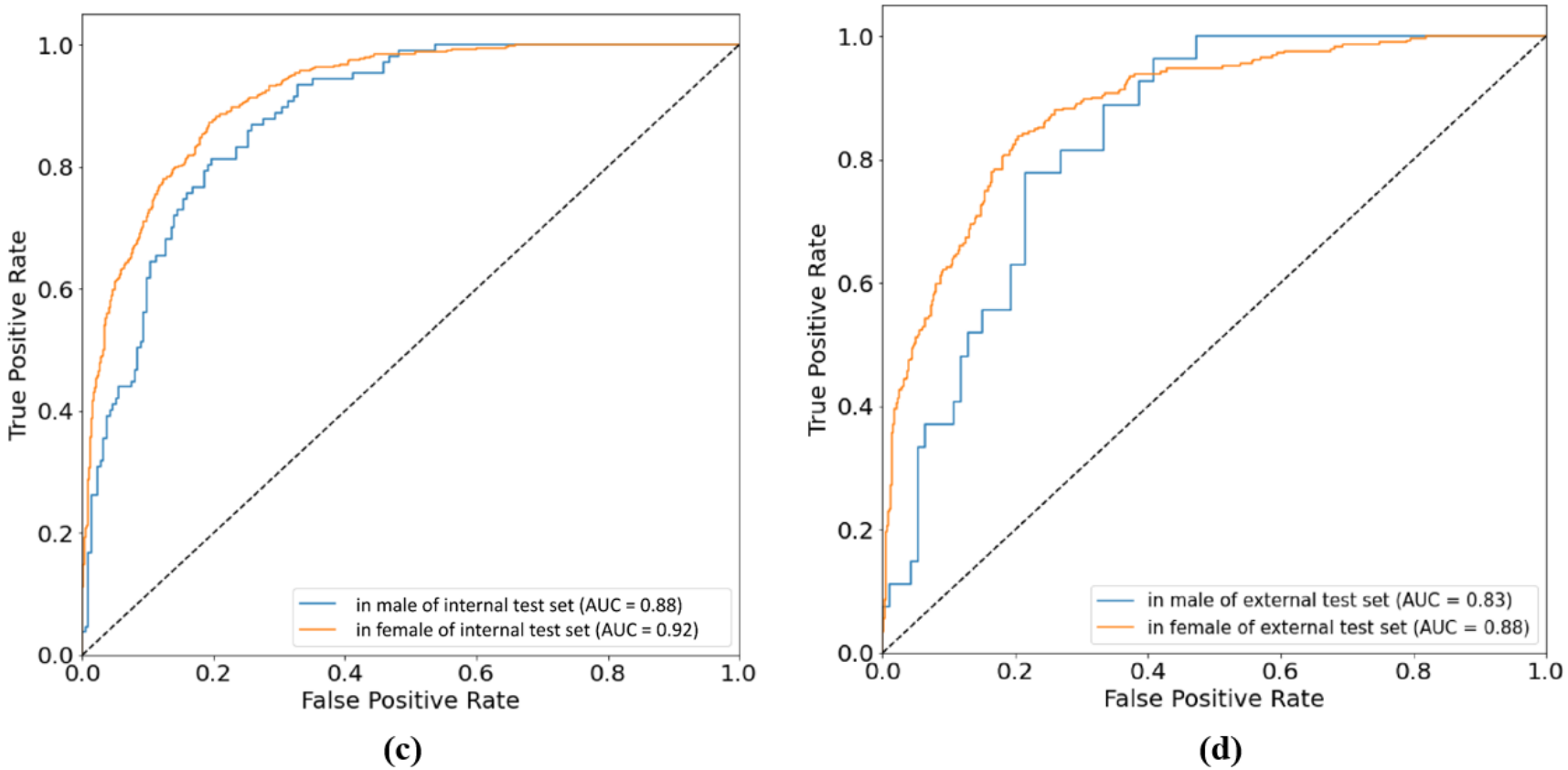


Figure 4.11. ROC curves by gender; (a) the baseline model in the internal validation dataset, (b) the baseline model in external validation dataset. (c) the transfer model from sub-group classification in the internal validation dataset, (d) the transfer model from sub-group classification in the internal validation dataset

4.6.3.2. Development of models using transfer learning methods of age and sex prediction model

These transferred models were trained with pre-trained weight from the best performance of age prediction and sex classification InceptionV3 models. In these cases, by using the pre-trained model, transfer learning was performed by training all layers of the InceptionV3 model for developing an osteoporosis screening model. These models were trained with training and tuning dataset and the results shows Table 4.10. and Table 4.11.

Table 4.10. Performance of the transfer model from age assessment model weights in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.91	81.90	71.07	83.33	71.07	90.20	76.10
external	0.87	78.33	84.59	75.75	58.99	92.26	69.51

Table 4.11. Performance of the transfer model from sex classification model weights in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.91	82.96	79.64	84.61	72.13	89.26	75.70
external	0.86	76.77	80.82	75.10	57.24	90.47	67.01

In the age assessment model, people over 20 years of age were included, and the osteoporosis dataset consisted of people over 40 years old, so there was a difference in age composition. Therefore, transfer learning of the age assessment model over 40 years of age was performed, and the results shows Table 4.12.

Table 4.12. Performance of the transfer model from over 40 age assessment model weights in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.89	80.95	70.89	85.97	71.65	85.52	71.27
external	0.86	78.70	72.01	81.45	61.56	87.59	66.38

4.6.4. Development of models using chest radiographs and demographic information

Extracted from the convolutional layers of the InceptionV3 of the chest radiograph, the one-dimensional reshaped result and 1x1 or 1x2 or 1x3 dimensional data created from age or sex or both or predicted age from CNN, the age assessment model, were combined. Table 4.13. and Table 4.14. presents the performance for chest radiographs combined with demographic information analysis. The model trained with addition of age or sex information showed some F1-score performance improvement in internal validation dataset however not in the external validation dataset compared with the baseline model.

Table 4.13. Performance of the CNN models trained with demographic information in the internal validation datasets.

	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
Add age	0.92	85.16	76.70	89.40	78.34	88.47	77.51
Add sex	0.92	85.20	77.53	89.03	77.94	88.80	77.73
Both	0.91	85.20	79.00	83.80	70.92	88.86	74.74
Pred-age	0.90	82.20	76.92	84.84	71.73	88.03	74.24

Table 4.14. Performance of the CNN models trained with demographic information in the external validation datasets.

	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
Add age	0.87	78.51	73.27	80.67	60.99	87.98	66.57
Add sex	0.87	79.06	76.10	80.29	61.42	89.06	67.98
Both	0.86	78.33	69.50	81.97	61.39	86.69	65.19
Pred-age	0.87	78.60	78.62	78.60	60.24	89.91	68.21

4.7. Experiment 3

4.7.1. Dataset of Stress tests and baseline model

We conducted stress study and ablation study on stress training to apprehend the effects of adding demographic information more clearly. Training and tuning datasets were constructed by matching the number and age distribution of male and female by BMD class shown as Figure 4.12. The numbers of training and tuning datasets were 1677 and 185, respectively. In a few published studies [88, 122-124] of deep learning models for screening osteoporosis, most of the training data were about 1000 cases.

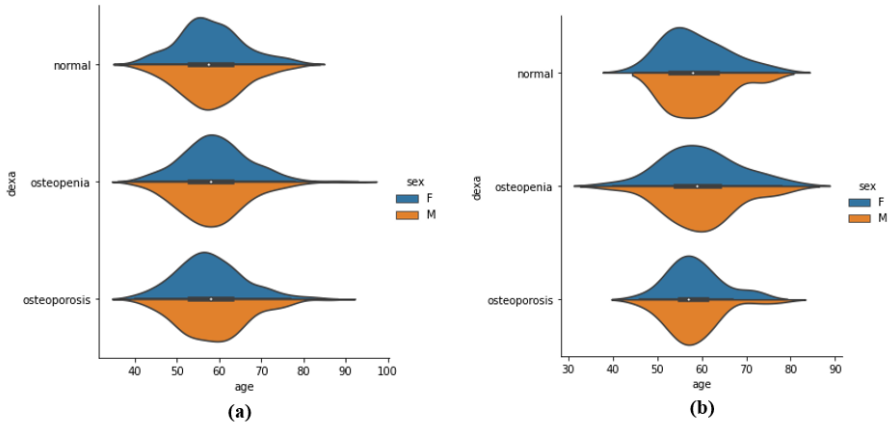


Figure 4.12. Male and female age distributions by BMD class; (a) training dataset of stress test, (b) tuning dataset of stress test.

Baseline model was developed with this dataset by trained using only images and Table 4.15 show the results.

Table 4.15 Performance of the baseline model trained stress test dataset.

	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
Internal	0.85	77.48	51.3	90.6	73.12	78.81	60.28
external	0.80	76.31	51.6	86.5	61.19	81.24	55.97

4.7.2. Transfer learning models from age and sex prediction models

Table 4.16. and Table 4.17. demonstrates the screening performances transfer model from age prediction model and sex prediction model in the internal and external validation datasets. In case of transfer model from age prediction model, F1-score performance was better than model trained only image and comparable to model trained with image and age in stress test. However, performance of transfer model from sex prediction model was lower than model trained only image in stress test.

Table 4.16. Performance of the transfer model from age prediction weight of stress test in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.87	74.26	87.18	67.80	57.51	91.36	69.30
external	0.83	70.71	85.85	64.46	49.91	91.70	63.12

Table 4.17. Performance of the transfer model from sex prediction weight of stress test in the internal and external validation datasets.

Datasets	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
internal	0.82	76.07	54.30	86.95	67.54	79.19	60.20
external	0.77	74.38	48.43	85.08	57.25	80.00	52.47

4.7.3. Development of models trained with demographic information.

Three CNN models were developed with this dataset such as trained using image and age, image and sex, image and both of age and sex. Each model was test in internal and external validation dataset. The performances of four models in internal and external datasets are presented in Table 4.18. and Table 4.19.

Table 4.18. Performance of the CNN models trained with demographic information of stress test in the internal validation datasets.

	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
Add age	0.86	79.24	63.2	87.3	71.26	82.58	66.99
Add sex	0.87	81.03	79.2	78.3	69.72	88.27	74.15
Both	0.86	79.13	66.4	85.5	69.61	83.57	67.95

Table 4.19. Performance of the CNN models trained with demographic information of stress test in the external validation datasets.

	AUC	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	F1-score (%)
Add age	0.81	75.30	58.49	82.23	57.59	82.77	58.03
Add sex	0.84	74.01	78.90	72.00	53.75	89.23	63.95
Both	0.83	76.58	65.4	81.2	58.92	85.05	62.00

The confusion matrixes of stress test models in internal and external validation datasets are showed in Figure 4.13 and Figure 4.14. ROC curves of stress test models in internal and external validation datasets are presented in Figure 4.15.

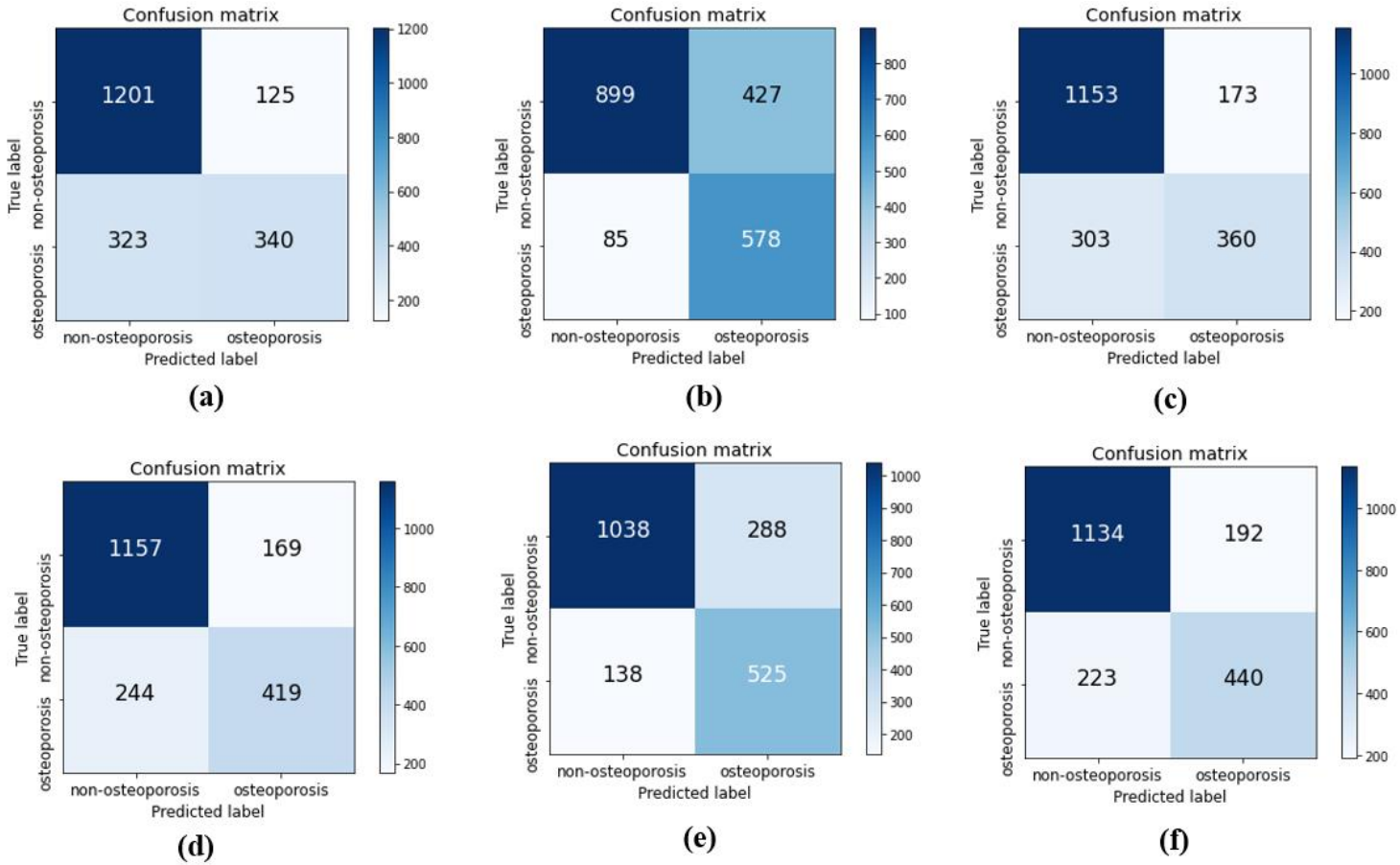


Figure 4.13. The confusion matrixes of CNN models of the stress test in the internal validation datasets; (a) model trained only images (baseline model), (b) transfer model from age assessment model, (c) transfer model from sex classification model, (d) model trained with age information, (e) model trained with sex information, (f) model trained with age and sex information.

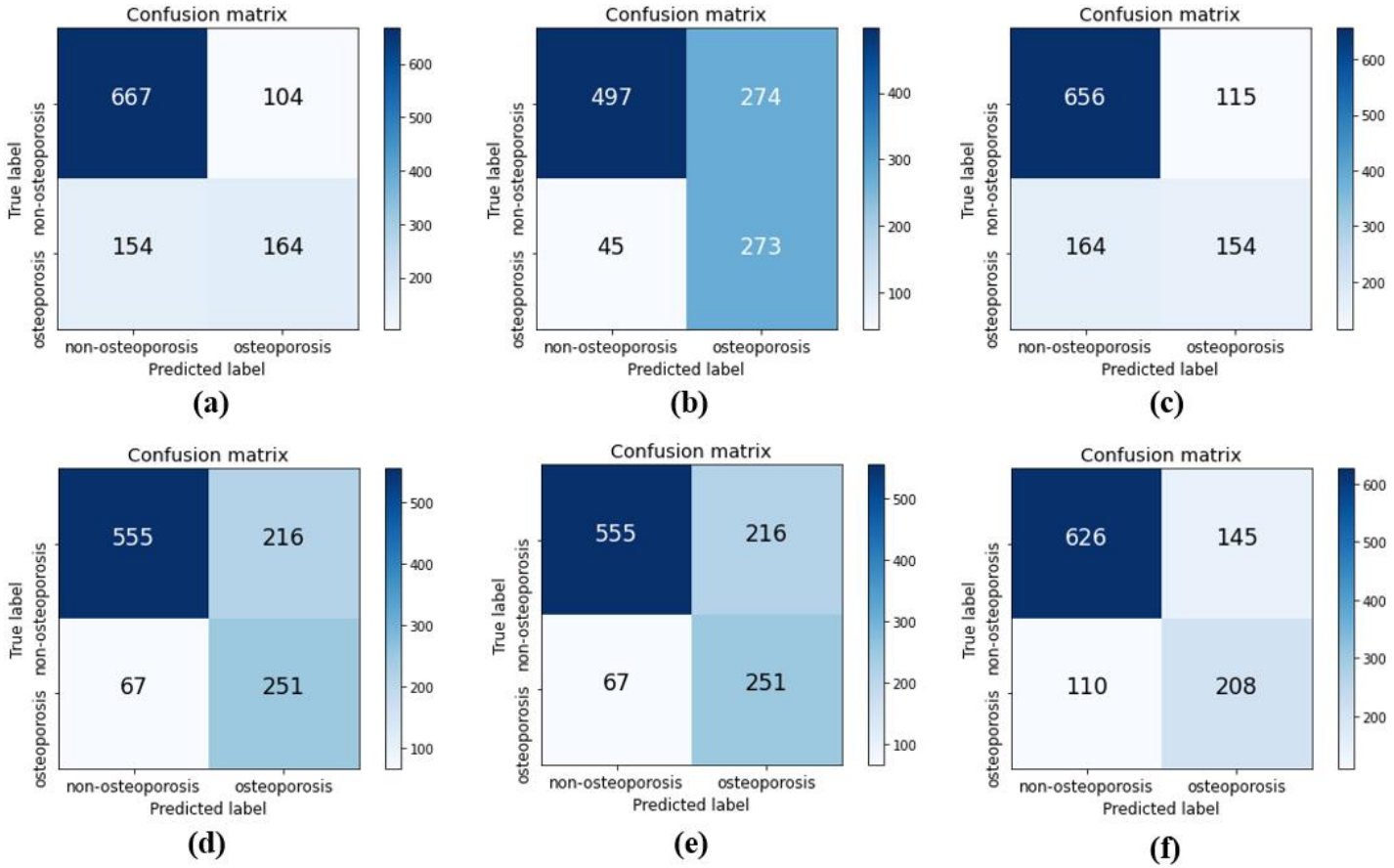


Figure 4.14. The confusion matrixes of CNN models of the stress test in the external validation datasets; (a) model trained only images (baseline model), (b) transfer model from age assessment model, (c) transfer model from sex classification model, (d) model trained with age information, (e) model trained with sex information, (f) model trained with age and sex information.

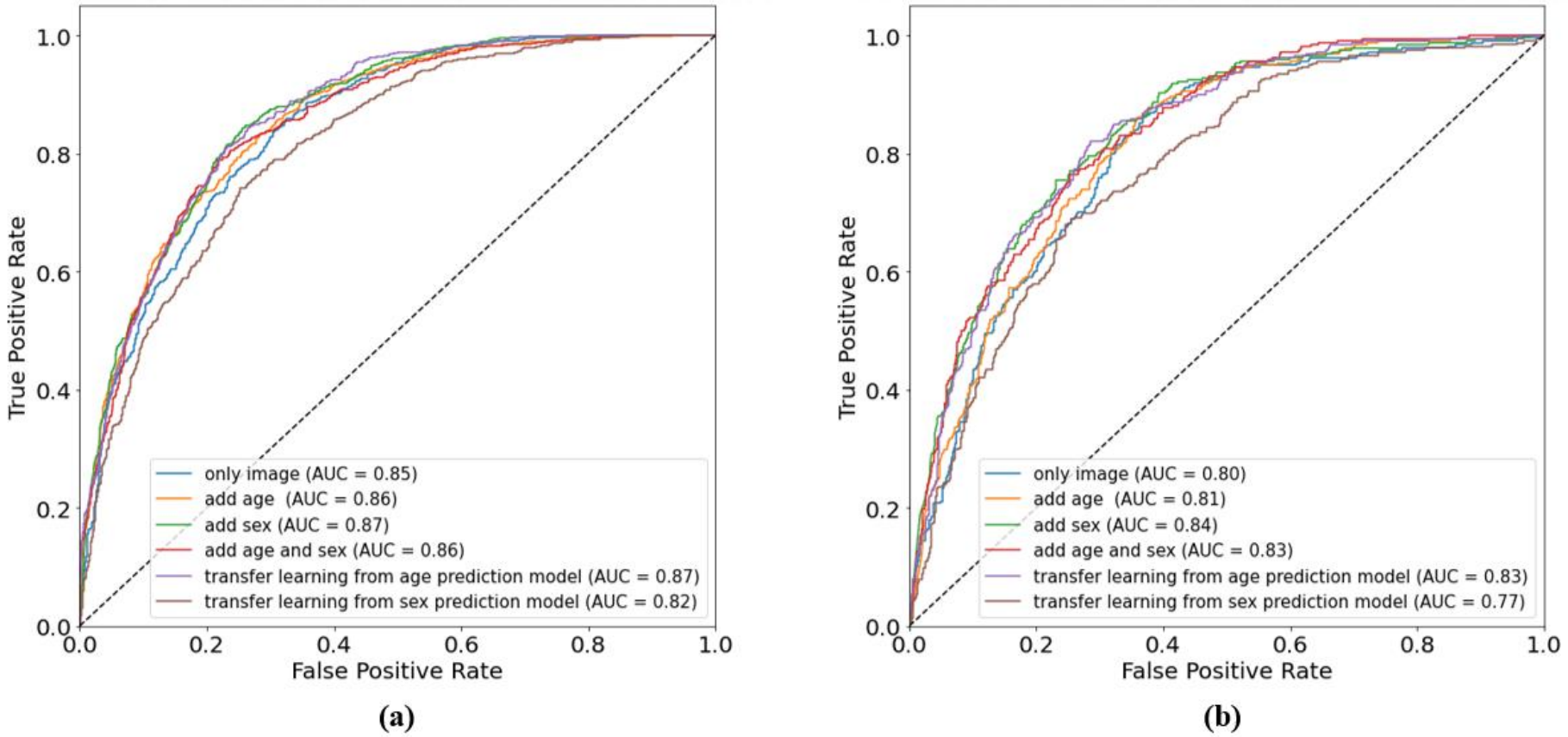


Figure 4.15. ROC curves of stress test; (a) in the internal validation dataset, (b) in the external validation dataset.

4.8. Interpretation of osteoporosis screening model

Deep learning models have often been referred to as non-interpretable black boxes because their prediction processes are difficult to determine. To determine the decision-making process of the model and the most important regions for the model to screen osteoporosis on chest radiographs, we used the gradient-weighted class activation mapping technique (Grad-CAM) [125] by overlaying the most significant regions for screening osteoporosis in the images with red color, in this study.

Because each chest radiograph has different spatial properties of anatomical structures, we registered all images based on the shape of lung. Therefore, lung segmentations for each image were done using deep learning-based lung segmentation developed in our institution. [126] Then, rigid registration parameters based on segmented lungs from target chest radiograph and reference chest radiograph was derived where reference chest radiograph was obtained from generation model. The derived registration parameters were then applied to corresponding Grad-CAMs. All Grad-CAMs for target images were registered in this manner. Finally, pixel-wise addition of registered Grad-CAMs were performed and converted to 8-bit scale image.

Transfer learning model from sub-group classification pre-trained weights showed best performance. InceptionV3 has 9 convolution layers in the last block. Grad-CAMs extracted from this model's 9 convolution layers of the last block. For a total 559 true positive case in internal validation dataset, the Grad-CAMs were obtained for each convolution layer. As shown in Figure 4.16, the average Grad-CAM for each layer was obtained by combining the images for each layer. The average Grad-CAMs in external validation were similar to Figure 4.16. 94-layer is the last convolutional layer, and the average Grad-CAM of 94-layer seems to show the previous Grad-CAMs combined.

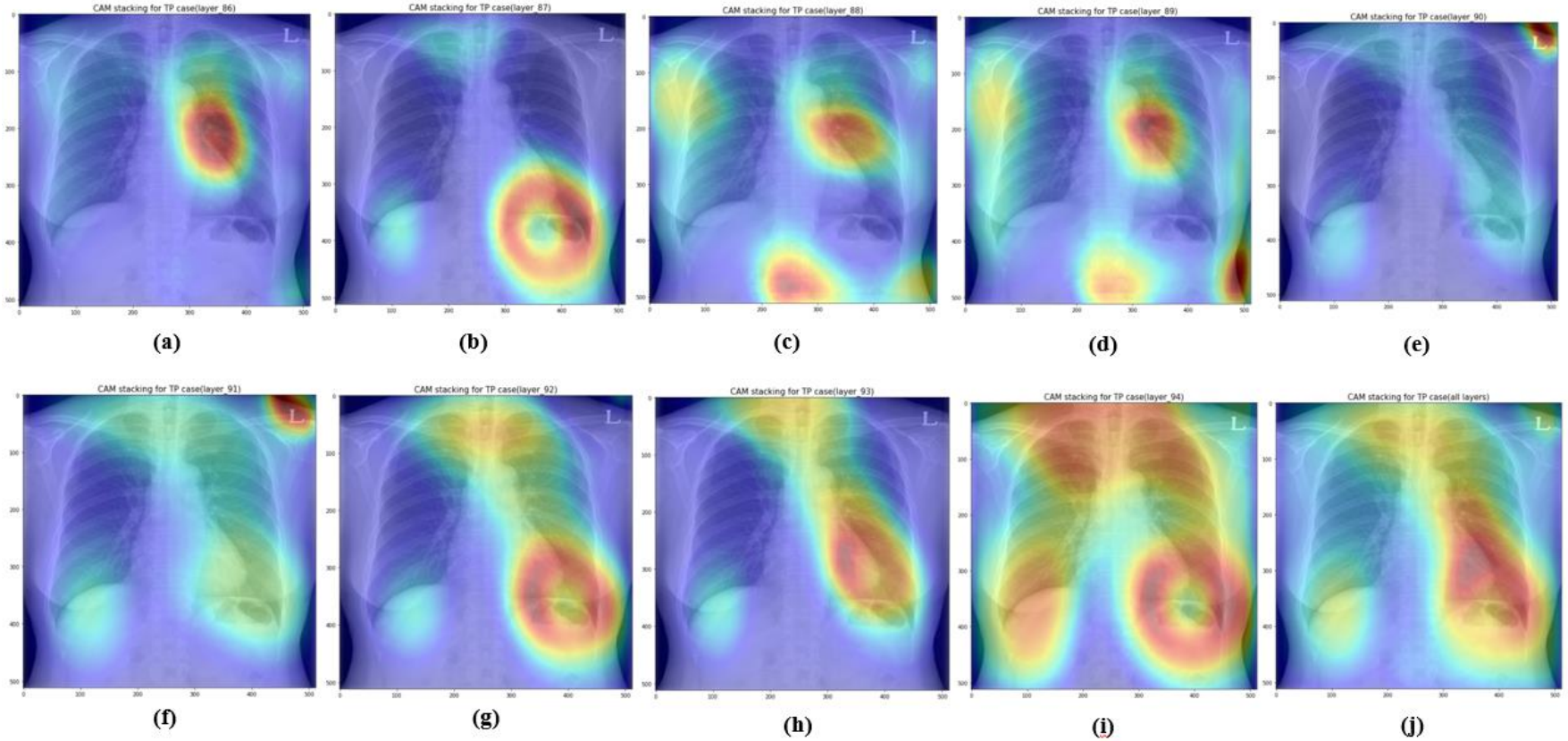


Figure 4.16. Average Grad-CAMs from subgroup transfer model of each convolution layer; (a) 86-layer, (b) 87-layer, (c) 88-layer, (d) 89-layer, (e) 90-layer, (f) 91-layer, (g) 92-layer, (h) 93-layer, (i) 94-layer, (j) average of 9 layers Grad-CAMs.

Transfer learning model from age assessment pre-trained weights showed similar performance compared to sub-group classification transfer model. The average 94-layer Grad-CAM of age assessment transfer model for a total 543 true positive case in internal validation dataset shows in Figure 4.17. Although the performance of sub-group classification transfer model and age assessment transfer model was similar, the average Grad-CAMs of the two models were very different.

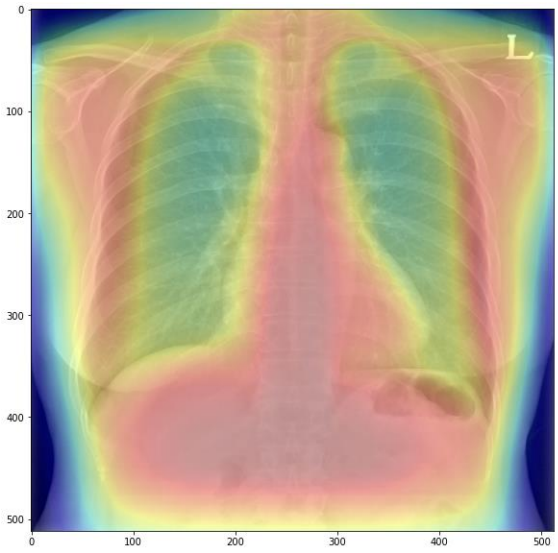


Figure 4.17. 94-layer average Grad-CAMs from age transfer model.

5. Integration in clinical workflow

A screening test is used to discover probable health issues or diseases in persons who are symptom-free. Early detection and lifestyle adjustments, as well as surveillance, are the goals in order to lower the risk of disease or to find it early enough to treat it effectively. Screening tests are not diagnostic, but they are used to identify a subset of the population who should undergo further testing to ascertain whether or not they have a disease. The ability of a screening test to discover possible problems while minimizing unclear, ambiguous, or

confusing outcomes is what makes it valuable. While screening tests are not 100 percent accurate in all circumstances, having them at the appropriate times, as indicated by a healthcare provider, is often more helpful than not having them at all.

We developed the deep learning model for opportunistic osteoporosis screening using chest radiographs. If a health checkup that does not include DXA was performed, or if a chest radiography was performed with suspicion of lung disease, and the findings were normal, this model can be additionally applied. Also, in Korea, DXA is only performed on women aged 55 and 66 in the national health checkup, so screening for osteoporosis is necessary for men. Because all major fractures were associated with increased mortality, especially in men. [127] This model can be easily used by simply uploading a chest radiograph DICOM file on cloud system from a primary care institution. Figure 5.1 shows an example of osteoporosis screening model using cloud system.

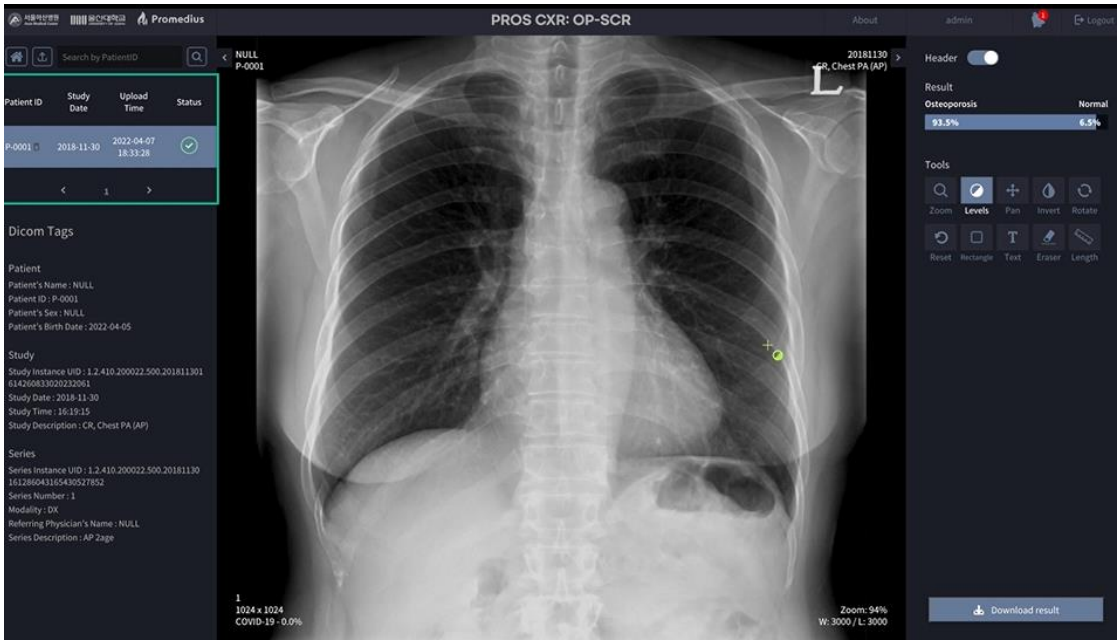


Figure 5.1. Example of osteoporosis screening model using cloud system.

6. Discussion

In this study, the deep learning-based model identified patients with osteoporosis on a single chest radiograph. In terms of performance, the screening model had an AUC of 0.88, sensitivity of 86.16%, and F1-score of 76.16% and thus could be considered useful for selecting patients at high risk of osteoporosis. This is the first study that aimed to screening for osteoporosis by using common deep learning models based on conventional chest radiographs with DXA measures and validated in other clinical indication sets. The results demonstrated that the deep-learning approach might have potential for automated osteoporosis screening on chest radiographs in the real-world clinical setting.

The same data were provided to train deep learning models with various image sizes of input matrix and various anatomical ROIs. The trained model with an input size of 512 × 512 pixels on the whole-chest image had better performance than on those with various input ROIs of the same batch size. In addition, this model with an input size of 512 × 512 pixels on the whole-chest image had better performance than that with an input size of 1024 × 1024 pixels with the same batch size. We believed that although a larger input image size provides more detailed information, the model performance becomes saturated at some point owing to its limited ability of receiving useful information and limited number of datasets. [128] As we tested sparse intervals of image sizes, we could not say that the optimal image size is 512 pixels. However, we observed that the input image size significantly affected the performance of the deep learning-based classification for screening osteoporosis on chest radiographs and the optimal image size of chest radiographs could enhance deep learning models performance. To confirm our observation, future studies are warranted to examine more various input image sizes with different kinds of deep learning models and dataset scales.

A significant proportion of the elderly population showed multifactorial age-related bone loss and consequent osteoporosis are multifactorial, accompanied by progressive loss of bone mass and quality. [129] The proportion of the cortical thickness to the total diameters of the clavicles, ribs and thoracic vertebrae decreased with age. [129-131] Kent et al. described that the one of ribs changes that occur with age and account for the decrease in injury tolerance in rib fractures in the elderly population was compositional changes (including cross sectional cortex area). [132] Several studies showed rib fractures should be considered to be osteoporotic fractures in older population. [133, 134] Although Grad-CAMs of the trained model with whole-chest images in sub-studies less highlight the lower thoracic and lumbar areas because the lower thoracic spine area overlapped the heart shade. However, the trained model with the patch areas of thoracic and lumbar spine also showed good performance in classifying normal conditions or osteoporosis. The osteoporotic changes of any bones on chest radiographs could provide osteoporosis screening with a deep learning-based model by comparing densities of bone and the surrounding soft tissues.

A few deep learning-based osteoporosis screening studies used radiographs such as lumbar spine, hip, and dental panoramic radiographs. [88, 122-124] Deep learning models based on the mandibular region on dental panoramic radiographs, femoral neck area on hip radiographs, and vertebrae on lumbar spine radiographs demonstrated the feasibility of BMD estimation and osteoporosis screening. The study of BMD measurement on low-dose chest CT using deep learning models showed good performances. [135]

In Yamamoto' research, they hypothesized that combining image features with patient data would the accuracy of osteoporosis classification using CNN. [124] They selected four clinical covariate types, namely age, sex, BMI, and history of hip fracture. They showed that the addition of patient variables offered important information, which improved the sensitivity and AUC score. In case of stress test in our study, CNN model trained with demographic

information showed better performance compared to CNN model trained only image. However, when the number of image data to be trained exceeds 10000, there was little difference in performance between the model which only the image was learned and the model that learned both the image and demographic information.

The appropriate number of samples is determined by the unique problem, and each case should be tested separately. For image classification using deep learning, a rule of thumb is 1,000 images per class. For effective generalization of the problem, a CNN algorithm may be trained with a data set bigger than 5000 samples. Generally, the larger the training set, the higher the classification performance. [136] Most of the data on osteoporosis screening using deep learning-based simple radiographs published so far were about 1000. [88, 122-124] In particular, there is a limit to collecting a lot of DXA data paired with lumbar radiographs or hip radiographs or dental panoramic radiographs. Therefore, there may be limitations in improving the performance of deep learning models using these images.

In addition, when processing images included in CNN training, such as cropping a specific area, an error may exist in this process, and it is impossible to obtain a lot of training data. In all of the four papers mentioned above, training and test data was secured by the doctor cutting a specific area of the radiographs. The previous studies were conducted using the ROI extracted from raw images. Although all radiography protocols have been established, there is no image that is completely captured in the image in the same shape as an adult chest X-ray. Deep learning research using chest radiographs can be automated in most cases even if specific image region cropping is required. This is also an advantage of this model using chest radiographs. In the real clinical setting, extraction of a ROI can be difficult and laboring task. However, our study shows that the model performance trained using whole-chest radiographs displayed even better performance than that trained using ROI-based chest radiographs in relatively many persons, over 1000 individuals.

Osteoporosis screening should, in theory, offer an estimate of the absolute risk of any fragility fracture occurring in the next 5 to 10 years. [137] Bone strength mostly reflects the integration of bone density and bone quality. [138] Since microarchitectural deterioration cannot be directly measured, BMD accounts for only 70% of bone strength. [139] DXA, which measures BMD at the lumbar spine and proximal femur, is a reliable and safe way to estimate the risk of fracture. However, DXA is not recommended universal screening. International Society for Clinical Densitometry recommends all women aged 65 years and older and men aged 70 years and older regardless of risk factors and Postmenopausal women and men age 50 to 70 years when risk factors are present. This model could screen persons 50 and older who are at high risk for osteoporosis in a timely and cost-effective manner. As a result, this model will aid in the identification of adults over the age of 50 who have unidentified risk factors for osteoporosis.

Grad-CAM needs to be interpreted with caution. [11] They localize the spatial regions, which could mainly contribute to predictions. Therefore, understanding the Grad-CAM images of correctly predicted osteoporosis cases is important. Visualization of the attention map in the final block layers is obtained by the weighted summation of all the feature maps on each layer where weight controls the importance of individual feature maps depending on the probability. In the average Grad-CAMs of osteoporosis cases, diffuse regions of the lung, the proximal ribs, the humerus head, the scapula, their nearby soft tissues, and part of waist were highlighted. This indicated that the model learned density differences between bone and the surrounding soft tissue to predict whether the input chest radiograph is osteoporosis or non-osteoporosis. This is the limitation of chest radiography which uses single X-ray contrary to the DAX which uses dual-energy X-rays with two distinct energies for subtracting soft tissue amount. For further study, we will work to explain individual predictions, which is critical for physicians and patients to accept deep learning results. [140]

Our study has several limitations. First, our model was trained solely in normal chest radiographs reported normal. We could not determine how the model predictions affect chest radiographs containing trivial abnormal findings such as small calcification nodules. Even though there could be chest radiographs with scar of old fractures. However, because the training and validation datasets were relatively large and radiologic reports of chest radiographs were known to include small human error [141], Osteoporosis screening model performance can be robust to chest radiographs having minor abnormal findings. Further research is needed with the more pair datasets of DXAs and chest radiographs including various chest radiographs, patient data on other diseases, implant devices, and anteroposterior chest radiographs, for screening purposes. Second, our datasets were established from one center. However, AMC is the largest hospital and most preferred hospital by patients in South Korea, with 2705 beds serving 11885 outpatients and 2540 inpatients on average per day. The Asan osteoporosis cohort dataset is comprised of the chest radiographs of outpatients, inpatients, and patients in the emergency department from 2006 to 2019 and patients in the health screening and promotion center from 2008 to 2011 (Table 4.1). Osteoporosis screening model was validated using this external test dataset collected over a 12-year period. Osteoporosis screening model needs more external datasets for verifying its performance exactly. In addition, training using more datasets is needed to develop better generalized models for osteoporosis screening. Thirdly, results interpretation of deep learning-based classification models is challenging. Although we used Grad-CAM to understand the specific regions used for the model prediction, it is limited to quantifying interpretable clinical features. Therefore, we presented average Grad-CAMs of all true positive cases in the internal and external sets. Furthermore, osteoporosis diagnosed from lower T-score of the lumbar spine, the femur neck, or the total femur, the disease implies systemic bone metabolic changes. CNN model can detect bone density changes on chest radiographs including bones such as the

humerus head, the ribs, and the part of spine compared with surrounding soft tissue. Further study is needed to identify the bone density of each bone shown on a chest radiograph by composing chest radiographs and chest CTs capable of Quantitative CT as a pair set. Through this study, an individual interpretation of the Osteoporosis screening model prediction will be possible. New CNN attribution method, class activation latent mapping (CALM) was introduced, and it can learn to predict the location of the cue for recognition. [142] In addition, more diverse interpretable tools of deep learning models need to be developed and new methods will be applied these models in further study. These tools could be explainable of individual predictions. Fourth, Osteoporosis screening model could have better performance with additional clinical information from electronic medical records (EMR), such as drug history, fracture history, laboratory results. EMR data cleansing for deep-learning training is very human labor task. Further study aims to train better performed deep-learning model with integrating EMR data by mining data with machine learning.

7. Conclusion

Chest radiograph-derived age using deep learning may have potential to predict prognosis of disease or diagnose of disease from one radiograph of individual. In this study showed chest radiograph-derived age correlated more with bone density than chronological age. Furthermore, it was shown that when training with a small number of data (less than 1000 cases), it is vital to check for sex imbalances in advance, which might lead to poor performance or could not be generalized. We also provided a deep learning-based model that achieves favorable performance on opportunistic osteoporosis screening using chest radiographs in middle-to-old aged persons. Routine chest radiographies obtained for other reasons in various clinical settings can be applied to identify patients at risk of osteoporosis without additional

radiation exposure or cost, which could improve osteoporosis screening. Although the deep learning models could not replace DXA for BMD screening, it could be used when chest radiography is readily available where DXA has not been performed. Therefore, identifying high-risk groups for osteoporosis using common chest radiographs will increase the disease recognition and prevent osteoporotic fractures.

8. Reference

1. Kim, M., et al., *Deep Learning in Medical Imaging.* Neurospine, 2020. **17**(2): p. 471-472.
2. Soffer, S., et al., *Convolutional neural networks for radiologic images: a radiologist's guide.* 2019. **290**(3): p. 590-606.
3. Litjens, G., et al., *A survey on deep learning in medical image analysis.* Med Image Anal, 2017. **42**: p. 60-88.
4. Chartrand, G., et al., *Deep Learning: A Primer for Radiologists.* Radiographics, 2017. **37**(7): p. 2113-2131.
5. Lee, J.G., et al., *Deep Learning in Medical Imaging: General Overview.* Korean J Radiol, 2017. **18**(4): p. 570-584.
6. Raghu, V.K., et al., *Deep Learning to Estimate Biological Age From Chest Radiographs.* JACC Cardiovasc Imaging, 2021.
7. Kisling LA, M.D.J. *Prevention Strategies*. Updated 2021 May 9 2021 Jan-.]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK537222.
8. Ron, E., *CANCER RISKS FROM MEDICAL RADIATION.* 2003. **85**(1): p. 47-59.
9. Irvin, J., et al. *Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison*. in *Proceedings of the AAAI conference on artificial intelligence*. 2019.
10. Wang, X., et al. *Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
11. Lu, M.T., et al., *Deep Learning to Assess Long-term Mortality From Chest Radiographs.* JAMA Netw Open, 2019. **2**(7): p. e197416.
12. Lu, M.T., et al., *Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model.* Annals of Internal Medicine, 2020. **173**(9): p. 704-713.
13. Sabottke, C.F., M.A. Breaux, and B.M. Spieler, *Estimation of age in unidentified patients via chest radiography using convolutional neural network regression.* Emerg Radiol, 2020. **27**(5): p. 463-468.
14. Li, D., et al., *Deep learning prediction of sex on chest radiographs: a potential contributor to biased algorithms.* Emerg Radiol, 2022. **29**(2): p. 365-370.
15. Yi, P.H., et al., *Radiology "forensics": determination of age and sex from chest radiographs using deep learning.* Emerg Radiol, 2021. **28**(5): p. 949-954.

16. Yang, C.Y., et al., *Using Deep Neural Networks for Predicting Age and Sex in Healthy Adult Chest Radiographs.* J Clin Med, 2021. **10**(19).
17. Perni, S., et al., *Association of a Deep Learning Estimation of Chest Imaging Age With Survival in Patients With Non-Small Cell Lung Cancers Undergoing Radiation.* International Journal of Radiation Oncology, Biology, Physics, 2021. **111**(3S): p. S114-S114.
18. *NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy, March 7-29, 2000: highlights of the conference.* South Med J, 2001. **94**(6): p. 569-73.
19. Curtis, E.M., et al., *The impact of fragility fracture and approaches to osteoporosis risk assessment worldwide.* Bone, 2017. **104**: p. 29-38.
20. Bliuc, D., et al., *Compound risk of high mortality following osteoporotic fracture and refracture in elderly women and men.* J Bone Miner Res, 2013. **28**(11): p. 2317-24.
21. Sozen, T., L. Ozisik, and N.C. Basaran, *An overview and management of osteoporosis.* Eur J Rheumatol, 2017. **4**(1): p. 46-56.
22. Melton, L.J., 3rd, et al., *Bone density and fracture risk in men.* J Bone Miner Res, 1998. **13**(12): p. 1915-23.
23. Melton, L.J., et al., *How Many Women Have Osteoporosis.* Journal of Bone and Mineral Research, 1992. **7**(9): p. 1005-1010.
24. Marcucci, G. and M.L. Brandi, *Rare causes of osteoporosis.* Clin Cases Miner Bone Metab, 2015. **12**(2): p. 151-6.
25. Hamdy, R.C., *Osteoporosis, the deafening silent epidemic.* South Med J, 2002. **95**(6): p. 567-8.
26. Dimai, H.P., *Use of dual-energy X-ray absorptiometry (DXA) for diagnosis and fracture risk assessment; WHO-criteria, T- and Z-score, and reference databases.* Bone, 2017. **104**: p. 39-43.
27. Filler, A., *The history, development and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, and DTI.* Nature precedings, 2009: p. 1-1.
28. Rowlands, J.A., *Current advances and future trends in X-ray digital detectors for medical applications.* IEEE Transactions on Instrumentation and Measurement, 1998. **47**(6): p. 1415-1418.
29. Bar, Y., et al. *Deep learning with non-medical training used for chest pathology identification*. in *Medical Imaging 2015: Computer-Aided Diagnosis*. 2015. International Society for Optics and Photonics.
30. Park, B., et al., *A Curriculum Learning Strategy to Enhance the Accuracy of*

Classification of Various Lesions in Chest-PA X-ray Screening for Pulmonary Abnormalities. Scientific Reports, 2019. **9**(1): p. 1-9.

31. Albahli, S., *Efficient GAN-based Chest Radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia.* Int J Med Sci, 2020. **17**(10): p. 1439-1448.
32. Briot, K., *DXA parameters: beyond bone mineral density.* Joint Bone Spine, 2013. **80**(3): p. 265-9.
33. Blake, G.M. and I. Fogelman. *Technical principles of dual energy x-ray absorptiometry*. in *Seminars in nuclear medicine*. 1997. Elsevier.
34. Watts, N.B., *Fundamentals and pitfalls of bone densitometry using dual-energy X-ray absorptiometry (DXA).* Osteoporosis international, 2004. **15**(11): p. 847-854.
35. McCulloch, W.S. and W. Pitts, *A logical calculus of the ideas immanent in nervous activity.* The bulletin of mathematical biophysics, 1943. **5**(4): p. 115-133.
36. Rosenblatt, F., *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychological review, 1958. **65**(6): p. 386.
37. Widrow, B. and M.E. Hoff, *Associative storage and retrieval of digital information in networks of adaptive "neurons"*, in *Biological Prototypes and Synthetic Systems*. 1962, Springer. p. 160-160.
38. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning internal representations by error propagation*. 1985, California Univ San Diego La Jolla Inst for Cognitive Science.
39. Ko, H., et al., *The emergence of functional microcircuits in visual cortex.* Nature, 2013. **496**(7443): p. 96-100.
40. Bengio, Y., A. Courville, and P. Vincent, *Representation learning: A review and new perspectives.* IEEE transactions on pattern analysis and machine intelligence, 2013. **35**(8): p. 1798-1828.
41. Bengio, Y. *Deep learning of representations: Looking forward*. in *International conference on statistical language and speech processing*. 2013. Springer.
42. Bengio, Y., *Learning deep architectures for AI*. 2009: Now Publishers Inc.
43. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 2012. **25**.
44. Farabet, C., et al., *Learning hierarchical features for scene labeling.* IEEE transactions on pattern analysis and machine intelligence, 2012. **35**(8): p. 1915-1929.
45. Tompson, J.J., et al., *Joint training of a convolutional network and a graphical model for human pose estimation.* Advances in neural information processing systems, 2014. **27**.

46. Szegedy, C., et al. *Going deeper with convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
47. Mikolov, T., et al. *Strategies for training large scale neural network language models*. in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. 2011. IEEE.
48. Hinton, G., et al., *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.* IEEE Signal processing magazine, 2012. **29**(6): p. 82-97.
49. Sainath, T.N., et al. *Improvements to deep convolutional neural networks for LVCSR*. in *2013 IEEE workshop on automatic speech recognition and understanding*. 2013. IEEE.
50. Ma, J., et al., *Deep neural nets as a method for quantitative structure–activity relationships.* Journal of chemical information and modeling, 2015. **55**(2): p. 263-274.
51. Ciodaro, T., et al. *Online particle detection with neural networks based on topological calorimetry information*. in *Journal of physics: conference series*. 2012. IOP Publishing.
52. Greenspan, H., B. Van Ginneken, and R.M. Summers, *Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique.* IEEE transactions on medical imaging, 2016. **35**(5): p. 1153-1159.
53. Lee, J.-G., et al., *Deep learning in medical imaging: general overview.* Korean journal of radiology, 2017. **18**(4): p. 570-584.
54. Xu, Y., et al., *Deep learning predicts lung cancer treatment response from serial medical imaging.* Clinical Cancer Research, 2019. **25**(11): p. 3266-3275.
55. Bello, G.A., et al., *Deep-learning cardiac motion analysis for human survival prediction.* Nature machine intelligence, 2019. **1**(2): p. 95-104.
56. LeCun, Y., et al., *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 1998. **86**(11): p. 2278-2324.
57. Goodfellow, I., *YoshuaBengio. Deep learning. Vol. 1, no. 2*. 2016, Cambridge: MIT press.
58. Lin, M., Q. Chen, and S. Yan, *Network in network.* arXiv preprint arXiv:1312.4400, 2013.
59. Nair, V. and G.E. Hinton. *Rectified linear units improve restricted boltzmann machines*. in *Icml*. 2010.
60. Xu, B., et al., *Empirical evaluation of rectified activations in convolutional network.* arXiv preprint arXiv:1505.00853, 2015.

61. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting*. The journal of machine learning research, 2014. **15**(1): p. 1929-1958.
62. Ioffe, S. and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. in *International conference on machine learning*. 2015. PMLR.
63. Shimodaira, H., *Improving predictive inference under covariate shift by weighting the log-likelihood function.* Journal of statistical planning and inference, 2000. **90**(2): p. 227-244.
64. Glorot, X. and Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks*. in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010. JMLR Workshop and Conference Proceedings.
65. Cortes, C., M. Mohri, and A. Rostamizadeh, *L2 regularization for learning kernels.* arXiv preprint arXiv:1205.2653, 2012.
66. Simard, P.Y., D. Steinkraus, and J.C. Platt. *Best practices for convolutional neural networks applied to visual document analysis*. in *Icdar*. 2003.
67. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
68. Huang, G., et al. *Densely connected convolutional networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
69. Chollet, F. *Xception: Deep learning with depthwise separable convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
70. Tan, M. and Q. Le. *Efficientnet: Rethinking model scaling for convolutional neural networks*. in *International conference on machine learning*. 2019. PMLR.
71. Haralick, R.M., K. Shanmugam, and I.H. Dinstein, *Textural features for image classification.* IEEE Transactions on systems, man, and cybernetics, 1973(6): p. 610-621.
72. Ausawalaithong, W., et al. *Automatic lung cancer prediction from chest X-ray images using the deep learning approach*. in *2018 11th Biomedical Engineering International Conference (BMEiCON)*. 2018. IEEE.
73. Chan, T.-H., et al., *PCANet: A simple deep learning baseline for image classification?* IEEE transactions on image processing, 2015. **24**(12): p. 5017-5032.
74. Zeng, R., et al., *Color image classification via quaternion principal component analysis network.* Neurocomputing, 2016. **216**: p. 416-428.
75. Tan, C., et al. *A survey on deep transfer learning*. in *International conference on*

artificial neural networks. 2018. Springer.

76. Morid, M.A., A. Borjali, and G. Del Fiol, *A scoping review of transfer learning research on medical image analysis using ImageNet.* Computers in biology and medicine, 2021. **128**: p. 104115.
77. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.
78. Alzubaidi, L., et al., *Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model.* Electronics, 2020. **9**(3): p. 445.
79. Shin, H.-C., et al., *Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning.* IEEE transactions on medical imaging, 2016. **35**(5): p. 1285-1298.
80. Alzubaidi, L., et al., *Novel transfer learning approach for medical imaging with limited labeled data.* Cancers, 2021. **13**(7): p. 1590.
81. Raghu, M., et al., *Transfusion: Understanding transfer learning for medical imaging.* Advances in neural information processing systems, 2019. **32**.
82. Gulshan, V., et al., *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.* Jama, 2016. **316**(22): p. 2402-2410.
83. Alzubaidi, L., et al., *Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis.* Electronics, 2020. **9**(3): p. 427.
84. Heker, M. and H. Greenspan, *Joint liver lesion segmentation and classification via transfer learning.* arXiv preprint arXiv:2004.12352, 2020.
85. Zhou, B., et al. *Learning deep features for discriminative localization*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
86. Selvaraju, R.R., et al. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
87. Raghu, V.K., et al., *Deep Learning to Estimate Biological Age From Chest Radiographs.* JACC Cardiovasc Imaging, 2021. **14**(11): p. 2226-2236.
88. Jang, R., et al., *Prediction of osteoporosis from simple hip radiography using deep learning algorithm.* Scientific reports, 2021. **11**(1): p. 1-9.
89. Poplin, R., et al., *Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning.* 2018. **2**(3): p. 158-164.
90. Pallagatti, S., et al., *Efficacy of Panoramic Radiography in the Detection of*

Osteoporosis in Post-Menopausal Women When Compared to Dual Energy X-Ray Absorptiometry. Open Dent J, 2017. **11**: p. 350-359.

91. Milea, D., et al., *Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs.* N Engl J Med, 2020. **382**(18): p. 1687-1695.
92. Force, U.S.P.S.T., et al., *Screening for Osteoporosis to Prevent Fractures: US Preventive Services Task Force Recommendation Statement.* JAMA, 2018. **319**(24): p. 2521-2531.
93. Park, E.J., et al., *Prevalence of osteoporosis in the Korean population based on Korea National Health and Nutrition Examination Survey (KNHANES), 2008-2011.* Yonsei Med J, 2014. **55**(4): p. 1049-57.
94. Unnanuntana, A., et al., *The assessment of fracture risk.* The Journal of bone and joint surgery. American volume, 2010. **92**(3): p. 743-753.
95. Haaland, D.A., et al., *Closing the osteoporosis care gap–Increased osteoporosis awareness among geriatrics and rehabilitation teams.* 2009. **9**(1): p. 1-9.
96. Sedlak, C.A., M.O. Doheny, and S.L. Jones, *Osteoporosis education programs: changing knowledge and behaviors.* Public Health Nurs, 2000. **17**(5): p. 398-402.
97. Mithal, A., et al., *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**(4): p. 449-54.
98. King, A.B. and D.M.J.H.A. Fiorentino, *Medicare payment cuts for osteoporosis testing reduced use despite tests' benefit in reducing fractures.* 2011. **30**(12): p. 2362-2370.
99. Amarnath, A.L., et al., *Underuse and Overuse of Osteoporosis Screening in a Regional Health System: a Retrospective Cohort Study.* J Gen Intern Med, 2015. **30**(12): p. 1733-40.
100. Kim, K.H., et al., *Prevalence, awareness, and treatment of osteoporosis among Korean women: The Fourth Korea National Health and Nutrition Examination Survey.* Bone, 2012. **50**(5): p. 1039-1047.
101. Karargyris, A., et al. *Age prediction using a large chest x-ray dataset*. in *Medical Imaging 2019: Computer-Aided Diagnosis*. 2019. SPIE.
102. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
103. Szegedy, C., et al. *Rethinking the inception architecture for computer vision*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
104. Gaser, C., et al., *BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease.* PLoS One, 2013. **8**(6): p. e67346.

105. Franke, K., et al., *Advanced BrainAGE in older adults with type 2 diabetes mellitus.* Front Aging Neurosci, 2013. **5**: p. 90.
106. Lowe, L.C., et al., *The Effect of the APOE Genotype on Individual BrainAGE in Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease.* PLoS One, 2016. **11**(7): p. e0157514.
107. Faulkner, K.G., *The tale of the T-score: review and perspective.* Osteoporos Int, 2005. **16**(4): p. 347-52.
108. Yang, C.-Y., et al., *Using Deep Neural Networks for Predicting Age and Sex in Healthy Adult Chest Radiographs.* Journal of Clinical Medicine, 2021. **10**(19): p. 4431.
109. Larrazabal, A.J., et al., *Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis.* Proceedings of the National Academy of Sciences, 2020. **117**(23): p. 12592-12594.
110. Seyyed-Kalantari, L., et al. *CheXclusion: Fairness gaps in deep chest X-ray classifiers*. in *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*. 2020. World Scientific.
111. Zech, J.R., et al., *Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study.* PLoS medicine, 2018. **15**(11): p. e1002683.
112. Larson, D.B., et al., *Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs.* Radiology, 2018. **287**(1): p. 313-322.
113. Cole, J.H., et al., *Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker.* Neuroimage, 2017. **163**: p. 115-124.
114. Cole, J.H., et al., *Brain-predicted age in Down syndrome is associated with beta amyloid deposition and cognitive decline.* Neurobiology of aging, 2017. **56**: p. 41-49.
115. Cole, J.H., et al., *Increased brain-predicted aging in treated HIV disease.* Neurology, 2017. **88**(14): p. 1349-1357.
116. Steffener, J., et al., *Differences between chronological and brain age are related to education and self-reported physical activity.* Neurobiology of aging, 2016. **40**: p. 138-144.
117. Luders, E., N. Cherbuin, and F. Kurth, *Forever Young (er): potential age-defying effects of long-term meditation on gray matter atrophy.* Frontiers in Psychology, 2015. **5**: p. 1551.
118. Ieki, H., et al., *Deep learning-based chest X-ray age serves as a novel biomarker for cardiovascular aging.* bioRxiv, 2021.

119. Ieki, H., et al., *Deep learning-based chest X-ray age serves as a novel biomarker for cardiovascular aging.* bioRxiv, 2021: p. 2021.03.24.436773.
120. Lee, S.H., et al., *High Circulating Sphingosine 1-Phosphate is a Risk Factor for Osteoporotic Fracture Independent of Fracture Risk Assessment Tool.* Calcified Tissue International, 2020. **107**(4): p. 362-370.
121. Park, E.J., et al., *Prevalence of Osteoporosis in the Korean Population Based on Korea National Health and Nutrition Examination Survey (KNHANES), 2008-2011.* Yonsei Med J, 2014. **55**(4): p. 1049-1057.
122. Zhang, B., et al., *Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study.* Bone, 2020. **140**: p. 115561.
123. Lee, K.S., et al., *Evaluation of Transfer Learning with Deep Convolutional Neural Networks for Screening Osteoporosis in Dental Panoramic Radiographs.* J Clin Med, 2020. **9**(2).
124. Yamamoto, N., et al., *Deep Learning for Osteoporosis Classification Using Hip Radiographs and Patient Clinical Covariates.* Biomolecules, 2020. **10**(11).
125. Nogueira, K., O.A. Penatti, and J.A. Dos Santos, *Towards better exploiting convolutional neural networks for remote sensing scene classification.* Pattern Recognition, 2017. **61**: p. 539-556.
126. Kim, S., et al., *An open medical platform to share source code and various pre-trained weights for models to use in deep learning research.* Korean journal of radiology, 2021. **22**(12): p. 2073.
127. Center, J.R., et al., *Mortality after all major types of osteoporotic fracture in men and women: an observational study.* The Lancet, 1999. **353**(9156): p. 878-882.
128. Kim, Y.G., et al., *Optimal matrix size of chest radiographs for computer-aided detection on lung nodule or mass with deep learning.* Eur Radiol, 2020. **30**(9): p. 4943-4951.
129. Chen, H., et al., *Age-related changes in trabecular and cortical bone microstructure.* Int J Endocrinol, 2013. **2013**: p. 213234.
130. Holcombe, S.A., S.C. Wang, and J.B. Grotberg, *Age-related changes in thoracic skeletal geometry of elderly females.* Traffic Inj Prev, 2017. **18**(sup1): p. S122-S128.
131. Kaur, H. and I. Jit, *Age estimation from cortical index of the human clavicle in northwest Indians.* Am J Phys Anthropol, 1990. **83**(3): p. 297-305.
132. Kent, R., et al., *Structural and material changes in the aging thorax and their role in crash protection for older occupants*. 2005, SAE Technical Paper.
133. Barrett-Connor, E., et al., *Epidemiology of rib fractures in older men: Osteoporotic*

Fractures in Men (MrOS) prospective cohort study. BMJ, 2010. **340**: p. c1069.

134. Sajjan, S.G., et al., *Rib fracture as a predictor of future fractures in young and older postmenopausal women: National Osteoporosis Risk Assessment (NORA).* Osteoporos Int, 2012. **23**(3): p. 821-8.
135. Pan, Y., et al., *Automatic opportunistic osteoporosis screening using low-dose chest computed tomography scans obtained for lung cancer screening.* Eur Radiol, 2020. **30**(7): p. 4107-4116.
136. Luo, C., et al. *How does the data set affect cnn-based image classification performance?* in *2018 5th International conference on systems and informatics (ICSAI)*. 2018. IEEE.
137. Raisz, L.G., *Screening for Osteoporosis.* New England Journal of Medicine, 2005. **353**(2): p. 164-171.
138. Melton, L.J., 3rd, et al., *Contributions of bone density and structure to fracture risk assessment in men and women.* Osteoporos Int, 2005. **16**(5): p. 460-7.
139. Ammann, P. and R. Rizzoli, *Bone strength and its determinants.* Osteoporos Int, 2003. **14 Suppl 3**: p. S13-8.
140. Holzinger, A., et al., *What do we need to build explainable AI systems for the medical domain?* 2017.
141. Jang, R., et al., *Assessment of the Robustness of Convolutional Neural Networks in Labeling Noise by Using Chest X-Ray Images From Multiple Centers.* Jmir Medical Informatics, 2020. **8**(8): p. e18089.
142. Kim, J.M., et al. *Keep CALM and Improve Visual Feature Attribution*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

9. Abstract (In Korean)

의료분야에서 컴퓨터 알고리즘 발전과 그래픽 처리장치 (GPU), 의료 빅데이터가 의료인공지능의 급속한 발전에 기여하고 있다. 특히 의료에서 패턴 인식 및 기계학습기술이 널리 사용되며 의료영상 분석시스템의 기반이 된다. 특히 딥러닝 (deep learning)은 기계학습에서 가장 중요한 기술로 의료 분야에서 활발히 개발되고 있으며 최근 몇 년간 이를 이용한 연구가 크게 증가하고 있다.

의료 영상에서 임상적으로 유용한 성능을 보여주는 딥러닝 모델이 실제 의료 현장에서 점차적으로 도입되고 있다. 소아청소년과에서 손 방사선 사진으로 골연령을 측정하는데, 골연령 평가에 대한 인공지능 보조프로그램을 사용하는 것이 한국에서는 흔한 일이 되었다. 또한 흉부 방사선 사진에서 질환의 유무를 일차적으로 선별하는 인공지능 프로그램도 많은 병원에서 도입하고 있다. 일상생활 곳곳에서 인공지능 맞춤형 서비스가 늘어나고 있듯이 의료분야에서도 임상 현장에서 사용가능한 인공지능 개발에 힘쓰고 있다.

의료분야에서는 질병의 진단, 치료의 결정, 예후의 예측, 질병의 예방도 또한 중요한 문제이다. 질병의 예방은 크게 3 단계로 구분하는데, 1 차 예방은 질환이나 손상 발생의 위험요인을 줄이는 것, 2 차 예방은 질환의 진행단계에서 조기 발견함으로써 상태를 완화하거나 진행을 차단하는 것, 3 차 예방은 발견된 질환을 치료함으로써 질병의 진행을 막는 것을 말한다. 1 차 예방이 가장 사회적인 손실을 줄이면서 비용 효과적인데, 건강한 개인을 대상으로 어떠한 질병의 고위험군을 선택하는 것이 중요하다.

흉부 방사선 사진은 의료분야에서 가장 흔하고 매우 적은 양의 전리방사선을 이용하기 때문에 비침습적인 영상이며 이를 포함하는 공개데이터셋이 가장 많다. 일반적으로 흉부 방사선 사진은 결핵, 폐렴, 폐암 등 폐에 있는 질환을 찾기 위한 1 차 검진 검사영상으로, 대부분은 특정 진단을 배제하는 방식으로 보고하며 많은 수의 영상이 정상 판독 영상이다. 영상의 촬영 방식이 전세계적으로 통일되어 있으며 적절하게 촬영된 흉부 방사선 사진은 견갑골(scapula) 내측 경계부가 폐첨부나 상부 폐야를 가리지 않고 흉곽 전체를 포함하여 후두부터 양측 갈비 가로막각 (costophrenic angle)까지 보여야 한다. 이 영상은 폐뿐만 아니라 심장, 흉곽 부위 쇄골, 늑골, 경추 일부와 흉추, 요추 일부의 뼈의 정보를 포함하고 있다.

의료에서 딥러닝을 이용한 기술로 다양한 의료영상에서 의료인공지능 모델을 개발할 때, 일반적으로 의사가 판단하는 영역 내에서 그 일을 도와주는 식의 역할을 목적으로 하는 경우가 많다. 이 또한 의료인공지능 연구의 중요한 분야이나 일부 연구에서 딥러닝 모델이 사람의 인지를 넘어서는 부분을 구분할 수 있는 성능을 보여주고 있다. 최근 연구에서 흉부 방사선 사진에서 사망 여부에 대하 예후 정보를 추출할 수 있음을 보여주었고 흉부 방사선 사진에서 나이와 성별 정보를 학습하고 추출하여 흉부 방사선 사진에서 추출된 생물학적 연령이 일반 인구군에서 심혈관 사망에 대한 예측을 하거나 비소세포폐암 환자에 대한 예후 바이오 마커로 사용할 수 있다고 제안되기도 하였다.

한편 골다공증은 골밀도가 낮고 골구조의 미세구조 악화로 골절 위험이 증가하는 전신 질환으로 골다공증 및 골다공증 골절은 한국을 비롯한 고령화 국가에서 연령 관절 골절과의 연관성으로 전세계적으로 주요 건강문제가 되었다.

골다공증은 증상이 없어 골절이 발생할 때까지 발견되지 않는 경우가 많기 때문에 조기 진단이 필요한 중요한 질환이며 이로 인한 고관절, 척추 및 손목 골절은 환자의 삶의 질을 저하시키고 심한 경우 사망 위험을 증가시킨다. 검진 목적으로 흔하게 검사하는 흉부 방사선 사진을 이용하여 골다공증의 고위험군을 선별하는 것은 의료의 1 차 예방목표에 부합하는 가치 있는 일이다.

이 연구에서는 9 만여건의 정상으로 판독된 흉부 방사선 영상에서 나이와 성별 정보를 학습하였고 흉부 방사선 사진에서 예측된 나이가 골다공증과 관련이 있음을 보여 주였다. 또한 딥러닝 모델은 적은 수의 데이터에서도 성별 분류가 쉽게 가능하기 때문에 성별의 불균형이 있는 데이터에서 모델을 개발할 때의 주의가 필요함을 설명하였다. 마지막으로 건강검진센터의 만 건이상의 데이터셋을 이용하여, 골밀도기준 골다공증 진단을 기반으로 분류된 흉부 방사선 사진을 이용하여 골다공증 검진을 위한 딥러닝 모델을 개발하고 평가하였다. 다양한 방식의 학습 모델을 이용하여 딥러닝 모델을 개발하였고 나이와 성별 정보, 흉부 방사선에서 예측된 나이 정보를 이용하는 모델 또한 개발하였다. 개발된 모델들을 비교 평가하고 외부 검증 세트에서도 그 성능을 검증했다. 이러한 모델이 실제 임상 현장에 사용되기 위해서는 적절한 임상 상황에서 이 모델이 사용가능해야 한다. 이 모델은 50 세 이상에서 골다공증 고위험군을 선별하기 위한 목적으로 개발된 것으로 흉부 X-ray 검사를 하는 1 차 진료 기관에서 골다공증 고위험군에게 골밀도 검사를 권고할 수 있도록, 진료 프로세스에 포함되는 방식으로 적용되어야 할 것이다. 이를 통해 질환의 인지율을 높이고 향후에 골다공증설 골절을 감소시키는데 기여할 수 있을 것이다.

10. Acknowledgements