FDA Submission

Your Name: Deep Learning Researcher

Name of your Device: PneumaTron-X

Algorithm Description

1. General Information

Intended Use Statement:
To assist radiologists in classification of presence or absence of pneumonia

Indications for Use:
This device is a deep neural network algorithm which should be used for classification of presence or absence of pneumonia in the age group of 1-100 years irrespective of any gender or racial bias.

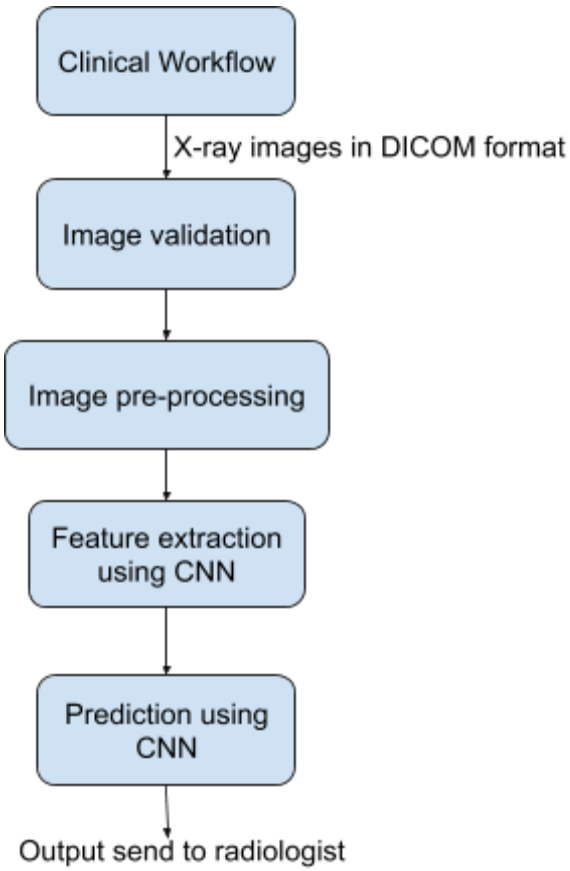
Device Limitations:
System requirement at the least is a CPU with 8GB RAM and 50 GB memory. The prediction from the device can be used by Radiologists for their assessment, but it is not 100% accurate. The accuracy is ~78% which is still high enough to assist Radiologists to provide timely treatment to the needed patients. Due to its performance based limitations it should only be used to assist a Radiologist.

Clinical Impact of Performance:
This device was designed to achieve a F1 score higher than 0.3, leading to a higher Recall rate. The algorithm gives a Recall score of 0.933, which is a good performance indicator. This will lead to a higher False positive rate, which means it can classify a prediction as Positive despite it being negative for Pneumonia. But this will assist radiologists get better at their X-ray readings with a good accuracy for positive tests.

The algorithm will also produce false negatives, which is not an intended outcome ideally, and can be dangerous if taken into consideration without any further assessment by a Radiologist.

2. Algorithm Design and Function

Algorithm Flowchart



The algorithm is integrated in the regular clinical workflow, where X-rays images in DICOM format will be fed. These images will be validated before it can be used.

DICOM validation Steps:
DICOM checking involves three tests

- Modality test to ensure image is of Digital X-ray type
- Patient position to ensure image has Anteroposterior(AP) or Posterior-Anterior(PA) position of patient
- Patient body to ensure image is a Chest X-ray

Once image passes all three DICOM checking steps, it is feed to the device

Preprocessing Steps:
First part of the device is pre-processing which ensures images fed to the deep neural network algorithm are processed well to ensure best algorithmic performance. This involves image resizing to 224 by 224 pixels, and normalized from 0,255 to 0,1

CNN Architecture:

The CNN architecture is based on VGG16 base model, and has following layers on the top of VGG16 model:

- Flatten
- Dropout with 50% probability
- Dense, 1024 units with ReLU activation
- Dropout Dropout with 50% probability
- Dense, 512 units with ReLU activation
- Dropout with 50% probability
- Dense, 256 units with ReLU activation
- Dense, 1 unit with Sigmoid activation

3. Algorithm Training

Parameters:

Types of augmentation used during training:
Data augmentation was only applied to the training set images, which included:

- A horizontal flip
- Random height shift of ±10% of image height
- Random width shift of ±10% of image width
- Random rotation transformation of ± 20 degrees
- Random shear transformation of ±10 degrees,
- Random zooming of ±10%

Batch size: 32 images per batch for training and 128 images per batch for validation

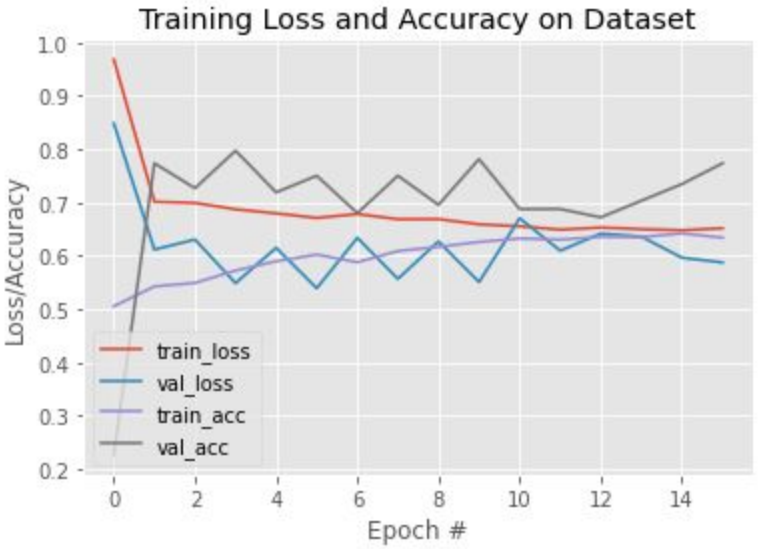
Optimizer learning rate: RMSprop optimizer with learning rate of 0.0005 was used while training

Layers of pre-existing architecture that were frozen:
All the layers except last output layer of the base VGG16 model was frozen, 17 in total

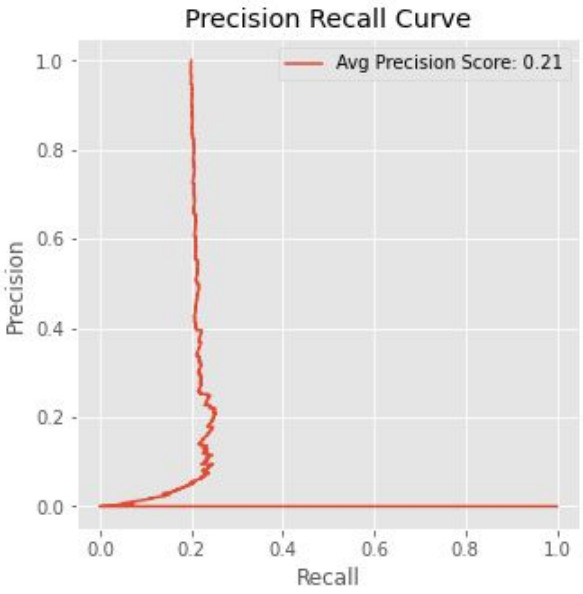
Layers of pre-existing architecture that were fine-tuned:
Last output layer of base VGG16 model was fine tuned to flatten it for new layer addition

Layers added to pre-existing architecture:
Already described in CNN architecture

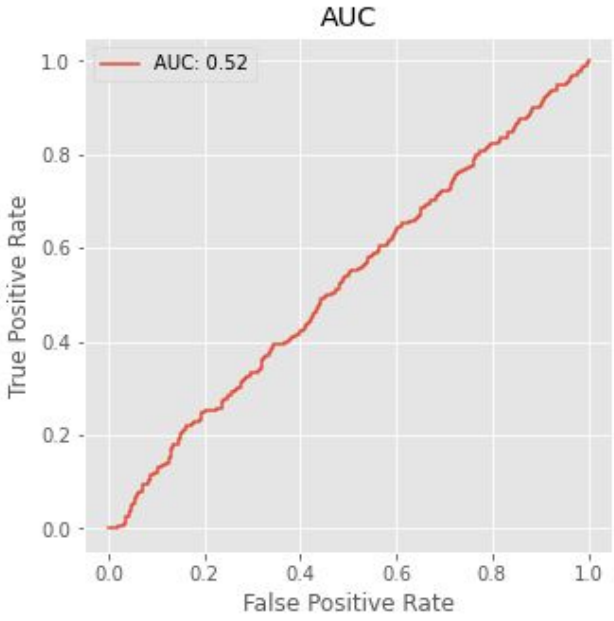
Algorithm training performance visualization



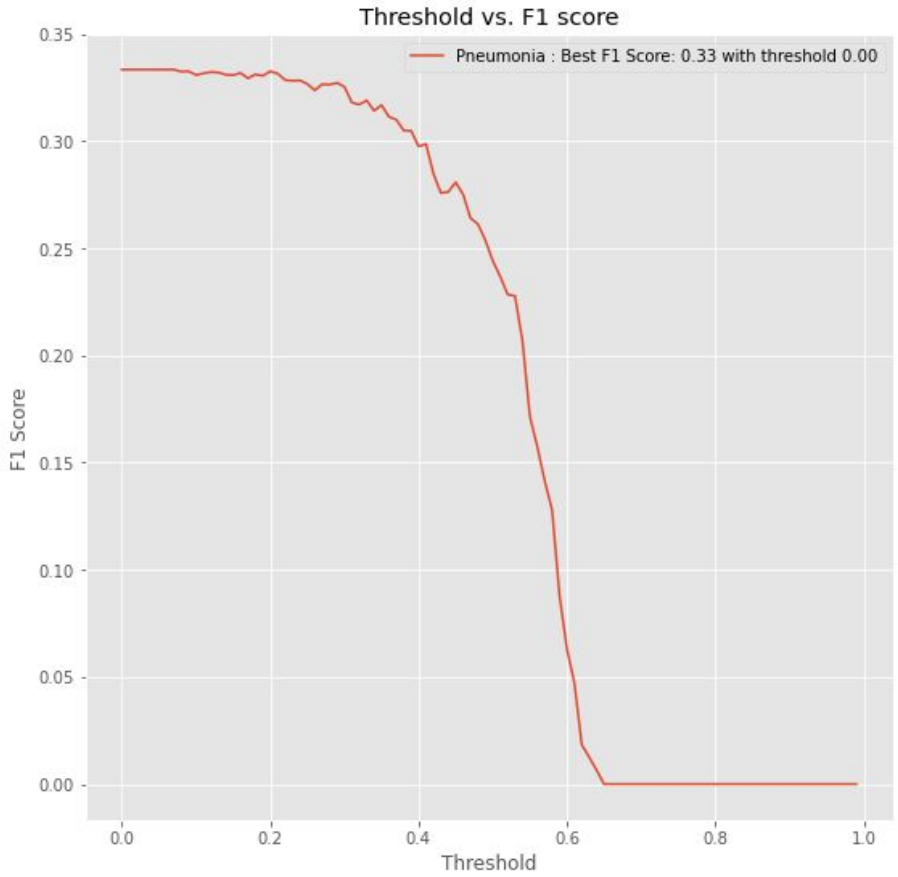
P-R curve



AUC Curve



Final Threshold and Explanation:



From the above graph, in order to get the highest F1 score, a threshold of 0.2 is selected. The F1 score is 0.3323 at 0.2 threshold.

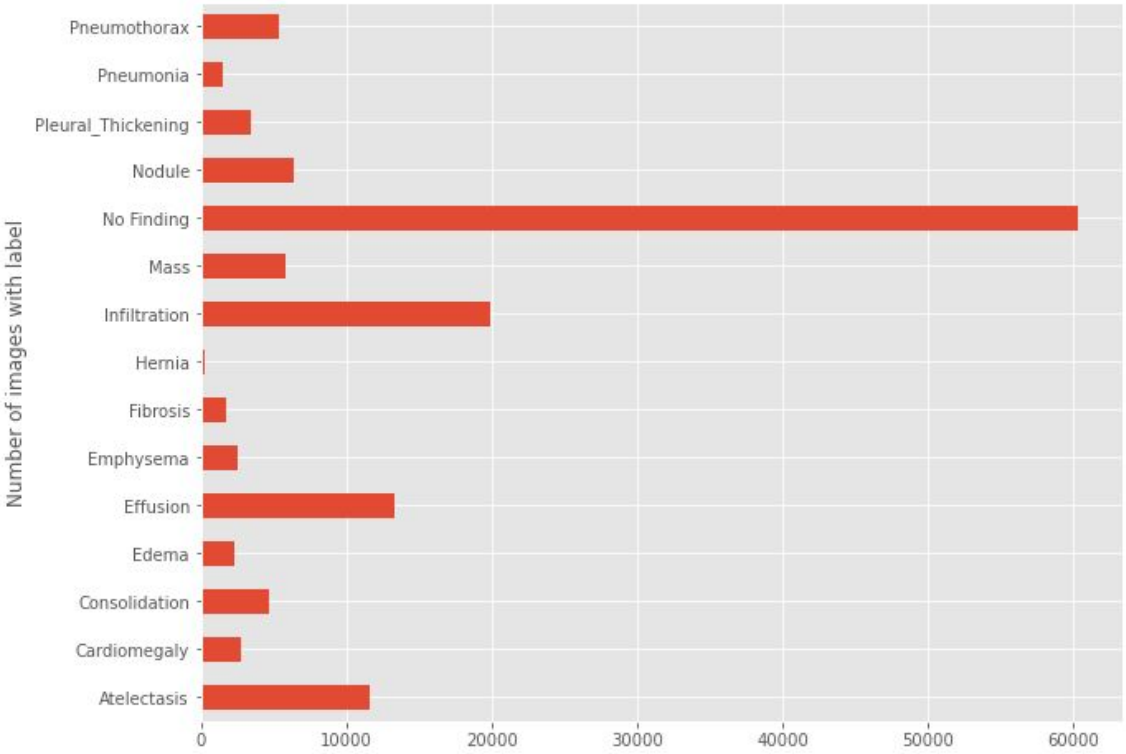
Selecting this threshold value gives higher recall rate:

```
Precision is:  0.20211960635881907
Recall is:  0.9335664335664335
Threshold is:  0.19970363
F1 Score is:  0.3322962041070317
```

4. Databases

National Institute of Health(NIH) Chest X-rays dataset was used for this project. This dataset is hosted on the Kaggle website. There are 112,120 X-ray images with disease labels from 30,805 unique patients in this dataset.The labels include 14 common thoracic pathologies:

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

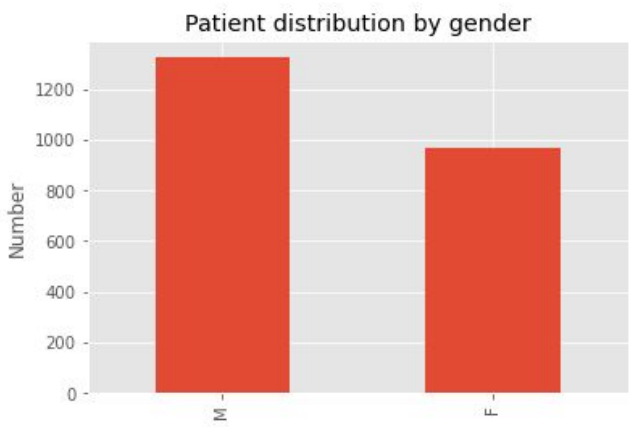


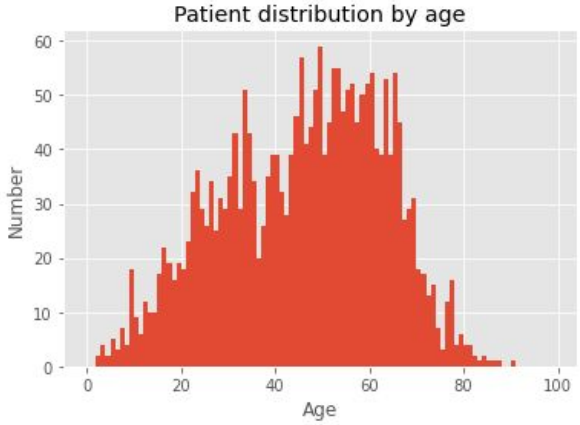
Disease wise distribution of the dataset can be seen in the above figure.

Description of Training Dataset

Training dataset had 1145 number of pneumonia cases and 88551 number of non-pneumonia cases. Percentage of pneumonia cases in the train set was 1.28% which was balanced to 50% by dropping non-pneumonia cases. Total representation of this updated train set:

Male: 1325
Female: 965

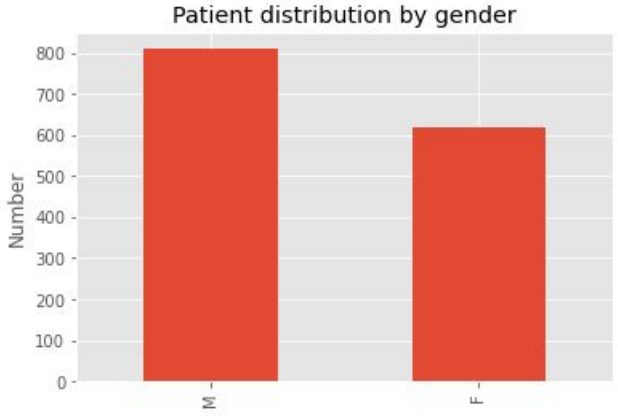


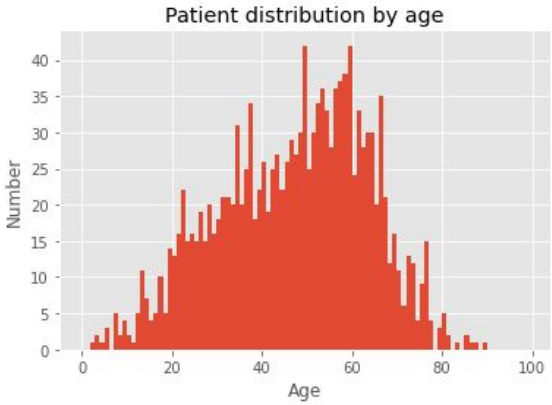


Description of Validation Dataset

Validation dataset had 286 number of pneumonia cases and 22138 number of non-pneumonia cases. Percentage of pneumonia cases in the validation set was 1.28% which was balanced to 20% to simulate the real world scenario by dropping non-pneumonia cases. Total representation of this updated train set:

Male: 810
Female: 620





5. Ground Truth

Ground truth for the disease labels was created using Natural Language Processing (NLP) to mine the associated radiological reports. Since the labels are NLP-extracted, there could be some erroneous labels but the NLP labeling accuracy is estimated to be more than 90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
For FDA validation set, the partner hospital or institute should create a dataset with both male and female cases in almost equal number, spread across age groups of 1 to 100 years. Dataset can contain patients with or without prior diagnosis of any of 14 common thoracic diseases. Ideally it should have a prior illness column, with prior diagnosis and None if no earlier illness.

Ground Truth Acquisition Methodology:
The ground truth should be acquired in AP or PA position of the chest part of a patient and should be of a Digital X-ray type.

Since this device is intended to be used in a non-emergency and non-life threatening situation, a sliver standard will be required to generate ground truth for the FDA validation set; which means a team of Radiologist's assessment is required to generate ground truth. This assessment can be a weighted assessment corresponding to the years of their experience, but not mandatory.

Algorithm Performance Standard:
Algorithm performance can be measured in terms of Recall score and it should be higher than 0.9 or 90%. Other performance metrics like F1 score and AUC curve can be used, but for our algorithm, we prefer to use a Recall Score for performance measurement.
(Ref: https://arxiv.org/pdf/1409.1556.pdf, Page 12, Cal-101 and Cal-256 mean class Recall score for VGG16)