FDA Submission

Your Name: Pranath Fernando

Name of your Device: Pneumonia detector

Algorithm Description

1. General Information

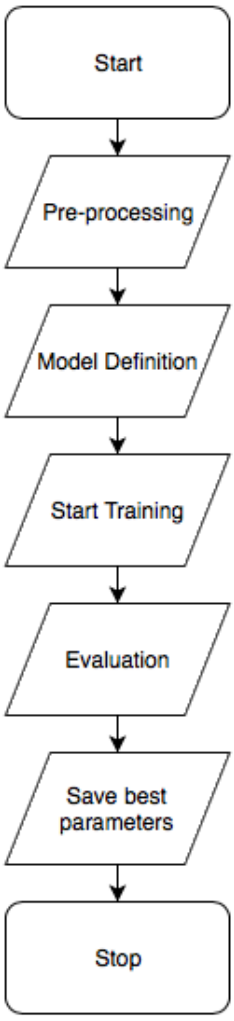
Intended Use Statement: Assisting a radiologist with identifying Pneumonia

Indications for Use: Indicated for use in screening chest X-ray images for males and females between the ages of 1-95.

Device Limitations: From our EDA we would expect lower performance for detection of Pneumonia in the presence of Effusion coupled with Infiltration.

Clinical Impact of Performance: We would expect approximately 50% of true Pneumonia cases to be classified correctly for the selected model threshold. In particular, a patient should expect a false positive rate of 27% and a false negative rate of 50% for predictions with this model.

2. Algorithm Design and Function



DICOM Checking Steps: Check diacom file has following meta-data set - BodyPartExamined is CHEST, Modality is DX, and PatientPosition is either PA or AP

Preprocessing Steps:

- Types of augmentation used during training: Optional horizontal flip, Height shift range 0.1, Width shift range 0.1, Rotation range 10, Sheer range 0.1, Zoom range 0.1
- Other image pre-processing (train & test): Images resized to 225x225 dimensions, image intensities normalised by division by 255,

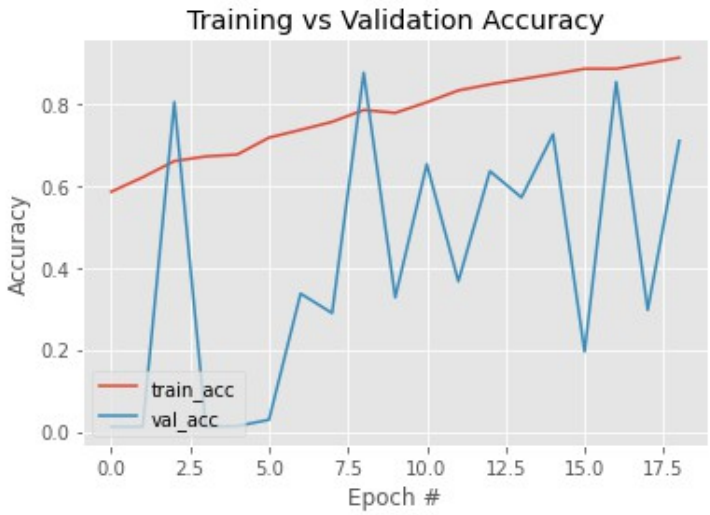
CNN Architecture: Resnet50

3. Algorithm Training

Parameters:

- Batch size: Training set – 16, Test set - 2000
- Optimizer learning rate: 0.0001
- Layers of pre-existing architecture that were frozen: All bar final
- Layers of pre-existing architecture that were fine-tuned: Final layer
- Layers added to pre-existing architecture: Single dense layer, with a sigmoid activation function and a binary cross-entropy loss function

Ttraining performance visualization

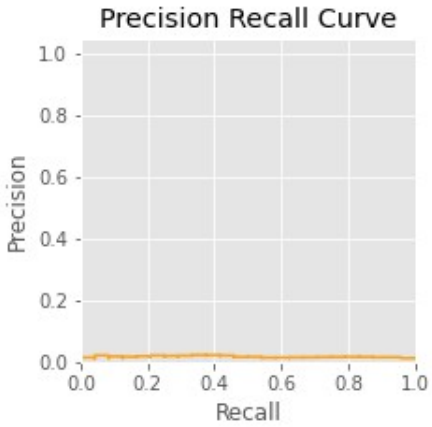


Model architecture & training

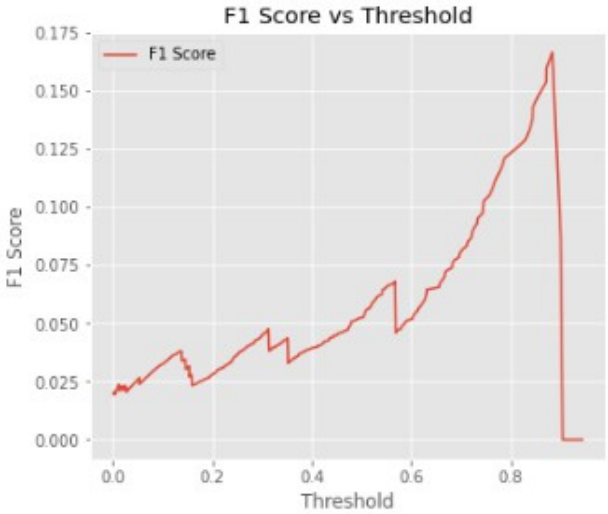
The model is based on a Resnet50 architecture pre-trained on the imagenet dataset. The pre-trained model final layer was removed and replaced with a single dense layer, with a sigmoid activation function and a binary cross-entropy loss function. Binary accuracy and loss metrics were output during training for both training and validation data.

Only the new final model layer was enabled for model fine tuning & training, the rest of the pre-trained network parameters were frozen for fine-tuning.

Precison-Recall Curve



F1 Score vs Threshold



F1, Recall, Precision scores & Confusion matrices for threshold values of 0.05 & 0.1

```
threshold: 0.05
----------
Predicted     0    1   All
Reality
0.0        1231  751  1982
1.0           8   10    18
All        1239  761  2000

Classification report:
              precision    recall  f1-score   support

         0.0       0.99      0.62      0.76      1982
         1.0       0.01      0.56      0.03        18

    accuracy                           0.62      2000
   macro avg       0.50      0.59      0.40      2000
weighted avg       0.98      0.62      0.76      2000


threshold: 0.1
----------
Predicted     0    1   All
Reality
0.0        1456  526  1982
1.0           9    9    18
All        1465  535  2000

Classification report:
              precision    recall  f1-score   support

         0.0       0.99      0.73      0.84      1982
         1.0       0.02      0.50      0.03        18

    accuracy                           0.73      2000
   macro avg       0.51      0.62      0.44      2000
weighted avg       0.99      0.73      0.84      2000
```

Final Threshold and Explanation

Generally we can see that regardless of threshold value, the model struggles to do a good job classifying positive Pneumonia cases - with roughly half getting mis-classified in all cases.

We can see from the above metrics that this is a difficult threshold value to balance. While the threshold value of 0.05 gives us the highest Recall value of 0.56 for the 1.0 Pneumonia cases - and the lowest false negatives, we can see this comes at a great cost of creating 751 false positives (as seen in the confusion matrix). While we want to priortise reducing false negatives, we still care about false positives.

If we look at the next threshold value of 0.1, while it has a slightly lower recall value of 0.50 and just one more false negative, this drastically reduces the false postives from 751 down to 526 false postives. So on balance, for this model I would suggest the best trade-off threshold value, would be 0.1.

At this threshhold of 0.1, we should expect a false positive rate of 526/(526+1456) = 0.27 = 27%.

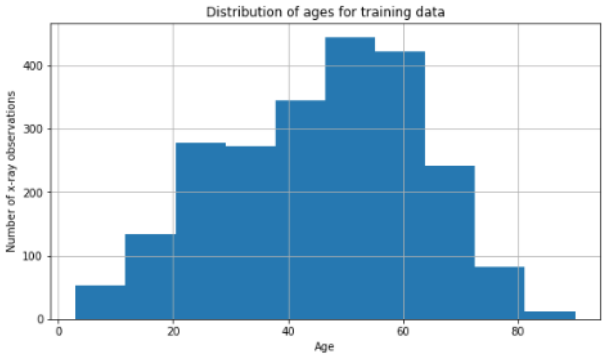
At this threshhold of 0.1, we should expect a false negative rate of 9/(9+9) = 0.5 = 50%.

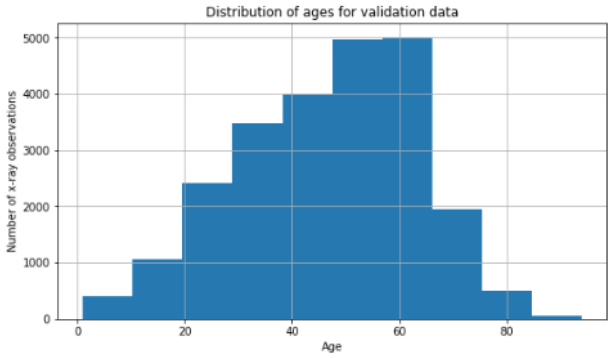
4. Databases

Description of Training Dataset:
Training set: 1,144 (50%) Pneumonia cases, 1,144 (50%) Non-Pneumonia cases - Total 2,288

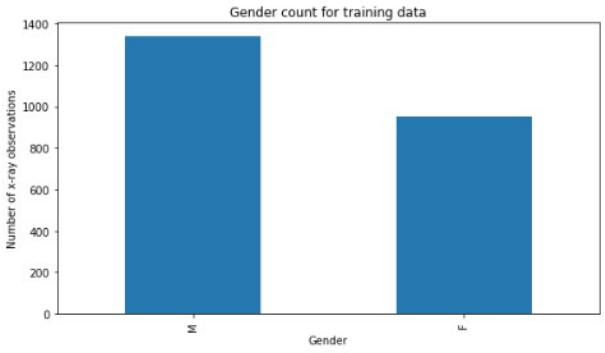
Description of Validation Dataset:
Test set: 286 (1.2%) Pneumonia cases, 23,547 (98.8%) Non-Pneumonia cases - Total 23,833

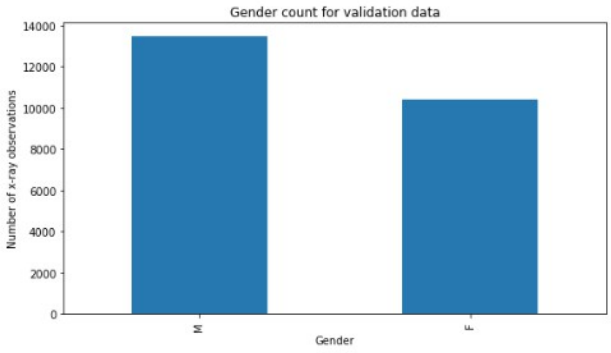
Distributions of ages in training vs validation data



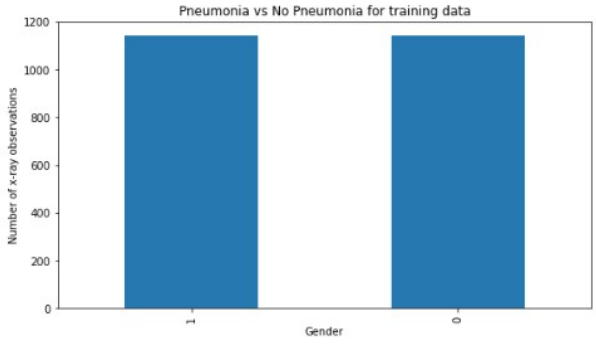


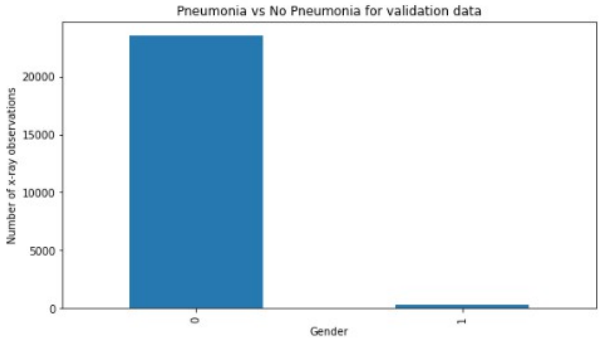
Propotions of gender in training vs validation data



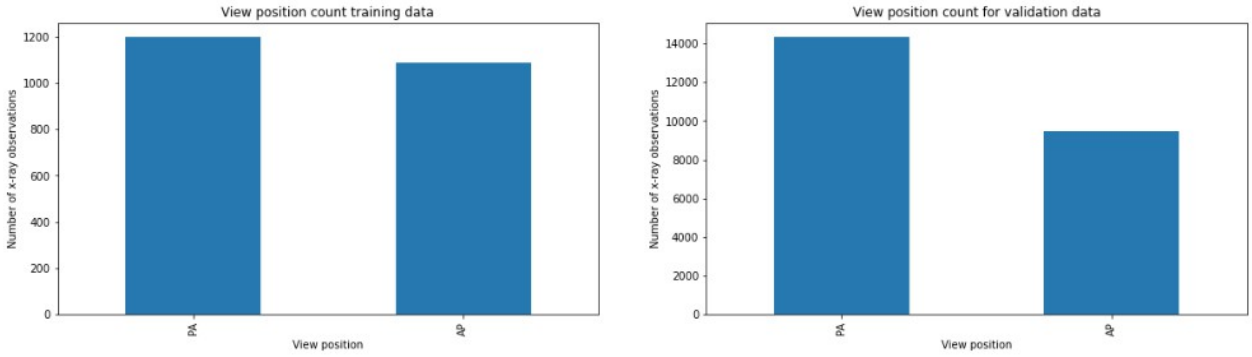


Proportions of Pnuemonia vs No Pneumonia cases in training vs validation data





Propotions of view positions in training vs validation data



So these proportions of key features are as we wished and expected for our training and validation data. The distributions of ages, proportions of gender and view position in the training and validation are roughly the same.

For the Pneumonia vs No Pneumonia cases, in our training set we have equal amounts of each case to give the model the best chance for training, while in the validation data we have a much smaller proportion of Pneumonia cases that matches the real world disease prevelance of approximately 0.1% that we observed in the EDA study.

This should mean we can expect the performance of the model on the validation set to be a good indicator of how the model will perform with unseen data in a real world clinical setting.

5. Ground Truth

To create these labels, the dataset creators used Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be >90% accurate and suitable for weakly-supervised learning.

The advantages of using this method to create these labels is that it is much more cost-effective using this automated method, as well as making it much easier and quicker to generate the dataset.

The dis-advantage of using this method to create this labels is it is not as high a standard of accuracy as say using human radiologists to label the dataset.

The lower reliability of this method say compared to using a human-labelled dataset, will probably impact on the model's ability to learn and achieve a better level of performance.

The silver standard approach of using several radiologists would be more optimal for establishing a more reliable ground truth if feasible and available.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
We would want a balanced distribution of males and females between the ages of 1-95.

Ground Truth Acquisition Methodology:

Use 3 Radiologists to score validation dataset and take highest vote.

Algorithm Performance Standard:

With the ground truth established, we would want to establish an F1 score for the ground truth, then compare the algorithm F1 score.