Name: Piush Vaish
Name of your Device: Pneumonia Classifier

Algorithm Description

General Information

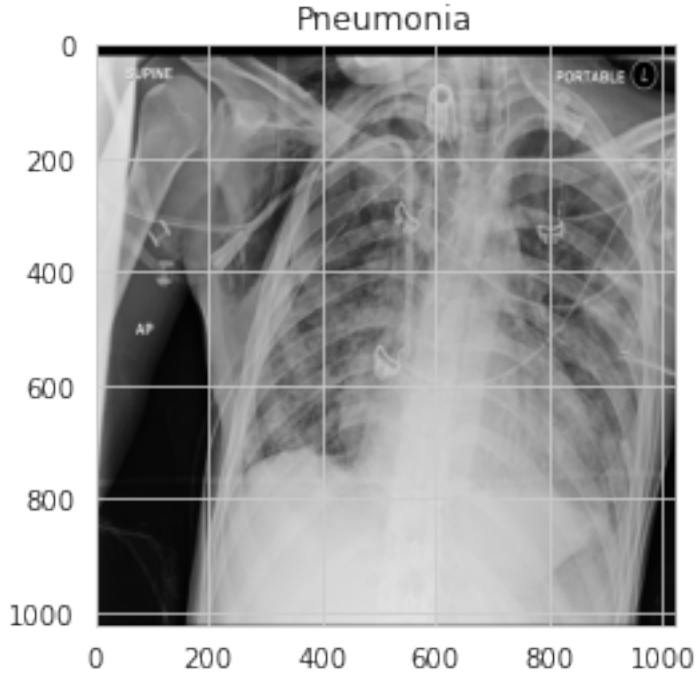
Intended Use Statement:
Assists radiologists to classify pneumonia or not using images

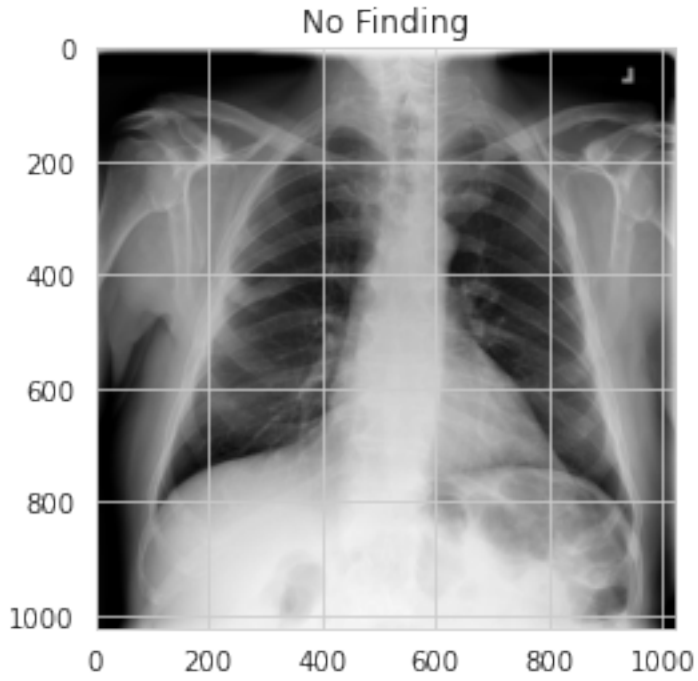
Indications for Use:
The trained algorithm should be integrated into the normal workflow of the diagnostic clinics to help radiologists in pre-screening of pneumonia from X-ray images. The X-ray images should be available in DICOM format, respecting the HIPAA rules. The target population's age is between 20 and 80 years old and for both males and females. The data is then sent through the algorithm and it first checks the following conditions for X-ray images:

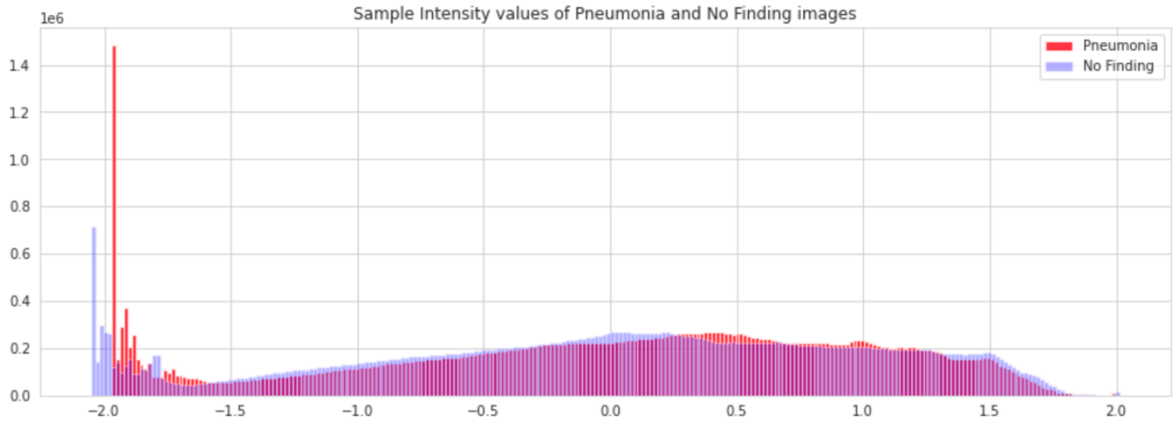
- Modality: DX (Digital Radiography)
- View Position: AP (Anterior/Posterior) or PA (Posterior/Anterior)
- Body part: Chest

If it satisfies the criteria, it makes the prediction. Once the prediction is complete, this data is sent to an expert radiologist/doctor, and then he/she will give the final verdict based on his/her independent diagnosis as well as the data supplied by the algorithm.

Device Limitations:
Deep learning model classifies pneumonia from X-ray images without any other diagnosis attached to the image in the meta data. It is also trained using GPU. However, there is no requirement for GPU during inference.

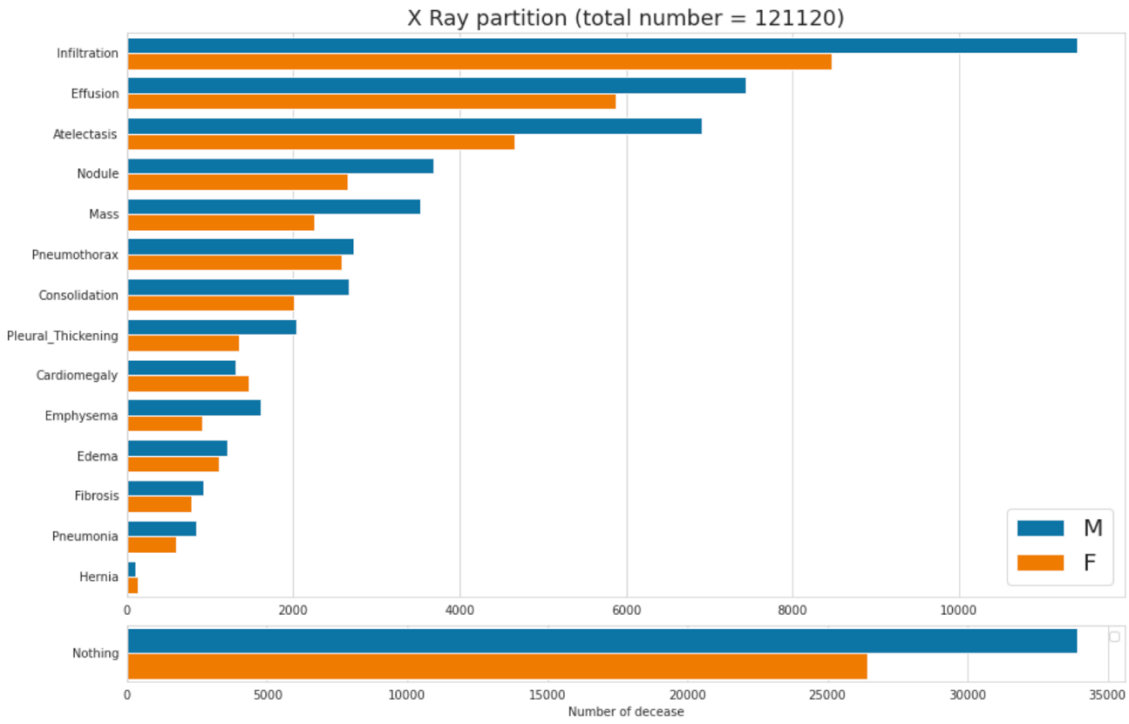






X-ray images can contain all kind of diagnosis:

- Effusion
- Aletectasis
- Infiltration Some others seem to have favorite couple:
- Cardiomegaly with Effusion
- Emphysema with Pneumothorax
- Nodule with Infiltration
- Edema with Infiltration
- Fibrosis with Infiltration
- Pneumonia with Infiltration



Clinical Impact of Performance:

The algorithm is trained for Pneumonia only and is better at classifying healthy patients than with pneumonia. The confusion matrix is as follows on the validation dataset:

```
Confusion Matrix :
[[131  65]
 [ 19  41]]
```

- True Negative: 131
- False Positive: 65
- False Negative: 19
- True Positive: 41

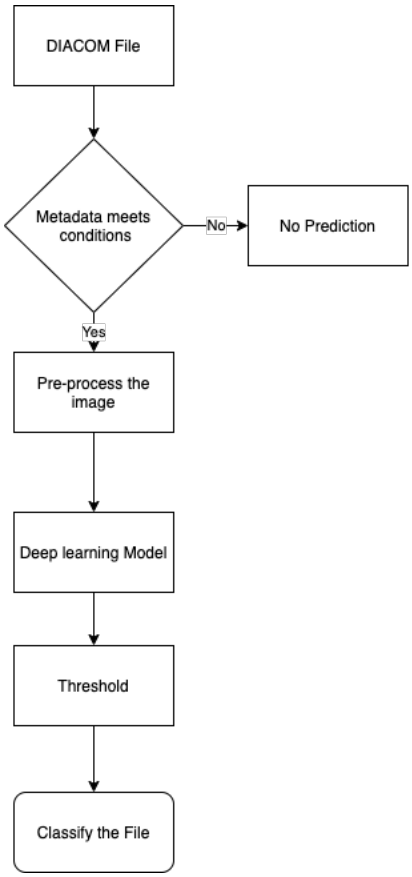
```
Report :
              precision    recall  f1-score   support

         0.0       0.87      0.67      0.76       196
         1.0       0.39      0.68      0.49        60

    accuracy                           0.67       256
   macro avg       0.63      0.68      0.63       256
weighted avg       0.76      0.67      0.70       256
```

The algorithm has both false positive and false negatives because it is not 100% accurate as shown in the classification report. A false positive result erroneously labels a person having pneumonia with consequences including unnecessary hospital visits and emotional stress. False negative results are more consequential, because person with pneumonia may not get treatment. Therefore, the algorithm should not be used in the absence of expert radiologist or a doctor. It is also worth mentioning that healthy cases should still be reviewed by the radiologist as this is supposed to be a supplementary tool.

Algorithm Design and Function



Flowchart

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model_2 (Model)              (None, 7, 7, 512)         14714688
_________________________________________________________________
flatten_2 (Flatten)          (None, 25088)             0
_________________________________________________________________
dense_5 (Dense)              (None, 1024)              25691136
_________________________________________________________________
dropout_4 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_6 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout_5 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_7 (Dense)              (None, 256)               131328
_________________________________________________________________
dropout_6 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_8 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_7 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_9 (Dense)              (None, 1)                 129
=================================================================
Total params: 41,094,977
Trainable params: 28,740,097
Non-trainable params: 12,354,880
_________________________________________________________________
```

DICOM Checking Steps:
The algorithm performs the following checks on the DICOM image:

- Check Examined Body Part is 'CHEST'
- Check Patient Position is either 'PA' (Posterior/Anterior) or 'AP' (Anterior/Posterior) Check Modality is 'DX' (Digital Radiography)

Preprocessing Steps:
We use the off-the-shelf ImageDataGenerator class from the Keras framework, which allows us to build a "generator" for images specified in a training data This class also provides support for basic data augmentation such as random horizontal flipping of images. We also use the generator to transform the values in each batch so that their mean is 0 and their standard deviation is 1. This will facilitate model training by standardizing the input distribution. We also convert RGB to Grayscale (if needed) and re-sizes the image to 244 x 244 (as required by the model).
We only normalize the images in validation dataset to resemble the real world data as closely as possible. The algorithm is evaluated using this dataset.

CNN Architecture:

The model uses transfer learning by taking the first layers of a VGG16 model trained on ImageNet data (classifying color images of dogs, airplanes, cats, ...) and retrains it on grayscale images of chests. The network output is a single probability value for binary classification.

Algorithm Training

Parameters:

Types of augmentation used during training

- rescale=1. / 255.0,
- samplewise_center=True,
- samplewise_std_normalization= True,
- shear_range=0.1,
- zoom_range=0.15,
- rotation_range=5,
- width_shift_range=0.1,
- height_shift_range=0.05,
- horizontal_flip=True,
- vertical_flip = False,
- fill_mode = 'reflect'

Batch size

- Training: 32
- Validation: 256

Optimizer learning rate

- 1e-4

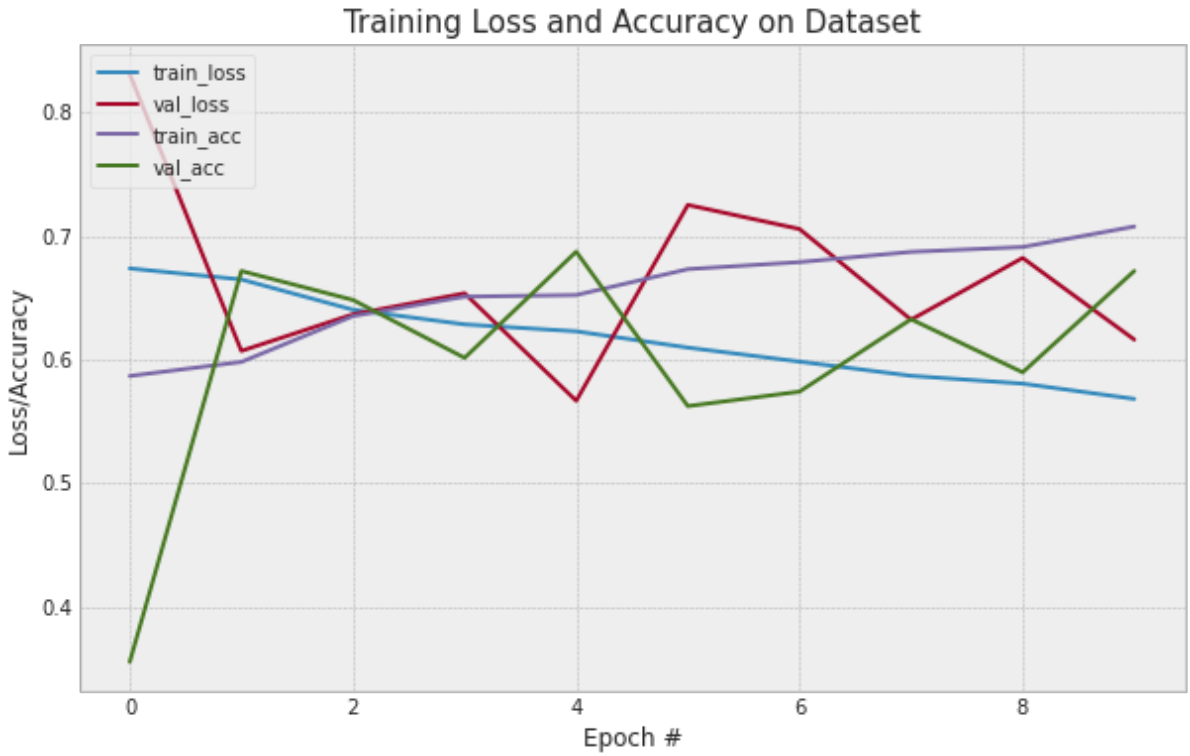
Layers of pre-existing architecture that were frozen

- 16 Layers

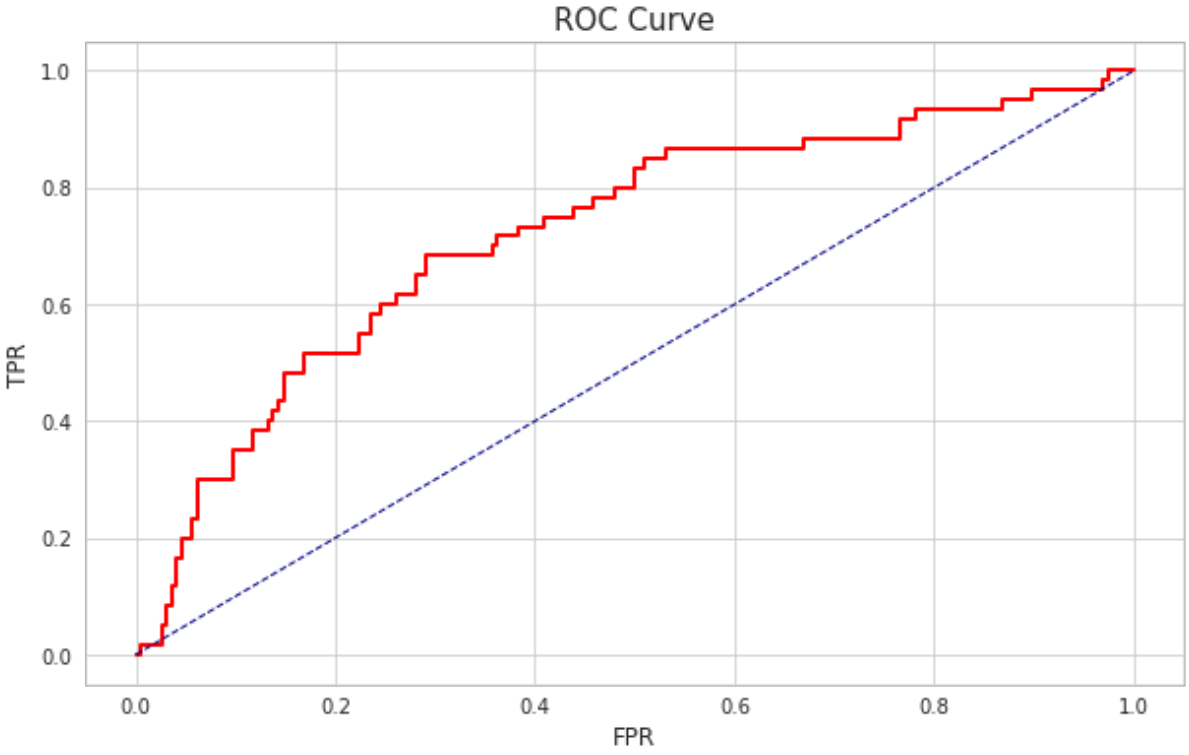
Layers added to pre-existing architecture

- 4 Layers

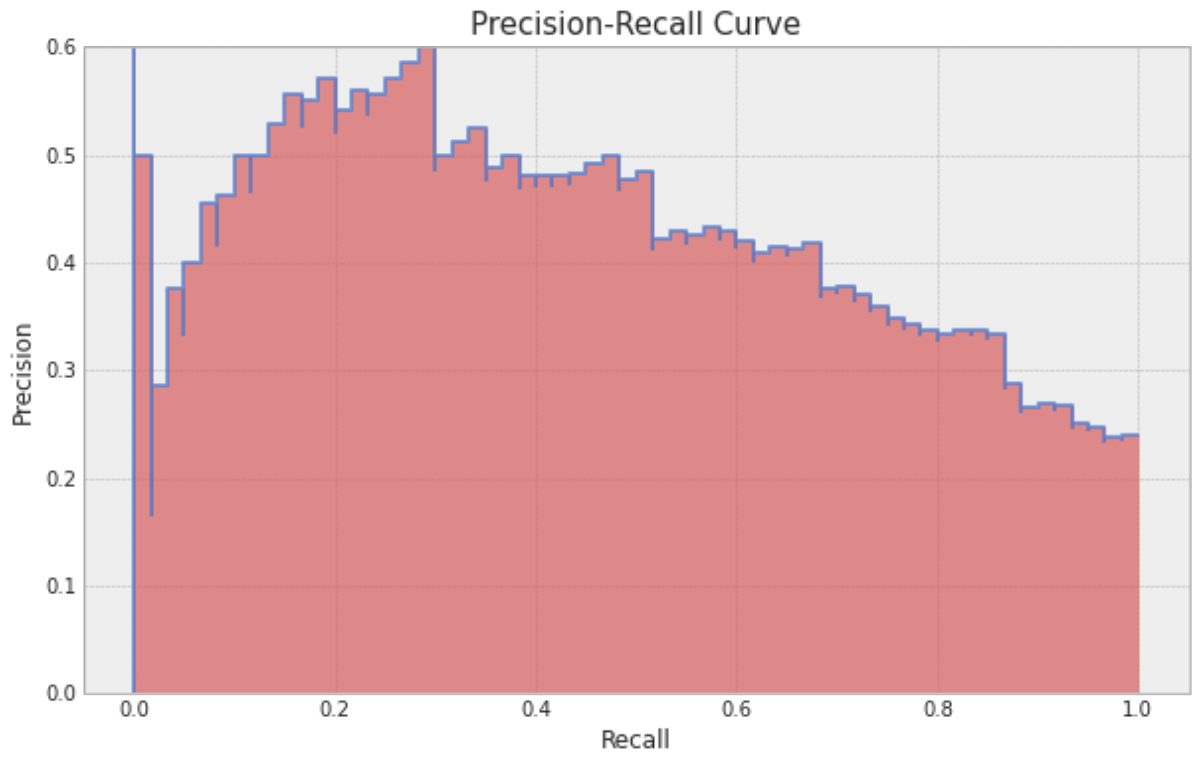
Insert algorithm training performance visualization



The plot above shows that the model is overfitting because the training loss is going down while the validation loss is going more after 8 epochs.



Insert P-R curve



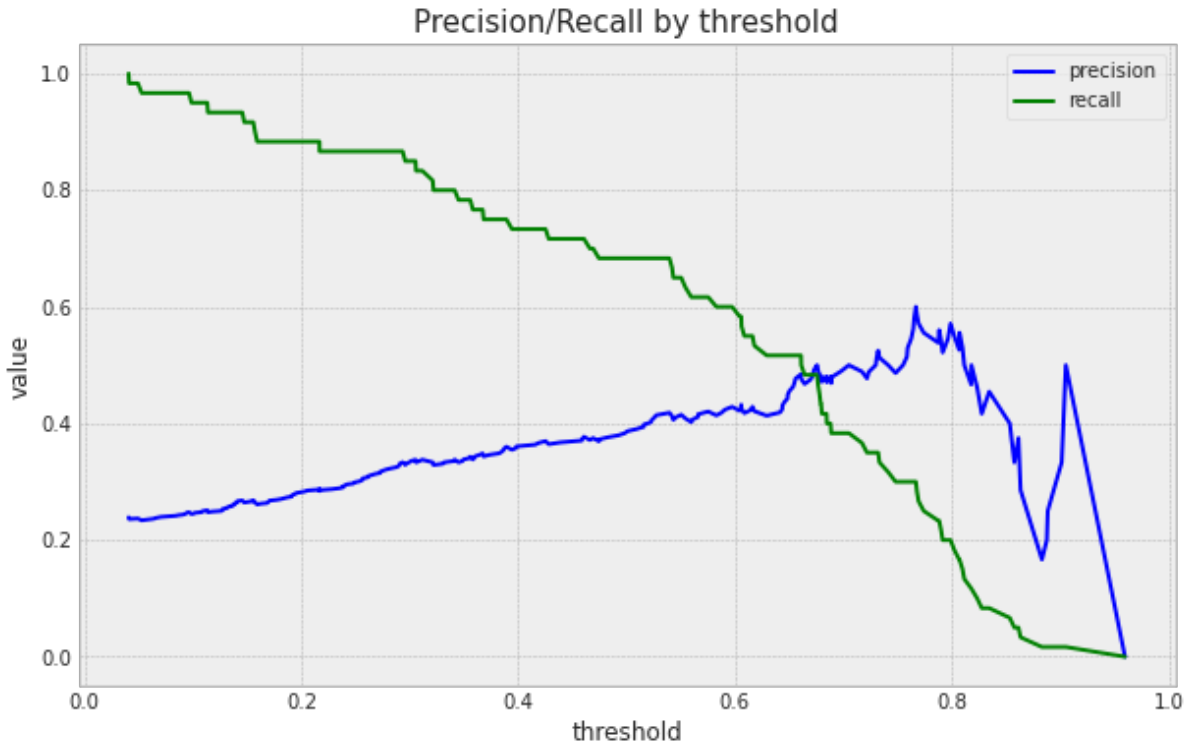
"A precision-recall curve shows the relationship between precision (= positive predictive value) and recall (= sensitivity) for every possible cut-off. The PRC is a graph with:

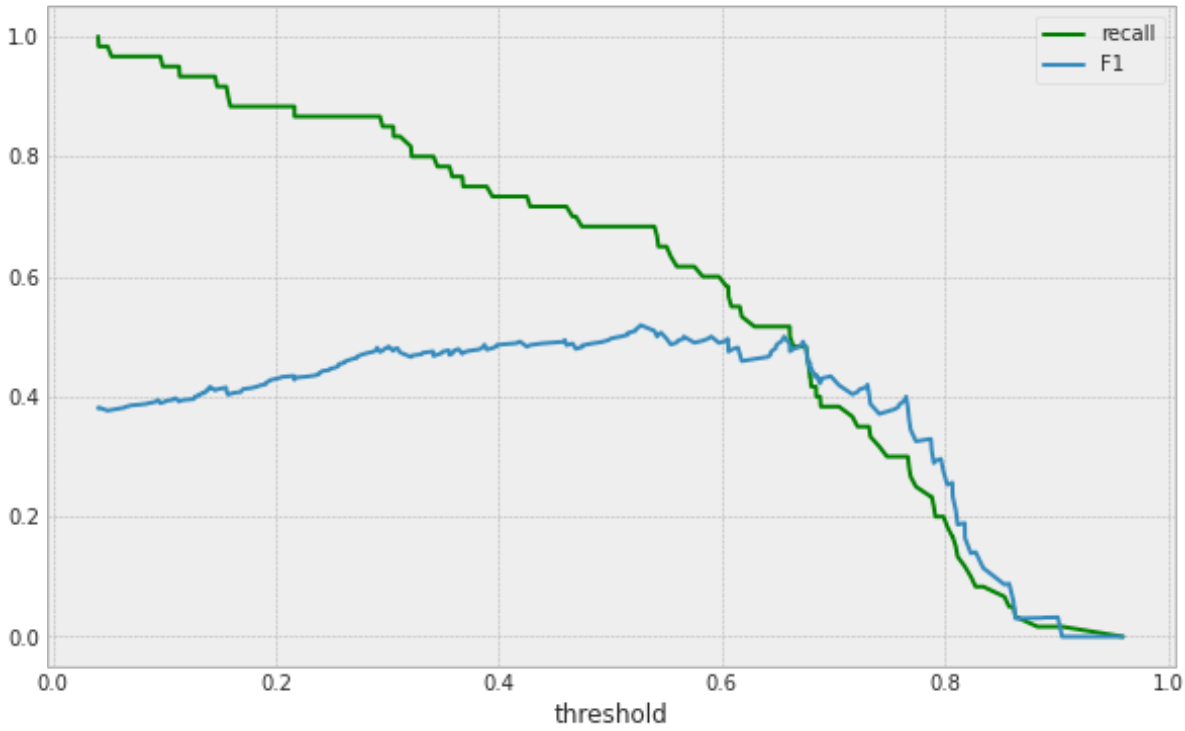
• The x-axis showing recall (= sensitivity = TP / (TP + FN))
• The y-axis showing precision (= positive predictive value = TP / (TP + FP))

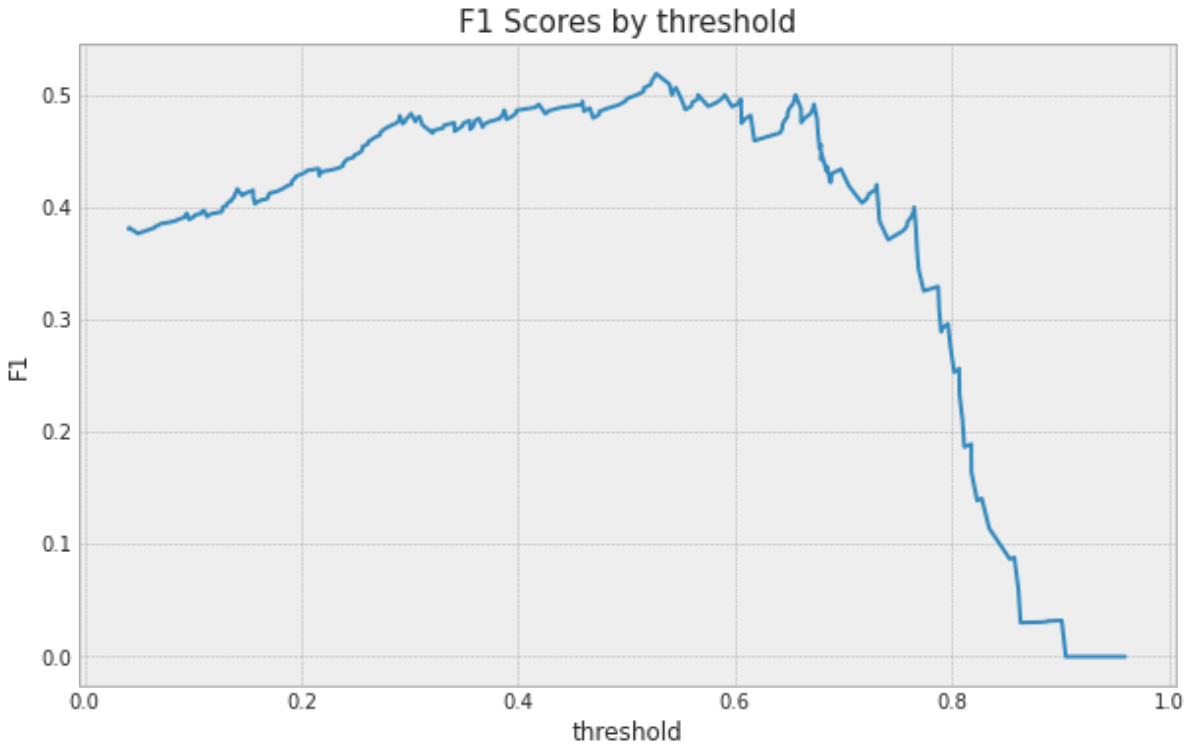
Thus every point on the PRC represents a chosen cut-off even. The precision-recall curves are not impacted by the addition of patients without disease and with low test results. "

The above plot shows sensitivity (= recall) to identify the vast majority of persons with a disease as having the disease. It also show positive predictive value (= precision) to identify a person actually having the disease. Thus, it discriminates persons with pneumonia from those without pneumonia.

Final Threshold and Explanation:


Precision/Recall by threshold
precision
recall
value
threshold
0.0
0.2
0.4
0.6
0.8
1.0


recall
F1
threshold
0.0
0.2
0.4
0.6
0.8
1.0



The maximum F1 score for the model is 0.5189 and it is achieved with a very high threshold value of 0.539.

The model's best F1-score is 0.5189. This is better than the F1 score from the paper (CheXNet: Radiologist-Level Pneumonia Detection). This is also better score than the radiologist's average as discussed in the same paper.

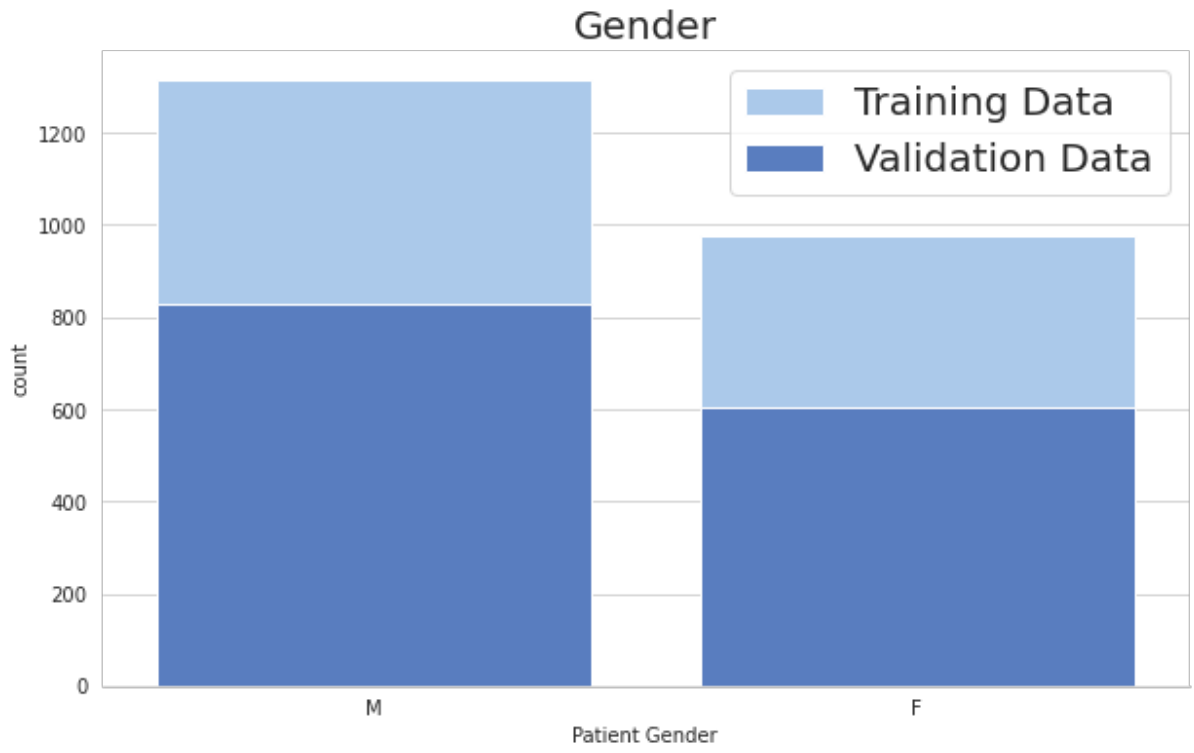
	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

However, F1 Score for the algorithm should be taken as a rough guide to the performance because the labels for training the algorithm were generated using Natural language processing and are expected to be >90% accurate.

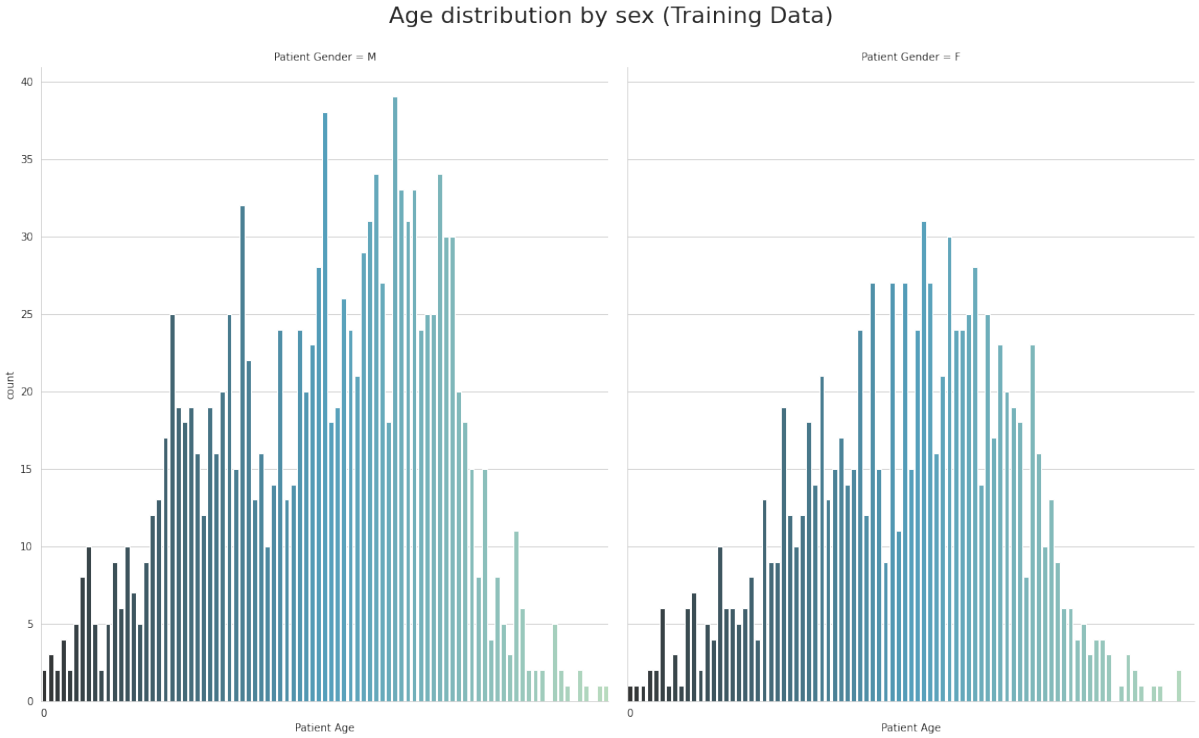
Databases

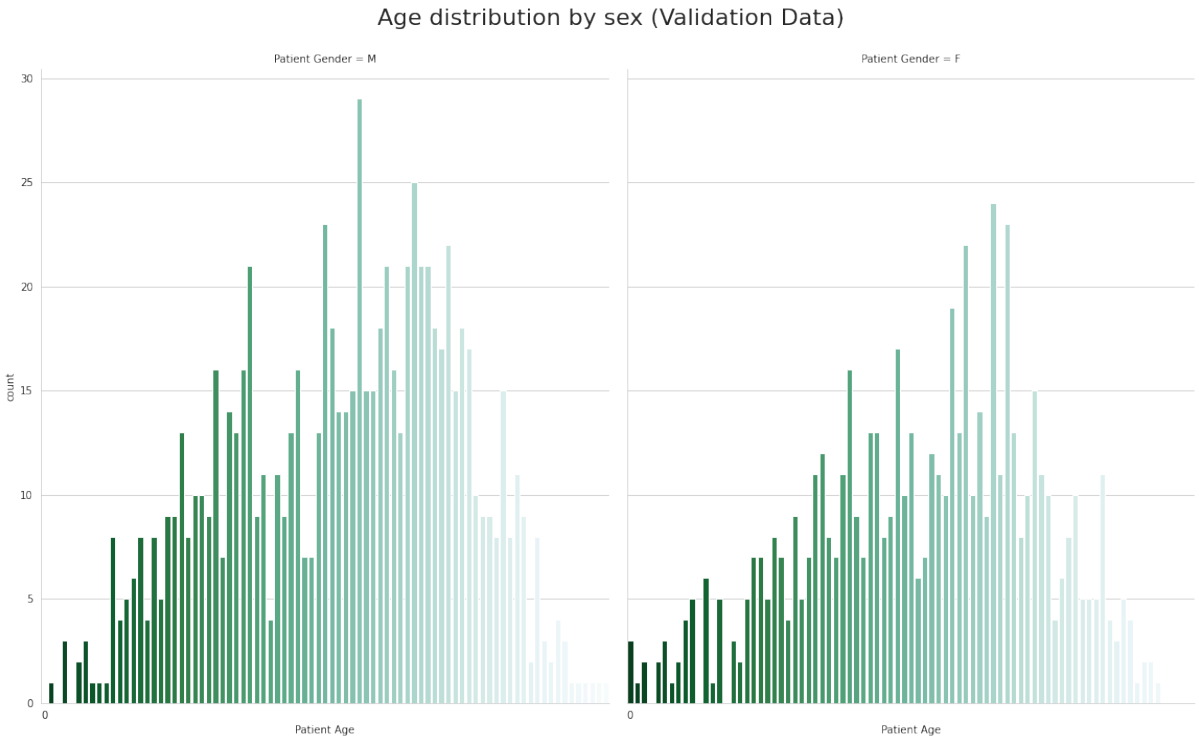
Description of Training Dataset:
Training dataset consisted of 2290 chest X-ray images, with a 50/50 split between positive and negative cases.
Description of Validation Dataset:
Validation dataset consisted of 1430 chest xray images, with 20/80 split

Gender distribution of patients in both training and validation datasets
There are more males than females in both the datasets. There are 1314 males and 976 females in training data and 828 males and 602 females in the validation dataset.



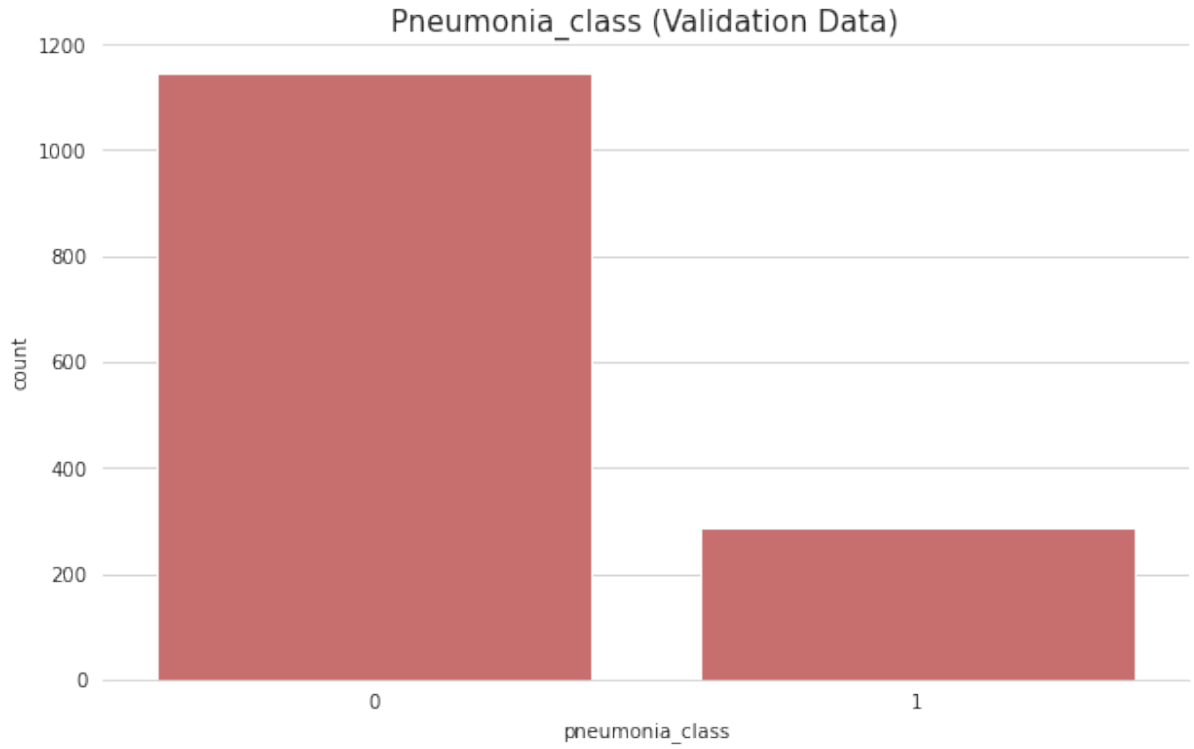
Age distribution of patients in both training and validation datasets
The age distribution shows that in both of the datasets the females have a normal distribution while males age is right skewed.





Prevalence of diseases in both training and validation datasets

There are 1145 records of both types of classes in the training data with 1144 records of healthy patients and 286 with pneumonia.



Ground Truth

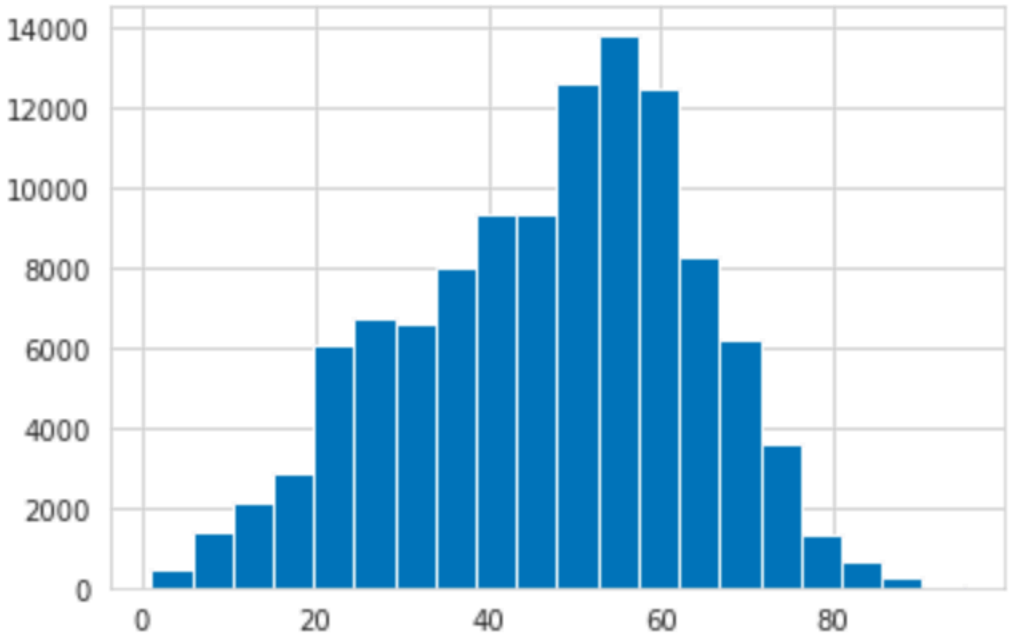
The NIH Clinical Centre recently released over 100,000 anonymized chest x-ray images and their corresponding data to the scientific community. ChestX-ray8 dataset contains 108,948 frontal-view X-ray images of 32,717 unique patients. Each image in the data set contains multiple text-mined labels identifying 14 different pathological conditions. These in turn can be used by physicians to diagnose 8 different diseases. This dataset has been annotated by consensus among four different radiologists for 5 of our 14 pathologies:

- Consolidation
- Edema
- Effusion
- Cardiomegaly
- Atelectasis

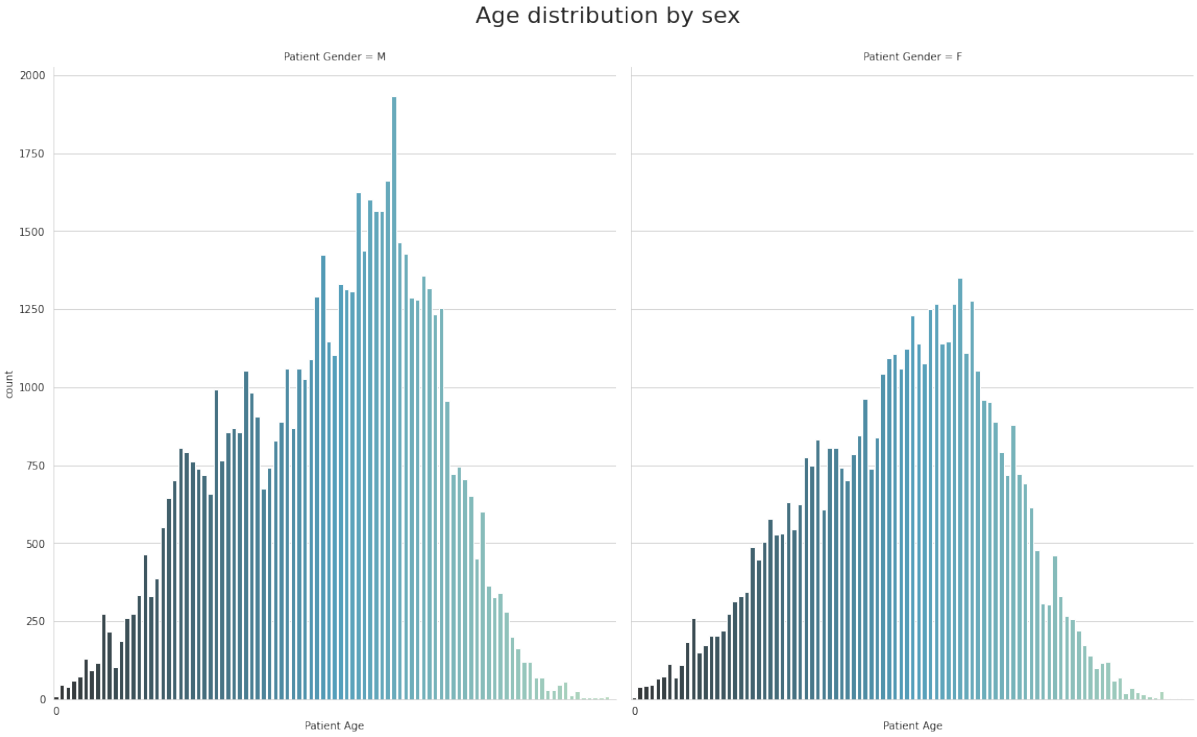
The release will allow researchers across the country and around the world to freely access the datasets and increase their ability to teach computers how to detect and diagnose disease. Ultimately, this artificial intelligence mechanism can lead to clinicians making better diagnostic decisions for patients. This data is used to develop a single model that provide binary classification predictions for pneumonia. In other words it will predict 'positive' or 'negative' for pneumonia.
The images are labelled using Natural Language Processing. The text is mined to classify disease from the associated radiological reports. The labels are expected to be >90% accurate and suitable for weakly-supervised learning. The original radiology reports are also not publicly available (Wang et al).
Therefore, the algorithm can learn to classify some images wrongly. It is also tested on the data from the same dataset. Hence, the metrics to evaluate the algorithm are also not correct.

FDA Validation Plan

Patient Population Description for FDA Validation Dataset
The following population subset is to be used for the FDA Validation Dataset:
Both men and women between the ages of 2 and 90 years old.



The data should have same ratios of both males and females.



The X-ray images should be available in DICOM format, respecting the HIPAA rules. The target population's age is between 20 and 80 years old and for both males and females. The data is then sent through the algorithm and it first checks the following conditions for X-ray images:

- Modality: DX (Digital Radiography)
- Position: AP (Anterior/Posterior) or PA (Posterior/Anterior)
- Body part: Chest

The data should include X-ray images with only pneumonia and should not contain any other common thoracic pathologies such as :

- Atelectasis
- Consolidation
- Infiltration
- Pneumothorax
- Edema
- Emphysema
- Fibrosis
  Effusion

Ground Truth Acquisition Methodology:
Doctors review medical history, perform a physical exam, and order diagnostic tests to diagnose pneumonia. The golden standard for obtaining ground truth would be to perform one of these diagnostic tests:

- Blood tests such as a complete blood count (CBC) to see if your immune system is actively fighting an infection.
- Sputum test.
- Chest computed tomography (CT) scan to see how much of your lungs is affected.
- Pleural fluid culture

The use of these tests along with X-ray images will allow the radiologists to diagnosis pneumonia. An average of radiologist's diagnosis can be used as a 'silver standard'.

Algorithm Performance Standard:

The model's best F1-score is 0.5189. It is achieved with a very high threshold value of 0.539. This is better than the F1 score from the paper (CheXNet: Radiologist-Level Pneumonia Detection). This is also better score than the radiologist's average as discussed in the same paper. However, F1 Score for the algorithm should be taken as a rough guide to the performance because the labels for training the algorithm were generated using Natural language processing and are expected to be >90% accurate.