# Supplementary materials

## Supplement description 1. Image preprocessing

All input images were resized to 512 × 512 pixels with bicubic interpolation and were normalized with z-scores. 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.

## Supplement description 2. OsPor-screen model development

Supervised learning is one of deep learning task and defined as training a model using input data and its corresponding labels. Deep learning produces its own algorithms by "learning" associations between inputs and outputs. Training of deep learning model is generally done in batches (subsets) randomly sampled from the training dataset. The training dataset is what is used for optimizing the network weights via backpropagation. Training is performed by updating the model parameters repeatedly until the model optimally fits the data. Validation dataset is used for parameter selection and tuning and is customarily also used to implement stopping conditions for training. Internal test dataset is used to confirm the classification performance of the model that has been trained. External test dataset is to evaluate the classification performance of the model using independent datasets composed from different clinical settings, different time periods.

This study used the Inception-v3 model for binary classification as an osteoporosis screening tool. Baseline model is a default CNN model with random initialization of network. Transferred model is training a model based on the pre-trained model, the best model of whole-chest images with 512 × 512 pixels in size in the sub-group studies, as an initial of network. In this way, OsPor-screen model was trained in the transfer learning manner by tuning the last two blocks of the Inception-v3(supplement figure 1). The classification performance of baseline model is shown in supplement table 4. To adjust age effect, CNN model trained with input of chest radiograph and age together by concatenating them at the last fully connected layer of the model (Supplement figure 4).

The model was implemented in Keras with a TensorFlow backbone and adaptive moment estimation (Adam) optimizer, which optimizes learning rates efficiently during training. The Inception-v3 model was chosen because it showed good performance in the 2014 ImageNet Large Scale Visual Recognition Competition, [54] where one of the Inception models was used. Among the

Inception models, Inception-v3 is computationally efficient. [55] Considering that our task was to develop a binary classifier, we modified the Inception-v3 model by replacing the last layer with a global average pooling layer, three dense layers and two dropout layers. 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.

### Supplement description 3. Average Grad-CAMs

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. [38] 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 developed in our institution. [56] 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.

**Supplement table 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 | |

**Supplement table 2. Odds ratios (95% CI) of OsPor-screen model likelihood score and risk factors for osteoporosis with one inter-quartile range (IQR) increase by using a multiple logistic regression.**

| Variable | Odds Ratio | 95% CI | p-value |
|---|---|---|---|
| Sex (M/F) | 4.124 | (1.595, 10.664) | 0.003 |
| Menopause | 2.107 | (0.897, 4.952) | 0.087 |
| Height (IQR) | 1.072 | (0.841, 1.367) | 0.575 |
| Weight (IQR) | 0.767 | (0.608, 0.968) | 0.025 |
| Age (IQR) | 1.187 | (1.002, 1.405) | 0.047 |
| Ca (IQR) | 0.927 | (0.787, 1.091) | 0.359 |
| P (IQR) | 1.153 | (0.947, 1.403) | 0.157 |
| ALP (IQR) | 1.124 | (0.969, 1.305) | 0.123 |
| FBS (IQR) | 1.004 | (0.912, 1.105) | 0.941 |
| hsCRP(IQR) | 1.048 | (1.000, 1.099) | 0.048 |
| Likelihood score (IQR) | 39.405 | (28.378, 54.717) | 0.000 |

Ca = serum calcium level, P = serum phosphorus level, ALP = serum alkaline phosphatase level, FBS = serum fasting blood sugar level, hsCRP = serum high-sensitivity C-reactive protein level, Likelihood score = output of OsPor-screeen model.

**Supplement table 3. Age of osteoporosis and non-osteoporosis in each dataset.**

| age (year) | Train set | Validation set | Internal test set | External test set |
|---|---|---|---|---|
| Osteoporosis | 59.79 ± 7.65 | 60.25 ± 7.83 | 61.33 ± 7.13 | 60.95 ± 6.81 |
| Non-osteoporosis | 54.82 ± 7.56 | 56.53 ± 7.15 | 57.39 ± 6.17 | 57.75 ± 6.26 |

Age: mean ± std

**Supplement table 4. Performances of the models trained with various image sizes on whole-chest images.**

| Input image size | Batch size | AUC (95% CI) | Accuracy (%) (95% CI) | Sensitivity (%) (95% CI) | Specificity (%) (95% CI) |
|---|---|---|---|---|---|
| 128 | 100 | 0.92 (0.91–0.94) | 84.24 (82.24–86.05) | 84.02 (81.13–86.55) | 84.45 (81.58–86.94) |
| 256 | 100 | 0.97 (0.96–0.98) | 90.37 (88.71–91.81) | 92.15 (89.93–93.92) | 88.59 (86.02–90.73) |
| 512 | 60 | 0.99 (0.98–0.99) | 94.01(92.64–95.13) | 95.15 (93.30–96.51) | 92.87 (90.72–94.55) |
| 512 | 15 | 0.99 (0.98–0.99) | 94.44(93.11–95.52) | 95.86 (94.12–97.10) | 93.01 (90.88–94.67) |
| 1024 | 15 | 0.98 (0.97–0.99) | 94.01(92.64–95.13) | 96.86 (95.29–97.92) | 91.16 (88.82–93.04) |

**Supplement table 5. The performance metrics of the baseline model in the internal and external test datasets.**

| Datasets | AUC (95% CI) | Accuracy (%) (95% CI) | Sensitivity (%) (95% CI) | Specificity (%) (95% CI) |
|---|---|---|---|---|
| Internal test | 0.91 (0.89–0.92) | 82.81 (81.08–84.40) | 76.77 (73.40–79.83) | 85.82 (83.84–87.60) |
| External test | 0.87 (0.84–0.89) | 79.06 (76.55–81.38) | 78.93 (74.12–83.05) | 79.12 (76.11–81.84) |

**Supplement table 6. Confusion matrixes of the OsPor-screen and baseline models in the internal and external test datasets.**

(a) Confusion matrix of the baseline model in the internal test dataset.

| | | Predicted label | |
|---|---|---|---|
| | | Non-osteoporosis | Osteoporosis |
| True label | Non-osteoporosis | 1138 | 188 |
| | Osteoporosis | 154 | 509 |

(b) Confusion matrix of the baseline model in the external test dataset.

| | | Predicted label | |
|---|---|---|---|
| | | Non-osteoporosis | Osteoporosis |
| True label | Non-osteoporosis | 610 | 161 |
| | Osteoporosis | 67 | 251 |

(c) Confusion matrix of the OsPor-screen model in the internal test dataset.

| | | Predicted label | |
|---|---|---|---|
| | | Non-osteoporosis | Osteoporosis |
| True label | Non-osteoporosis | 1080 | 246 |
| | Osteoporosis | 104 | 559 |

(d) Confusion matrix of the OsPor-screen model in the external test dataset.

| | | Predicted label | |
|---|---|---|---|
| | | Non-osteoporosis | Osteoporosis |
| True label | Non-osteoporosis | 572 | 199 |
| | Osteoporosis | 44 | 274 |

**Supplement table 7. Confusion matrixes of the models trained with various inputs ROIs.**

(a) Confusion matrix of the model trained with right shoulder area

| | | Predicted label | |
|---|---|---|---|
| | | Normal | Osteoporosis |
| True label | Normal | 641 | 60 |
| | Osteoporosis | 42 | 659 |

(b) Confusion matrix of the model trained with cervical and thoracic area

| | | Predicted label | |
|---|---|---|---|
| | | Normal | Osteoporosis |
| True label | Normal | 641 | 60 |
| | Osteoporosis | 52 | 649 |

(c) Confusion matrix of the model trained with thoracic and lumbar area

| | | Predicted label | |
|---|---|---|---|
| | | Normal | Osteoporosis |
| True label | Normal | 642 | 59 |
| | Osteoporosis | 62 | 639 |

**Supplement table 8. Confusion matrixes of the models trained with various inputs image sizes.**

(a) Confusion matrix of the model trained with 128 × 128 pixel-sized image and batch size 100

| | | Predicted label | |
|---|---|---|---|
| | | Normal | Osteoporosis |
| True label | Normal | 592 | 109 |
| | Osteoporosis | 112 | 589 |

(b) Confusion matrix of the model trained with 256 × 256 pixel-sized image and batch size 100

| | | Predicted label | |
|---|---|---|---|
| | | Normal | Osteoporosis |
| True label | Normal | 621 | 80 |
| | Osteoporosis | 55 | 646 |

(c) Confusion matrix of the model trained with 512 × 512 pixel-sized image and batch size 60

| | | Predicted label | |
|---|---|---|---|
| | | Normal | Osteoporosis |
| True label | Normal | 651 | 50 |
| | Osteoporosis | 34 | 667 |

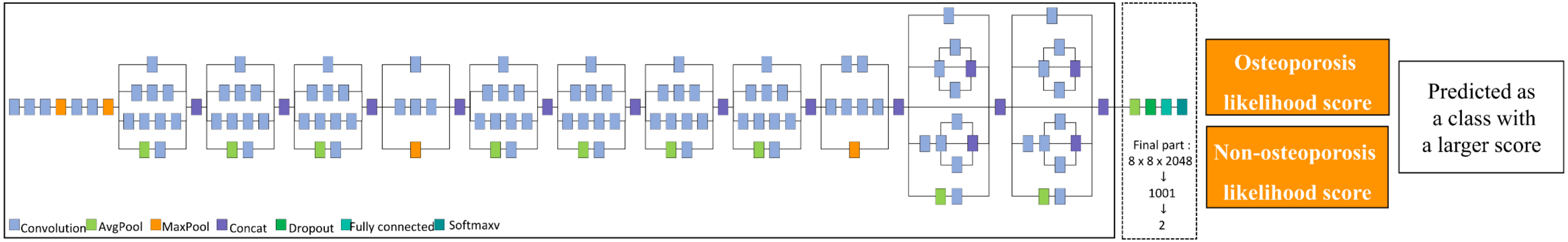
(d) Confusion matrix of the model trained with 512 × 512 pixel-sized image and batch size 15

| | | Predicted label | |
|---|---|---|---|
| | | Normal | Osteoporosis |
| True label | Normal | 652 | 49 |
| | Osteoporosis | 29 | 672 |

(e) Confusion matrix of the model trained with 1024 × 1024 pixel-sized image and batch size 15

| | | Predicted label | |
|---|---|---|---|
| | | Normal | Osteoporosis |
| True label | Normal | 639 | 62 |
| | Osteoporosis | 22 | 679 |

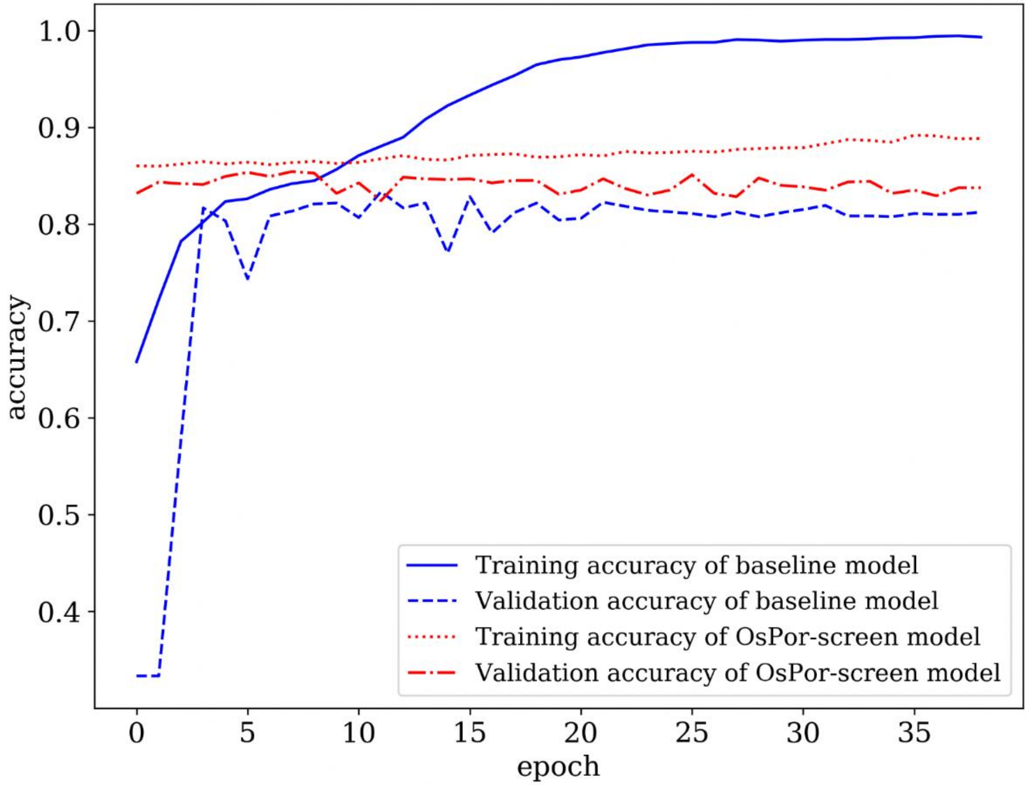
**Supplement figure 1. OsPor-screen model architecture.**


Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmaxv
Final part :
8 x 8 x 2048
↓
1001
↓
2
Osteoporosis likelihood score
Non-osteoporosis likelihood score
Predicted as a class with a larger score

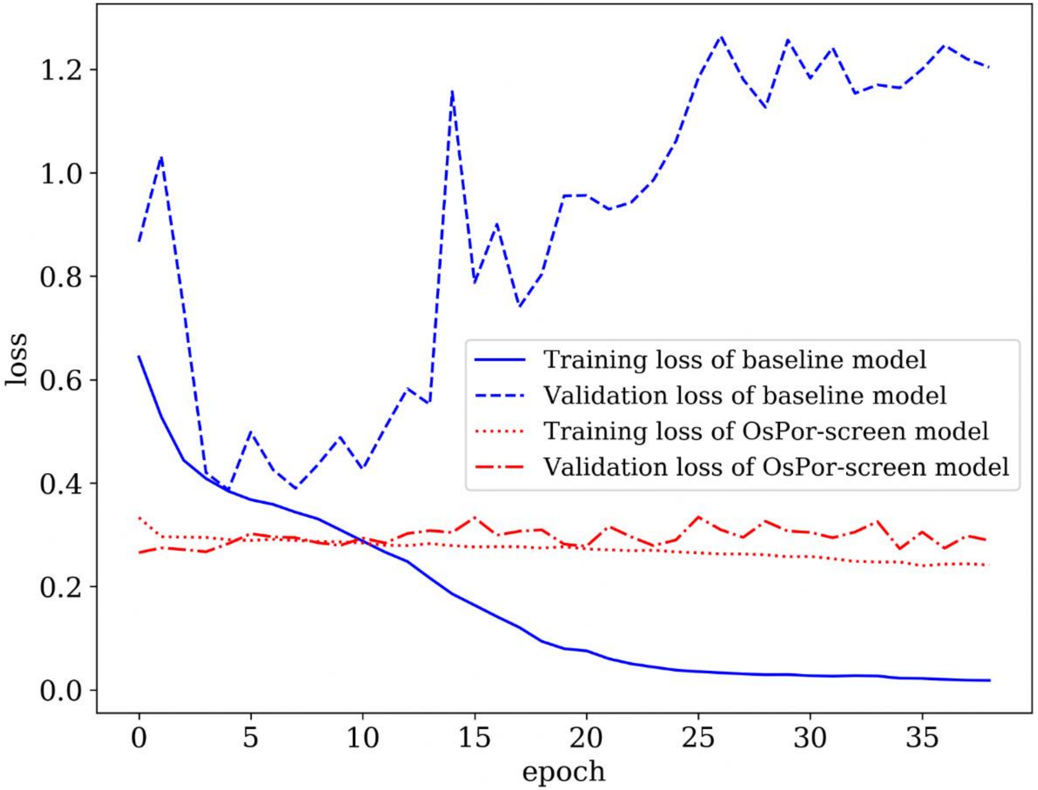
**Supplement figure 2. The training processes of the OsPor-screen and baseline models. (a) Accuracy and (b) loss.**

The training process shows the training performance in terms of accuracy and loss function along the training epochs for the baseline and OsPor-screen models.

(a)



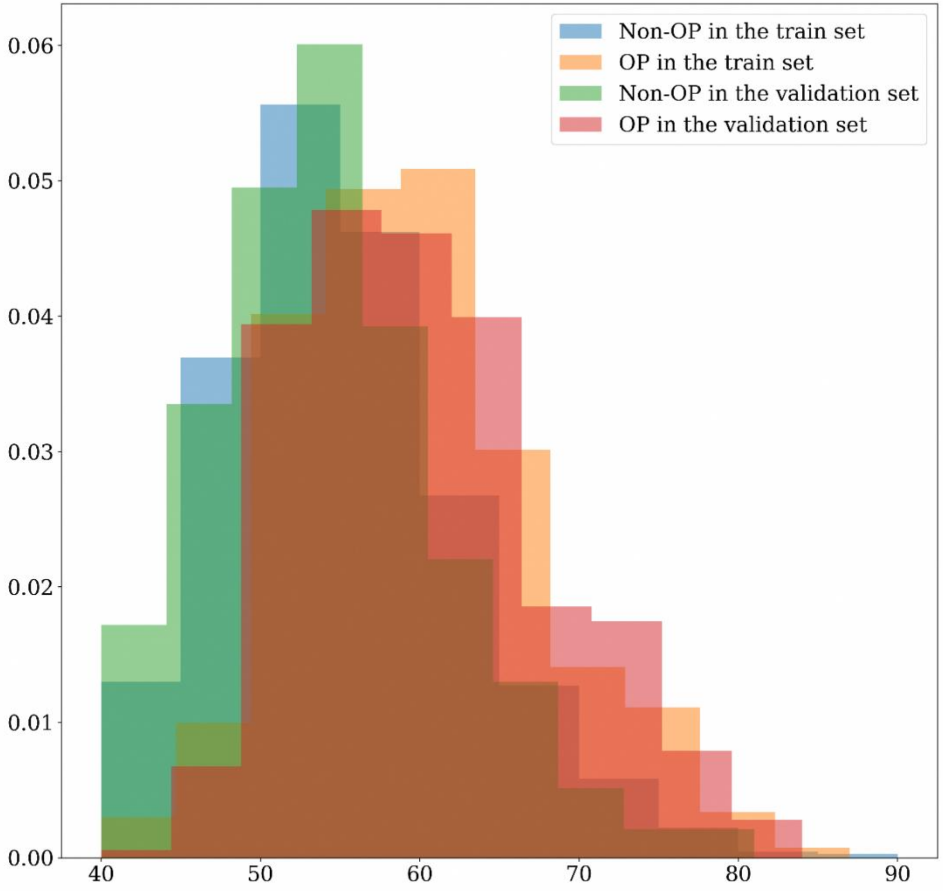
(b)



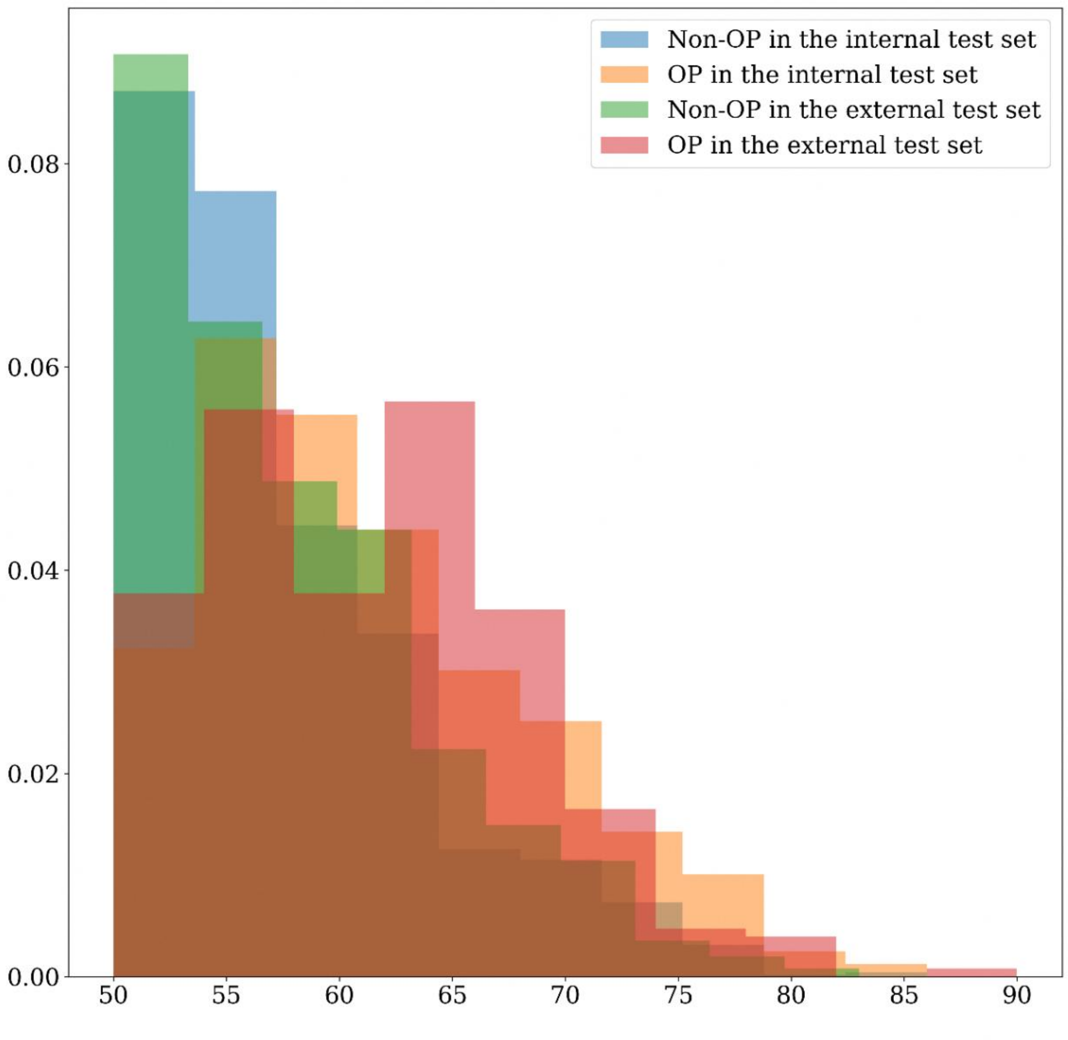
1.2
1.0
0.8
0.6
0.4
0.2
0.0
loss
0
5
10
15
20
25
30
35
epoch
Training loss of baseline model
Validation loss of baseline model
Training loss of OsPor-screen model
Validation loss of OsPor-screen model

**Supplement figure 3. Age distribution of osteoporosis and non-osteoporosis in each dataset. (a) the train and validation sets and (b) internal and external test sets.**

(a)

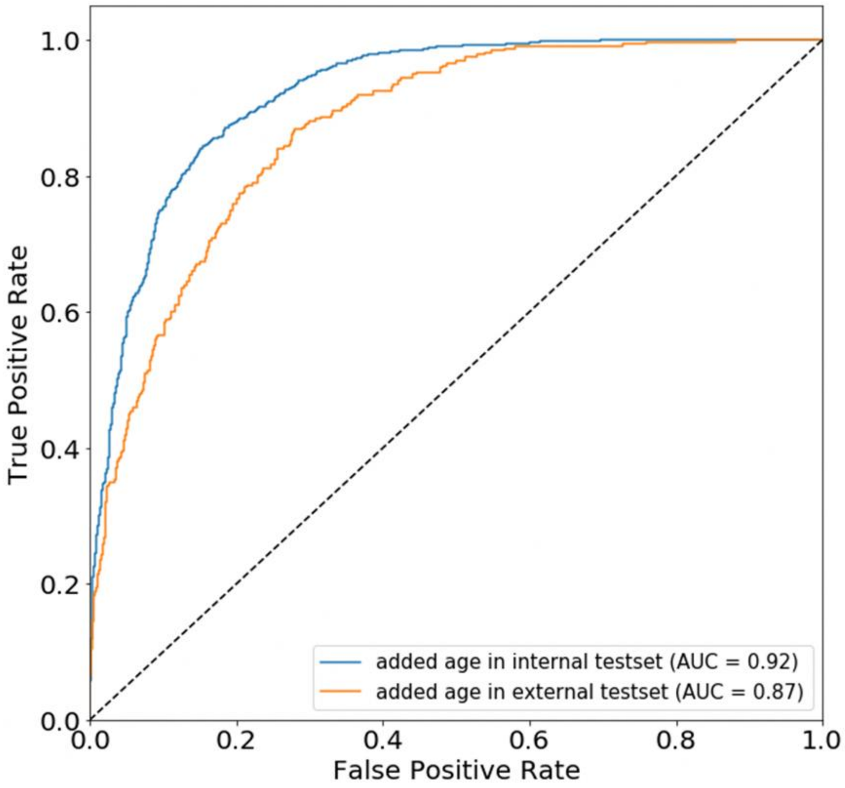


(b)



X-axis is age(year). This age distribution showed weighted distribution.

**Supplement figure 4. The ROC curves of CNN model trained with age in the internal and external test set**


True Positive Rate
False Positive Rate
added age in internal testset (AUC = 0.92)
added age in external testset (AUC = 0.87)