FDA Submission

Your Name: Sahika Betul Yayli

Name of your Device: DSS - Pneumonia Detector

Algorithm Description

1. General Information

Intended Use Statement:
Intented to use assist the medical doctors in the screening Pneumonia using chest X-Rays.

Indications for Use:

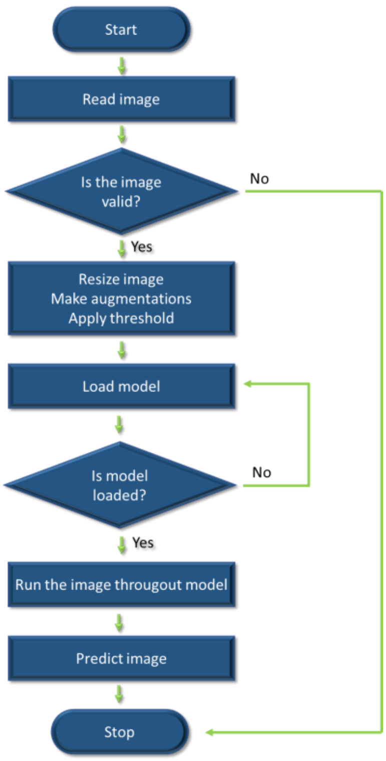
- Use in PA and AP chest X-Rays screening studies.
- Use in men and women with ages between 1 and 100.
- Modality should be only "DX".

Device Limitations:
Pneumonia can be coexist with other pectoral diseases such as atelectasis, infiltration, etc. It therefore may be difficult to detect coexisting pneumonia because combination list has wide variety.

Clinical Impact of Performance:
As for screening, we are more interested in reducing the number of patients sent home with mis-diagnosed as non-pneumonia. So we reduce the False Negatives. Hence, we maximize the recall was our goal.

By the way, precision increased when maximizing the recall. As our goal detecting pneumonia patients with minimum mi-sdiagnosis, it is okay to send a healthy patient to a new screening.

2. Algorithm Design and Function



DICOM Checking Steps:

1. The modality: Should be 'DX'
2. Body part: Should be 'CHEST'
3. The patient position: Should be 'PA' or 'AP'

Preprocessing Steps:

1. Take image pixel array from .dcm
2. Normalize the image by mean and standard deviation of image array
3. Resize the image to 224*224

CNN Architecture:

Pretrained VGG16 used with change in last layer. We added 1 flatten layer; then 2 relu activated dense layer and 1 sigmoid activated dense layer. So the model makes predictions for 2 classes.

3. Algorithm Training

Parameters:

*Types of augmentation used during training:

- horizontal_flip = True,
- vertical_flip = False,
- height_shift_range= 0.1,
- width_shift_range=0.1,
- rotation_range=20,
- shear_range = 0.2,
- zoom_range=0.2

* Batch size: 64

* Optimizer learning rate: Adam(lr=5e-5)

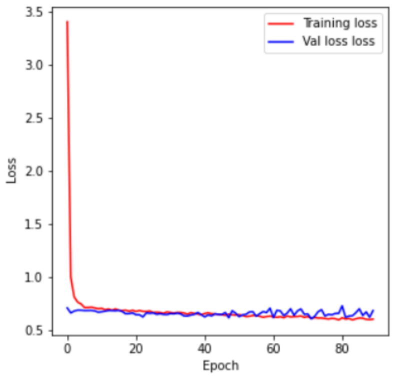
* Layers of pre-existing architecture that were frozen: VGG16 layers till dense layers.

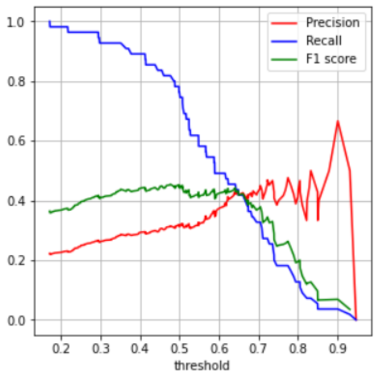
* Layers of pre-existing architecture that were fine-tuned: None

* Layers added to pre-existing architecture:
2 dense relu layers, 1 dense sigmoid layer for classification

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688
_________________________________________________________________
flatten_1 (Flatten)          (None, 25088)             0
_________________________________________________________________
dense_1 (Dense)              (None, 4096)              102764544
_________________________________________________________________
dropout_1 (Dropout)          (None, 4096)              0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               524416
_________________________________________________________________
dropout_2 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 129
=================================================================
Total params: 118,003,777
Trainable params: 103,289,089
Non-trainable params: 14,714,688
_________________________________________________________________
```

Final Threshold and Explanation:

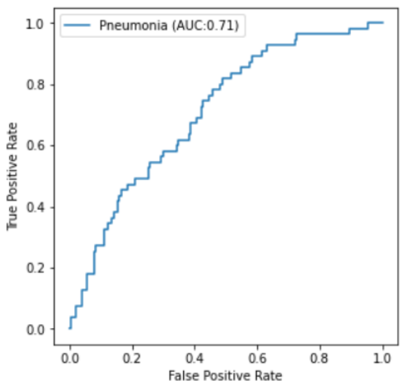


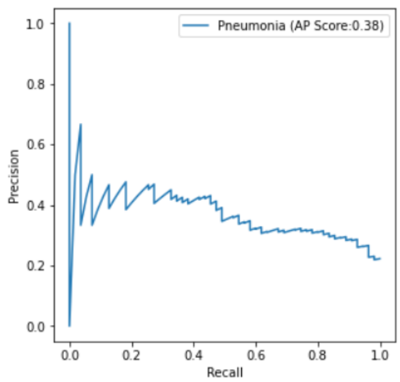


First graph shows training and validation loss. Values seems OK.

Second graph shows balance between threshold and precision, recall, f1-score. Our device for screening, so recall is more important than precision for determining the best threshold. According to recall, we find optimized threshold as 0,501. With this value, precision, recall and f1-score values are below:

Threshold is: 0.50120884
Precision is: 0.31297709923664124
Recall is: 0.7454545454545455
F1 Score is: 0.44086021505376344



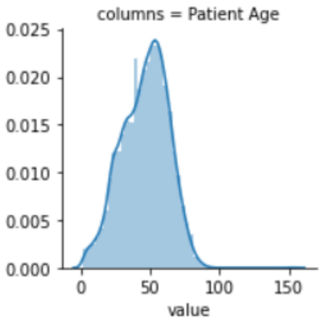


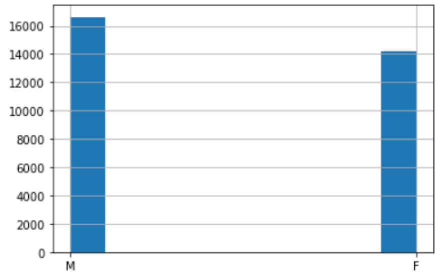
Third graph shows FP-TP relationship. AUC much below than 1.

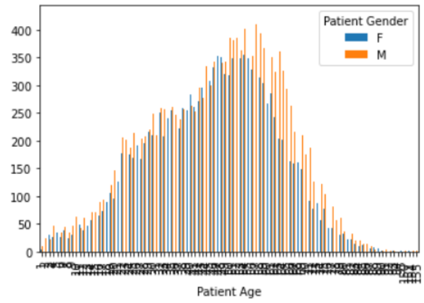
Fourth graph shows when recall increasing, precision decreasing.

4. Databases

Description of Training Dataset:



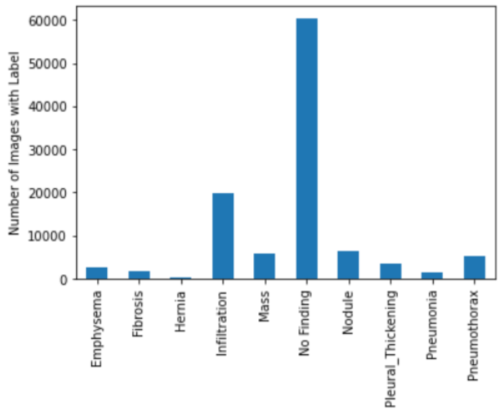


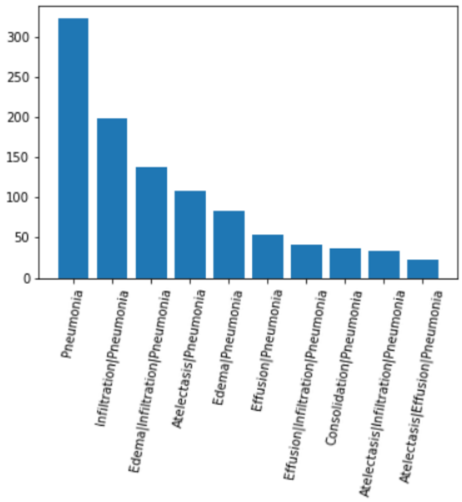


First graph shows age distribution. One vpatient had age 155 is removed from training set after analysis.

Second graph shows gender distribution.

Third graph shows patient age distribution by gender. There is no much difference between two groups.



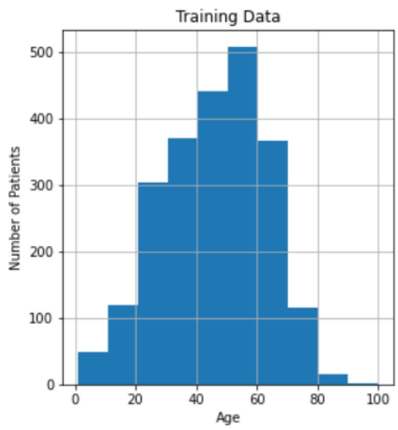


Fourth graph shows disease distribution on training dataset.

Fifth graph shows comorbid disease distribution with pneumonia.

Training dataset:

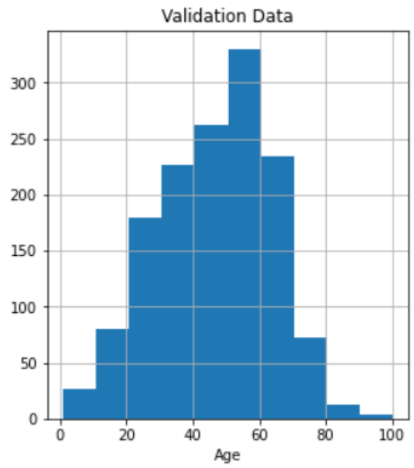
- We balaced pneumonia-nonpneumonia cases 1:1
- One patient with 155 years old deleted.
- Number of images:2289
- mean 46.178603
  std 18.689923
  min 2.000000
  25% 33.000000
  50% 48.000000
  75% 59.000000
  max 412.000000



Description of Validation Dataset:

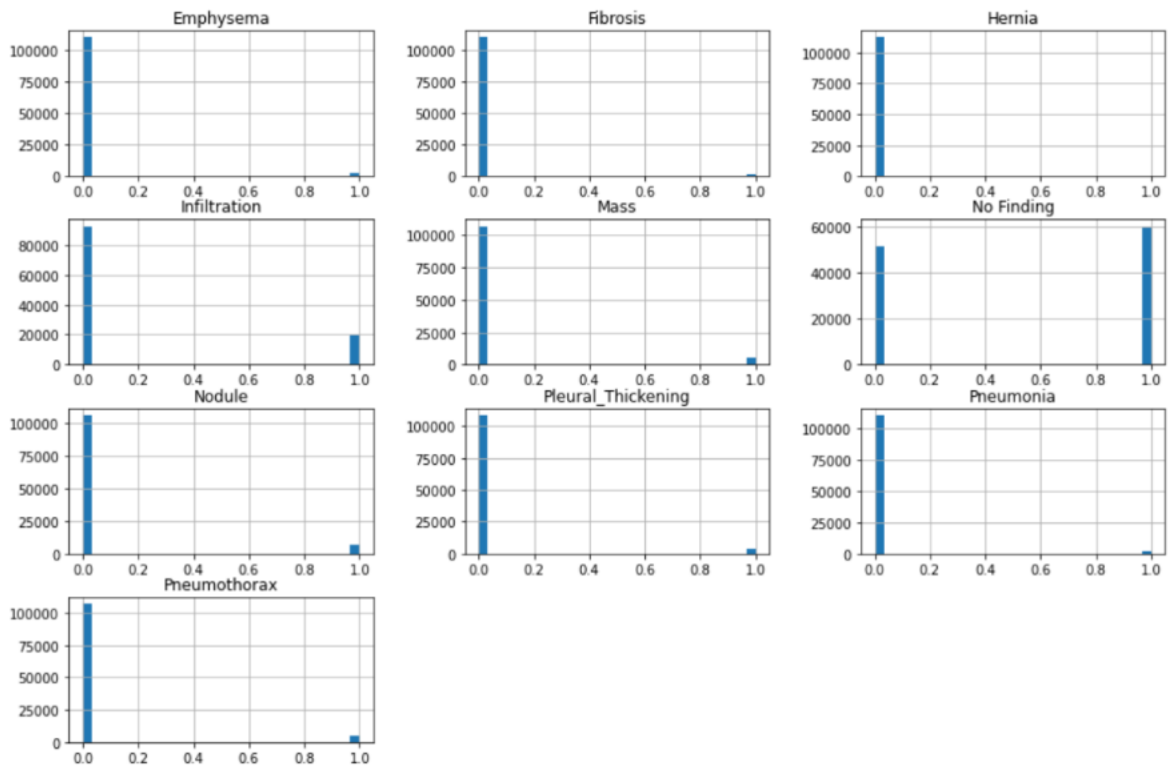
Validation dataset:

- Ratio between pneumonia-nonpneumonia cases 1:4
- Number of images:1430
- mean 46.855944
  std 19.641723
  min 2.000000
  25% 33.000000
  50% 49.000000
  75% 59.000000
  max 412.000000



5. Ground Truth

The NIH Chest X-ray Dataset is highly imbalanced in relation with patients with pneumonia.



NLP-derived labels from the NIH are sub-optimal since they are more general than only the case of Pneumonia. Oakden-Rayner published a article about false labeling with labeling on this dataset. (Oakden-Rayner L (2017) Exploring the ChestXray14 dataset problems. https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14- dataset-problems/)

Other lung diseases could seem like pneumonia. This dataset also includes this disease images. This could confuse when training.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
Ideally validation dataset have same distribution with my training dataset:
This should be a balanced dataset with population of men and women with ages between 1 and 100 with no previous history of Pneumonia.

Ideal dataset:

- Imaging modality: DX
- Body parts examined: CHEST
- Positions: ‘PA’ or ‘AP’
- Gender distribution: 1:1
- Age distribution: 1-100
- Presence of Atelectasis, Caridomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumothorax is acceptable.

Ground Truth Acquisition Methodology:
Gold standard would be bronchoscopy, pleural fluid culture or sputum test but this are expensive and time consuming. Also this are not routine for every pneumonia patient.

Silver standard, expert(radiology or pulmonologist) labeled X-Ray images would be better option.

Algorithm Performance Standard:
We focused on recall because of this device’s aim is screening the pneumonia. We tried to achieve recall value higher than 0.7 and we achieved it.
Average of radiologists f1-score was 0.387 according to the CheXNet paper. Our algorithm achieved 0.44. This score is better than radiologists’ achievement.