FDA Submission

Your Name: Jayanthi Suryanarayana

Name of your Device: Pneumonia Detector

Algorithm Description

1. General Information

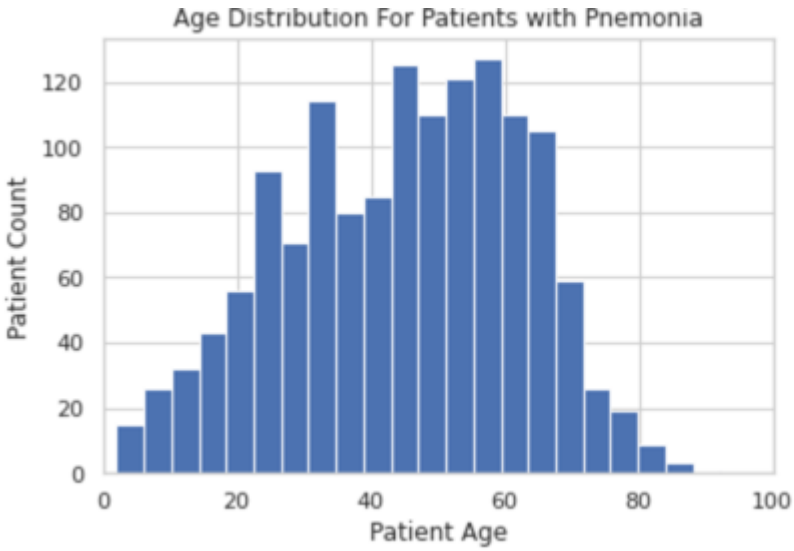
Intended Use Statement: Help Radiologists with Pneumonia detection in Chext X-Rays

Indications for Use:

X-Ray image properties:

Body part: Chest Position: AP (Anterior/Posterior) or PA (Posterior/Anterior) Modality: DX (Digital Radiography)The algorithm should be integrated to the workflow of the diagnostic clinics to assists radiologists to be effective

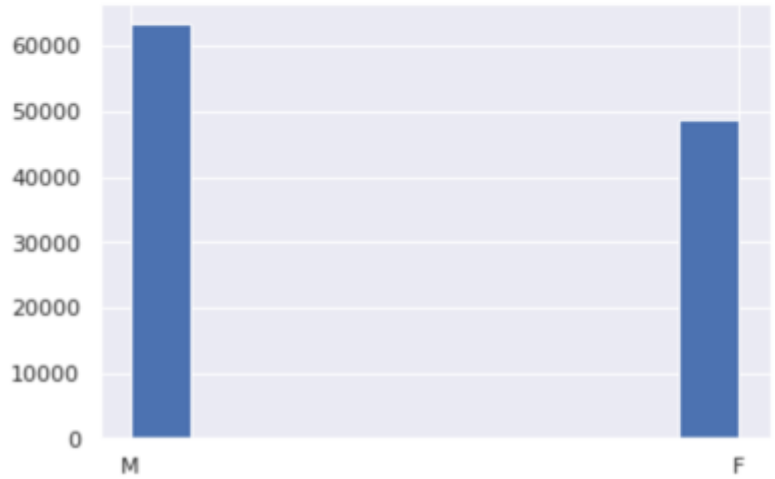
Age: As we can see from the EDA, the data set has more population in 30-75 age and should be best used for that population. and gender



Gender: The data set has reasonable gender distribution and can be used on any gender

```
In [13]: #plt.figure(figsize=(6,6))
         #all_xray_df['Patient Gender'].value_counts().plot(kind='bar')
         plt.hist(all_xray_df['Patient Gender'])

Out[13]: (array([63340.,     0.,     0.,     0.,     0.,     0.,     0.,     0.,
                 0., 48780.]),
          array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
          <a list of 10 Patch objects>)
```



Device Limitations: Inferences did not take much time and thus does not need any specific devices to run. The age and gender distributions do not specifically call out any limitations. It is not recommended to use for prioritization workflow

Clinical Impact of Performance: Checking for Pneumonia cannot be considered screening like mamogram for brest cancer. It is more for diagnostic that means false positives should be minimized. The threshold accounts for this by choosing higher threshold. Should be used to assist clinicians and not as a final diagnostic.

2. Algorithm Design and Function

Use Transfer learning. Use image net and fine tune it as described in the project

DICOM Checking Steps:

1. View the images
2. Preprocess sort of do one hot encoding for the diseases

Preprocessing Steps:

1. while splitting the training and validation, make sure the distribution is similar

2. also make sure the data is not imbalanced by imputing some data in training set
3. Augument the training set.

CNN Architecture:

1. take imagenet as basis (VGG16 and freeze initial 17 layers)
2. modify and below is the model summary Model: "sequential_1"

Layer (type) Output Shape Param

model_1 (Model) (None, 7, 7, 512) 14714688

---

flatten_1 (Flatten) (None, 25088) 0

---

dropout_1 (Dropout) (None, 25088) 0

---

dense_1 (Dense) (None, 1024) 25691136

---

dropout_2 (Dropout) (None, 1024) 0

---

dense_2 (Dense) (None, 512) 524800

---

dropout_3 (Dropout) (None, 512) 0

---

dense_3 (Dense) (None, 256) 131328

---

dense_4 (Dense) (None, 1) 257

Total params: 41,062,209 Trainable params: 28,707,329 Non-trainable params: 12,354,880

3. Algorithm Training

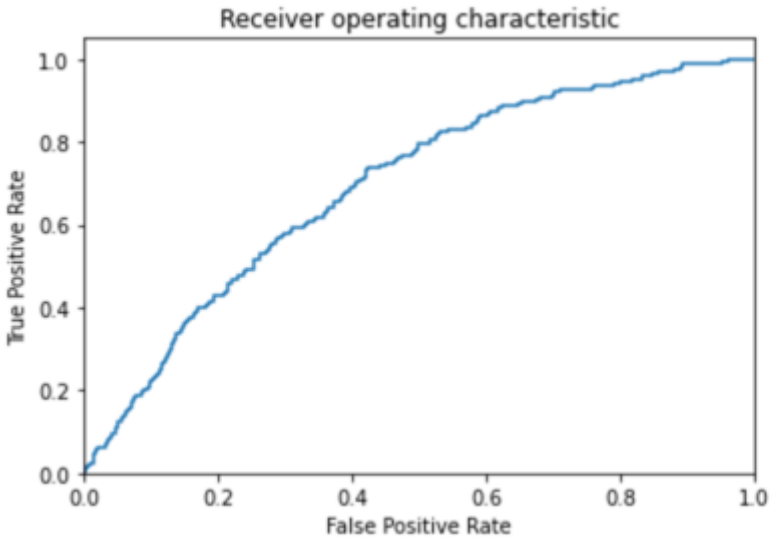
Parameters:

- Types of augmentation used during training - ImageGenerator
- Batch size - 32
- Optimizer learning rate - lr=1e-4

- Layers of pre-existing architecture that were frozen - VGG16 initial 17 layers were frozen
- Layers of pre-existing architecture that were fine-tuned
- Layers added to pre-existing architecture : Three dense layers and drop out layers in between were added as shown above in the model summary

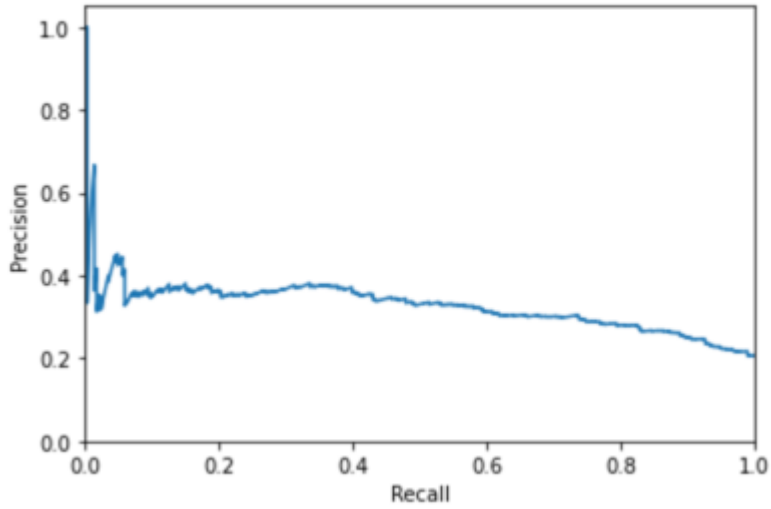
Training and Validation Loss:

```
<matplotlib.legend.Legend at 0x7f207c08c790>
```



ROC:

```
[<matplotlib.lines.Line2D at 0x7f51a0dcda50>]
```



Recall Precision:

Final Threshold and Explanation: F1 score was calculated and from there best threshold was selected Best Threshold=0.419119, F-Score=0.430

** Classification report precision recall f1-score support

0.0	0.89	0.58	0.70	1144
1.0	0.30	0.73	0.43	286
accuracy			0.61	1430

macro avg 0.60 0.65 0.57 1430 weighted avg 0.78 0.61 0.65 1430

4. Databases

(For the below, include visualizations as they are useful and relevant)

Description of Training Dataset: After EDA, a training/validation sets were created with 80/20 split. The train set images were imputed to account for imbalance After that, the training set images were normalized and augumented.

Description of Validation Dataset: From the original set took 20% of the rows. A random sample of non-pneumonia data that's 4 times as big as the pneumonia sample was added to account for imbalance of data

The validation set was normalized but not augumented.

5. Ground Truth

Silver, labels done from NLP. NLP extraction is not 100% accurate and will lead to some bias in the label. This is cost effective, but does not have the accuracy of a radiologist actually labelling.

The data is taken from a larger xray dataset, with disease labels created using Natural Language Processing (NLP) mining the associated radiological reports. The labels include 14 common pathologies :

Atelectasis Consolidation Infiltration Pneumothorax Edema Emphysema Fibrosis Effusion Pneumonia Pleural thickening Cardiomegaly Nodule Mass Hernia

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset: For Pneumonia detection, all genders mostly in the age range of 30-70 is suitable.

Ground Truth Acquisition Methodology: The golden standard for obtaining ground truth would be to perform one of these tests see this Mayo Clinic Link:

For this , silver standard which is labelling by NLP was used.

Algorithm Performance Standard: It can help the clinicians, it can be easily integrated with the device.

In terms of Clinical performance, the algorithm's performance can be measured by calculating F1 score against 'silver standard' ground truth as described above. Thealgorithm's F1 score should exceed 0.387 which is an average F1 score taken over three human radiologists, as given in CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rayswith Deep Learning, where a similar method is used to compare device's F1 score to average F1 score over three radiologists. THis algorithm has an F1 score of .43. So it is comparable to average F1 score over three human radiologists.