FDA Submission

Name: Nivedita Bijlani

Device: Assistive AI algorithm to help detect pneumonia from chest X-rays.

Algorithm Description

General Information

Intended Use Statement:

This algorithm is intended for assisting the radiologist in the detection of pneumonia from 'AP' or 'PA' chest X-rays from patients between the ages of 30-70 years.

Indications for Use:

Screening 'AP' or 'PA' chest X-rays of patients, both men and women, between the ages of 30-70 years for presence of pneumonia, either when pneumonia alone is present, or when pneumonia is co-present with Edema, Infiltration or Atelectasis.

Assistive device to flag patients with potential pneumonia to focus radiologist attention to these cases; may be useful for prioritising their workload.

Device Limitations:

Not indicated for use in patients outside of the age range listed in the "Indications for Use" above.

Not indicated for use in detecting any other lung condition such as edema, effusion, infiltration, atelectasis etc. from chest X-rays, whether or not these are comorbid with pneumonia.

If device is used for workload prioritization for a radiologist, then a GPU will be required for fast performance on streaming chest X-ray images.

Device has a lower than listed precision performance if pneumonia is co-present with one or more of the following morbidities:

- Cardiomegaly
- Effusion
- Emphysema
- Fibrosis
- Hernia
- Mass
- Nodule
- Pleural Thickening

- Pneumothorax

Clinical Impact of Performance:

Given an X-ray, the algorithm first calculates the probability of pneumonia. It then uses a decision threshold, computed based on the model training dataset, of classifying the chest X-ray as a positive or negative pneumonia case. The threshold was chosen based on maximising the algorithm's F1-score on the internal hold-out or validation dataset. The F1-score is the harmonic mean of precision and recall and reflects a score that aims to balance the performance of the algorithm for false positive and false negatives.

To find the optimal threshold, a precision recall curve was used to get an idea of the optimal threshold that balances both precision and recall. This was followed by an F1 score vs. threshold curve to pick the threshold that maximises the F1-score.

A false positive for a patient means the patient's chest X-ray is flagged as a pneumonia case when in reality it isn't. This will prompt the radiologist to spend a bit more time than usual in examining the X-ray to rule out pneumonia, but a false positive will not act to harm the patient in any way.

A false negative for a patient means the patient's chest X-ray is identified as not showing pneumonia when in reality the patient has pneumonia. This will prompt the radiologist to maybe spend less time than usual in examining the X-ray to rule out pneumonia, but as the radiologist is still the final authority on chest X-ray classification here, they should be able to correctly identify pneumonia, and report the algorithm's incorrect finding to the Medical Device Reporting system.

Algorithm Design and Function

The algorithm or model uses a fine tuned VGG16 CNN architecture to classify chest X-rays.

The algorithm was trained on 2288 chest X-rays, half of which are pneumonia positive and half pneumonia negative, as labelled by an expert radiologist. The overall chest X-ray dataset was provided by the NIH. The algorithm was then tested on a separate validation dataset of X-rays that have a 1:5 distribution of pneumonia-positive to negative cases. This reflects the real-world distribution of pneumonia positive and negative cases found in a typical clinical setting.

Having been trained on the training dataset where the algorithm has equal access to what a pneumonia positive and negative case looks like, it was optimized so as to achieve minimum loss over the validation dataset. After training for a maximum of 12 epochs (passes over the training set), the model weights were saved and performance measures computed on the validation dataset.

Algorithm Flowchart:

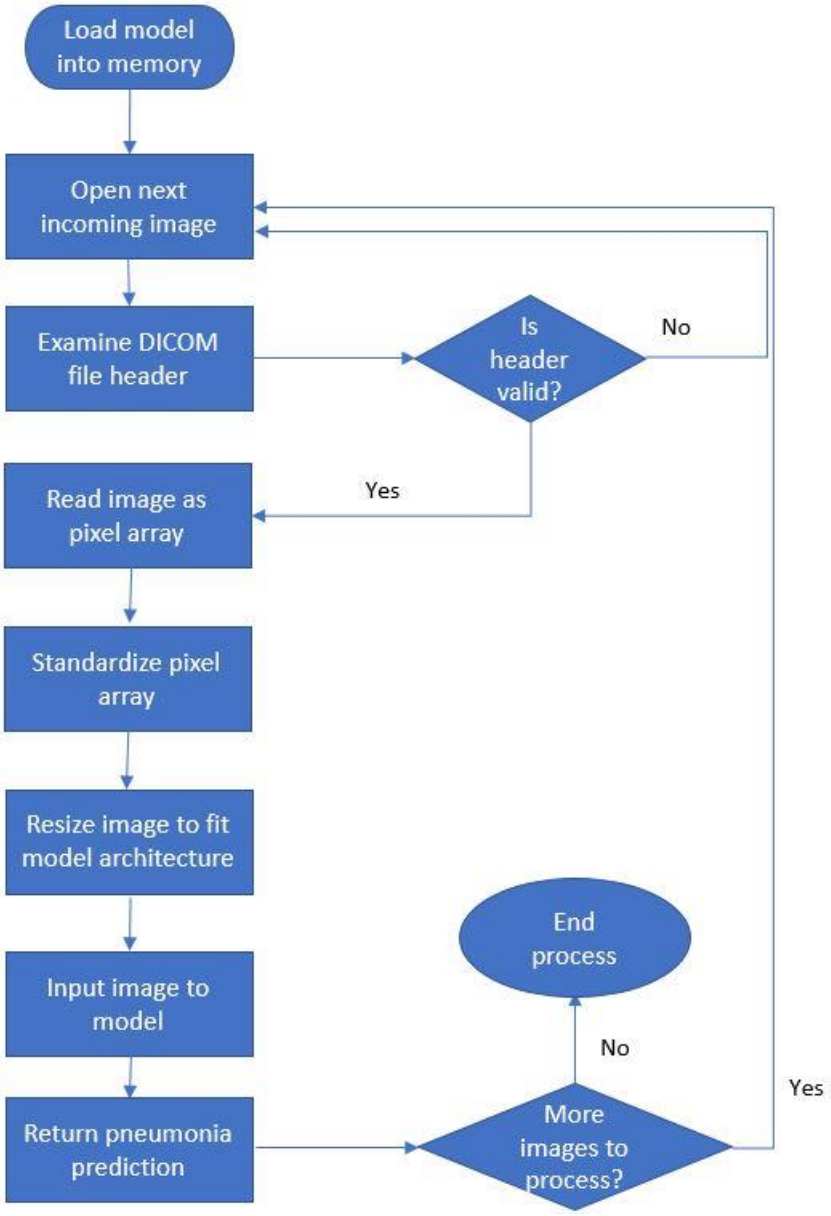

Figure 1 - Algorithm flowchart

DICOM Checking Steps:

Each time a DICOM image is received, the following two checks are performed:

1. Examine the header – The header of the DICOM file is examined. Specifically, the "View Position" and "Body Part Examined" headers are checked. The "View Position" should be "PA" or "AP" and the "Body Part Examined" should be "CHEST" or "RIBCAGE". This ensures that the algorithm will be looking at chest X-rays in the correct position for which the algorithm is trained.

2. Read the image pixels – The image part of the DICOM file is read in and inspected to ensure that there are non-zero pixels, i.e. there is a finite length image available to the algorithm to inspect.

Pre-processing Steps:

1. The mean and standard deviation are first calculated for the pixel intensity values for the incoming image.
2. Each pixel intensity value is normalized by subtracting each pixel's intensity value from its mean and then dividing this by the standard deviation computed above.
3. The normalized image pixel array is then resized from its current size (e.g. 1024, 1024) to (1, 224, 224, 3). This is required because the input layer to the CNN architecture that the algorithm uses is the base layer of a VGG16 model that requires its inputs to be of size (1, 224, 224, 3). This means that a chest X-ray that is a gray scale image will have the gray scale image pixels repeated for each of the 3 channels. A color X-ray, on the other hand will retain the 3 color channels, with each color channel forming one of the 3 dimensions of the (1, 224, 224, 3) image.

CNN Architecture:

The CNN architecture has, as its foundation, the well tested and benchmarked VGG16 CNN model. This architecture was pre-trained and pre-tested on the ImageNet database – a database with over 14 million images belonging to 1000 image classes. The VGG16 model was specifically created for image recognition/classification problems.

The algorithm for pneumonia detection utilizes all of the convolution blocks of the VGG16, and replaces the last 3 dense layers with a sequence of 3 alternating dropout and dense layers followed by a final single node Dense layer that outputs the final pneumonia/not pneumonia classification. The final architecture is shown in Figure 2.

Algorithm Training

Parameters:

Types of augmentation used during training

To enrich the training dataset with even more examples than the ones obtained by partitioning the chest X-ray dataset into train and validation sets, image augmentation was applied to the existing training data for model training. The augmentation techniques used reflect some common image contortions or variations seen in a clinical setting and are taken from standard available augmentation functionality offered by the Keras deep learning package:

1. Rotation range = 30 degrees: Not all images chest X-rays are perfectly vertical. Sometimes the X-ray scanner may not be placed directly in line with the AP or PA positions, resulting in images that are slightly tilted to normal. A realistic value here would be a maximum of 30 degrees of rotation. So we present such rotationally augmented images to our model when training.
2. Width shift range = 0.1: This setting reflects an X-ray stretched wide by a maximum of 10% over the original image.
3. Height shift = 0.1: This setting reflects an X-ray stretched lengthwise by a maximum of 10% over the original image.
4. Brightness range = [0.5, 1.3]: This setting augments existing training images by lightening or darkening them by 50% to 130% respectively, a common range seen in a clinical setting depending on the X-ray machine setting and make.
5. Shear range = 0.1: This setting acts to shift the top and bottom edges of the X-ray by a maximum of 10% in opposite directions.
6. Zoom range = 0.2: This augmentation acts to magnify the original image by a maximum of 20%. It reflects X-rays taken at close range.
7. Horizontal flip = True: This setting reflects the fact that some X-rays might present organs flipped left to right depending on how the X-ray was recorded.
8. Vertical flip = False: This setting explicitly switches off vertical flipping of the original images, as the presentation of organs upside down is not a valid or acceptable presentation of an X-ray.

Batch size

A batch size of 64 was used when model training. Considering the training set size of 2288 images, this would result in approximately 36 batches of 64 images. 64, being a whole number exponent of 2, ensures that a full unit of memory can fit these images nicely. It also offers a large and diverse enough set of images for the algorithm to learn enough from in one pass. It is also a good enough size to compute gradient updates over. Benchmark implementations using similar size of datasets also use a batch size of 64, as it tends to mitigate overfitting.

Optimizer learning rate

The Adam optimizer with a constant learning rate of 0.0001 (or 1e -4) was used for model training. This learning rate was chosen to be small enough not to overshoot the point of minimum loss or get stuck on a local minimum optimal training, but large enough for the training to proceed and reach the point of minimum loss quickly and prevent overfitting.

A batch size of 64 was used when model training. Considering the training set size of 2288 images, this would result in approximately 36 batches of 64 images. 64, being a whole number exponent of 2, ensures

Layers of pre-existing architecture that were frozen

The first four convolutional blocks of the pre-existing VGG16 model trained on the ImageNet database, were all frozen. Referring to the figure below, this comprises “Conv 1-1” through to and including “Conv 5-2”. The weights for these layers were frozen, not to be updated during model training.

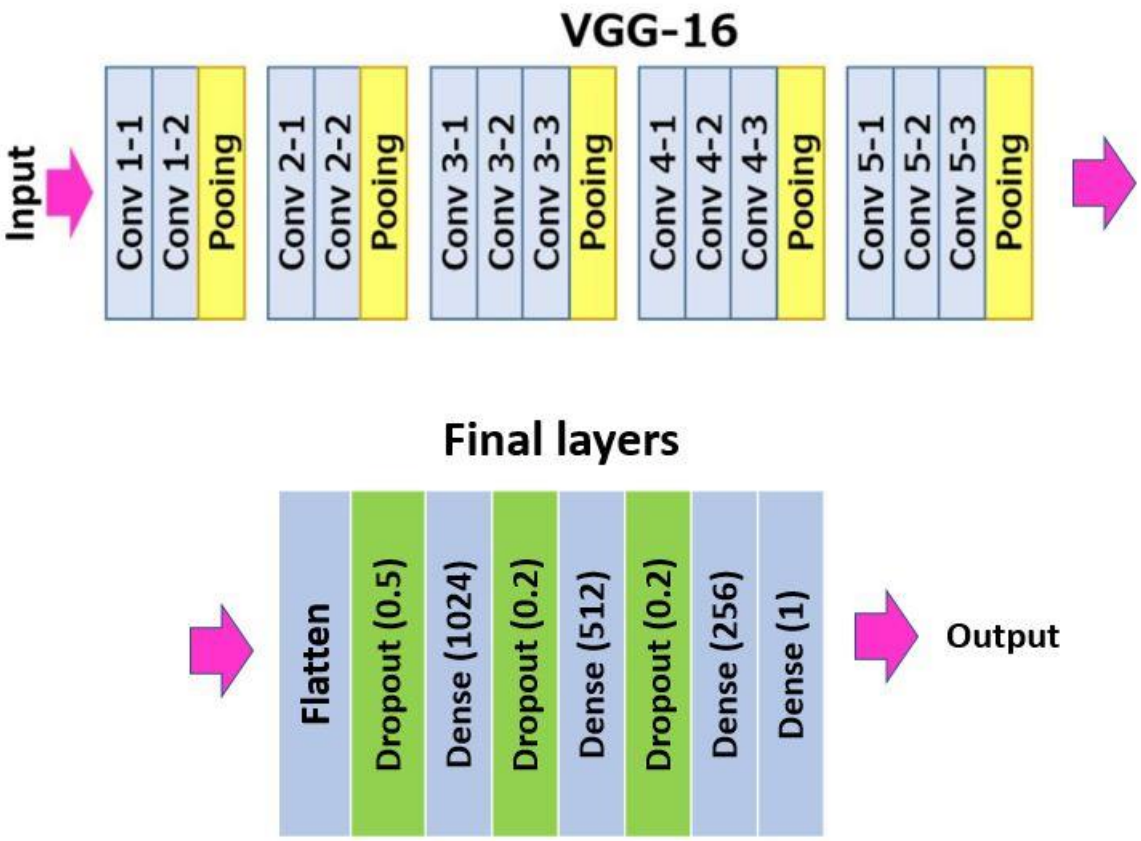

Figure 2: CNN Architecture

Layers of pre-existing architecture that were fine-tuned

The last convolutional layer of the VGG16 model, i.e. the Conv5-3 + Pooling layers were fine tuned when training the overall model.

Layers added to pre-existing architecture

As shown in the figure above, the output from the final convolutional block of the pre-existing VGG16 model was first flattened. This was then fed to the following sequence of dense layers to perform the final classification for the existence of pneumonia:

Dropout (rate 50%) -> Dense (1024 nodes) -> Dropout (20%) -> Dense (512 nodes) -> Dropout (20%) -> Dense (256 nodes) -> Dense (1 node)

The dropout nodes were added to mitigate overfitting in the final layers, and the final dense node was implemented with a single node or neuron with sigmoid activation so that the output from the overall network would be a probability of the chest X-ray showing pneumonia.

The model summary is evidenced below:

```
Downloading data from https://github.com/fchollet/deep-learning-models/releases/download/v0.1/vgg16_weights_tf_dim_ordering_tf_
kernels.h5
553467904/553467096 [==============================] - 6s 0us/step
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
```

Figure 3 - Model summary

Performance summary:

The algorithm was trained for 12 epochs with the loss function set to “binary cross entropy” and the training process set to monitor the loss over the complete validation set, with a checkpoint set to save the model at the end of the epoch where the minimum validation loss in training was obtained.

The loss graphs are shown below:

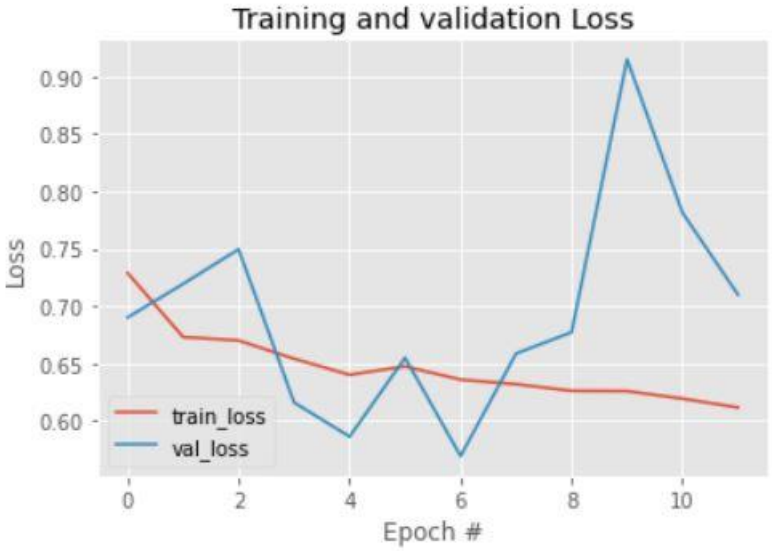

Figure 4 - Training and validation loss

We see that the model reached its lowest training loss of 0.57 at the end of the epoch 6. This is the model that was saved for use.

The corresponding accuracy curves are shown below:

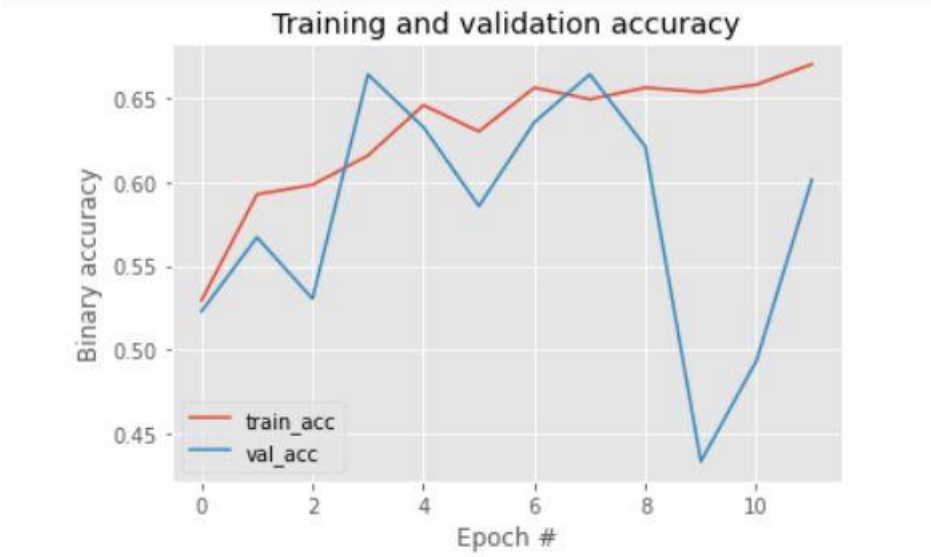

Figure 5 - Training and validation accuracy

The AUC curve over the complete validation set is shown below:

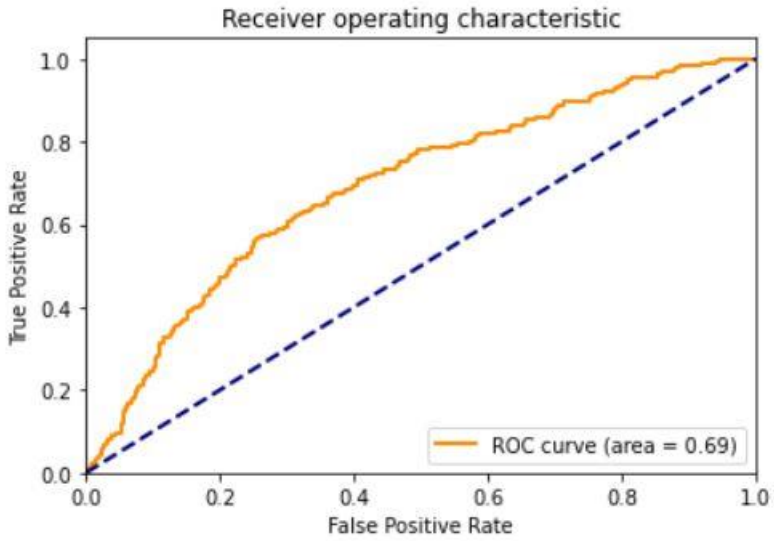

Figure 6 - AUROC curve

The Precision-Recall curve over the complete validation set is shown below:

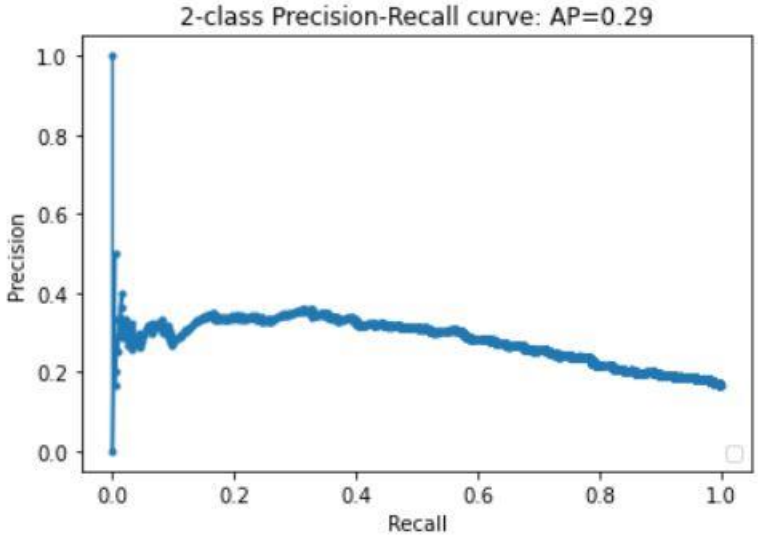

Figure 7 - Precision Recall Curve

The precision-recall curves as a function of the decision threshold are shown below:

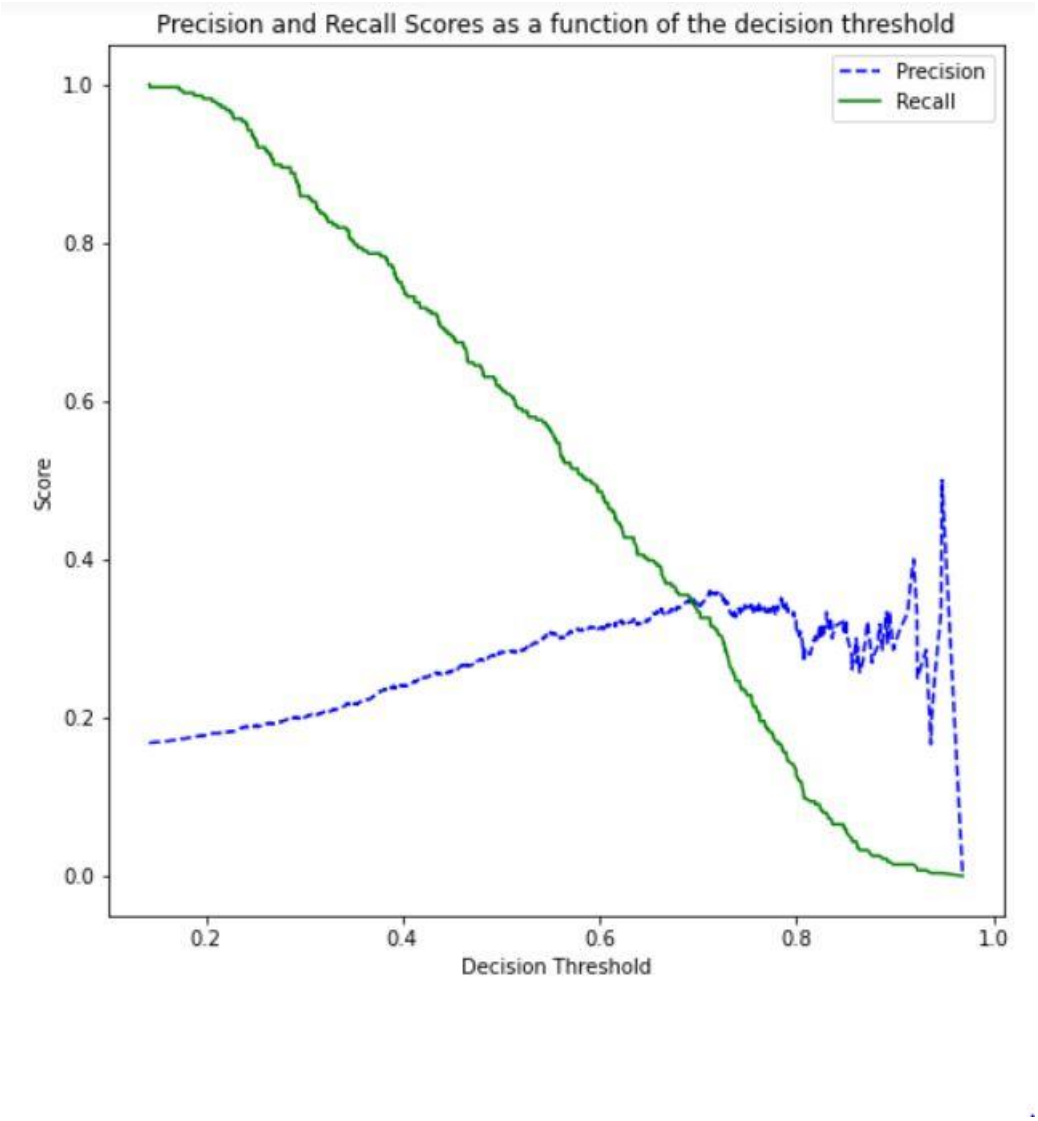

Figure 8 - PR curve as a function of threshold

The following curve plots the F1 scores as a function of the decision thresholds:

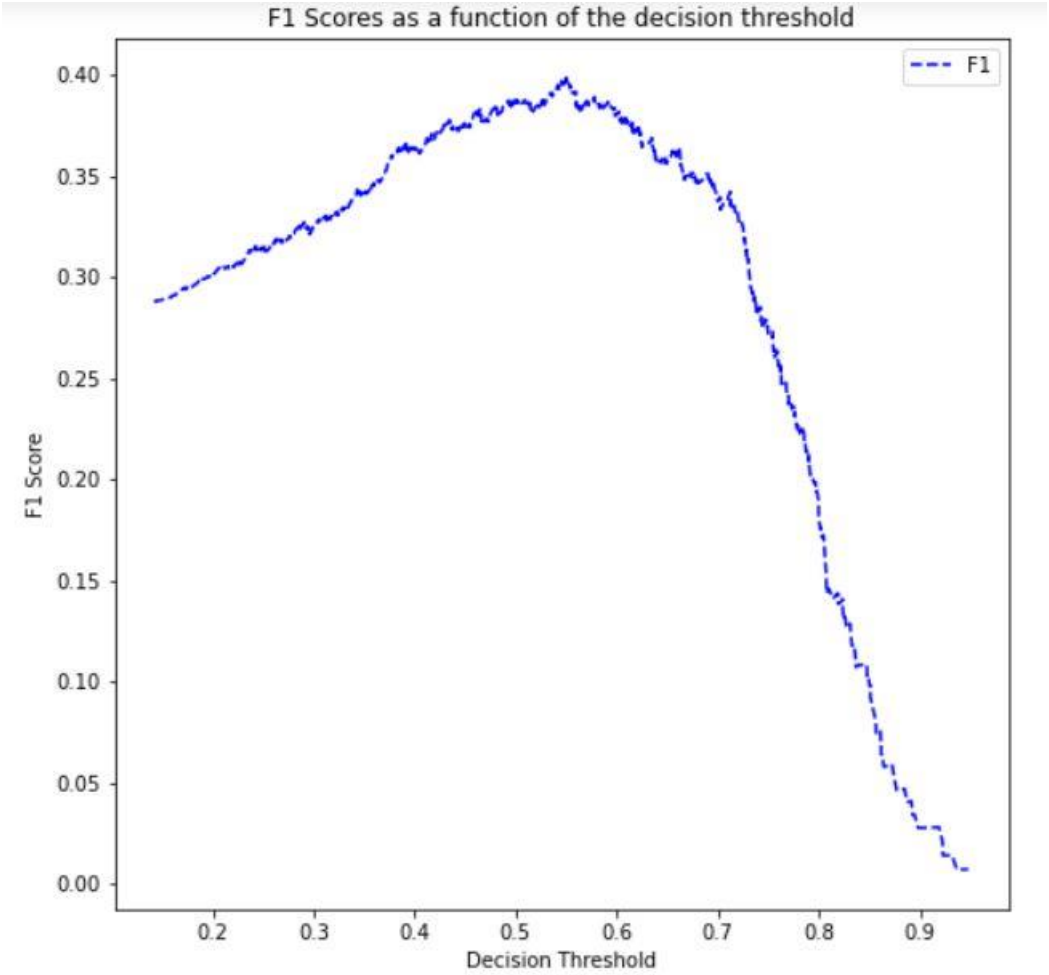

Figure 9 - F1 score as a function of threshold

Final Threshold and Explanation:

The final threshold was chosen based on the maximum F1 score. The F1 score is the harmonic mean of the precision and recall and balances them optimally. Given this is a highly imbalanced dataset -only 1.27% of the X-rays were labelled as showing pneumonia – we could choose either Precision or F1 score. Note from Figure 8 that based on precision only, the threshold that yields the maximum value of precision is between 0.95 and 1. However at this threshold, the recall or sensitivity of the algorithm is extremely poor – between 0.1 and 0.2 only. Therefore we do not use precision but the F1 score as the criterion for deciding on the final threshold.

The final threshold based on the maximum F1 score, as shown in Figure 9 is **0.50763667**. The performance figures computed at this threshold are:

F1 score: 0.387

Precision: 0.284

Sensitivity: 0.695

Specificity: 0.609

The performance figures for pneumonia classification in the presence of various comorbidities is as follows:

Condition	Precision	Recall	F1-score

Infiltration	0.303	0.982	0.463
Effusion	0.207	1.0	0.343
Edema	0.731	1.0	0.845
Atelectasis	0.284	0.983	0.44
Cardiomegaly	0.137	1.0	0.241
Consolidation	0.307	0.964	0.465
Emphysema	0.189	0.875	0.311
Mass	0.203	1.0	0.337
Nodule	0.113	1.0	0.204
Pleural Thickening	0.108	1.0	0.194
Pneumothorax	0.103	1.0	0.1875
Pneumonia only (no other comorbidity)	1.0	0.986	0.993

Table 1 - Performance figures for pneumonia classification

Legend: Green implies the F1-score is as high or above the F1-score of 0.387 computed over the entire dataset, at the chosen threshold.

Databases

Overall dataset

The dataset for chest X-rays was provided by the National Institute of Health (NIH). It comprises **112,120** chest X-rays along with the information for each patient whose X-ray it is. The information includes the Patient ID, Age, Gender, View Position, list of diseases and image pixel related information and the file path to the corresponding chest X-ray file. Pneumonia is one of these 14 diseases. A "No Finding" implies no disease for the patient. A summary of the data revealed 16 patients whose ages were recorded to be > 110 years and being outliers, these were removed. This left us with **112,104** records. There are **63328 Male** and **48776 Female** patient records. There are **67299 chest X-rays with view position "PA"** and **44805 with view position "AP"**. The age distribution is shown below:

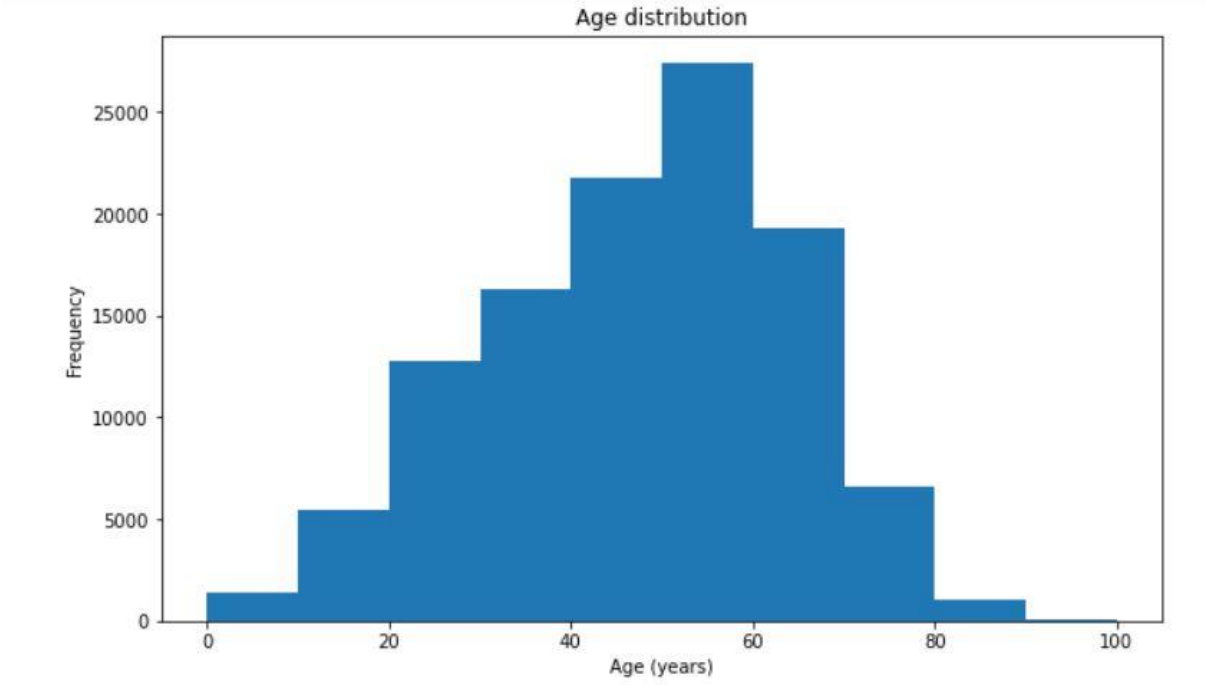

Figure 10 - Age distribution within overall dataset

The disease prevalence is as follows:

Pneumonia: 1430

Non-pneumonia: 110674

Percentage of pneumonia records: 1.276%

The age distribution for pneumonia patients is shown below:

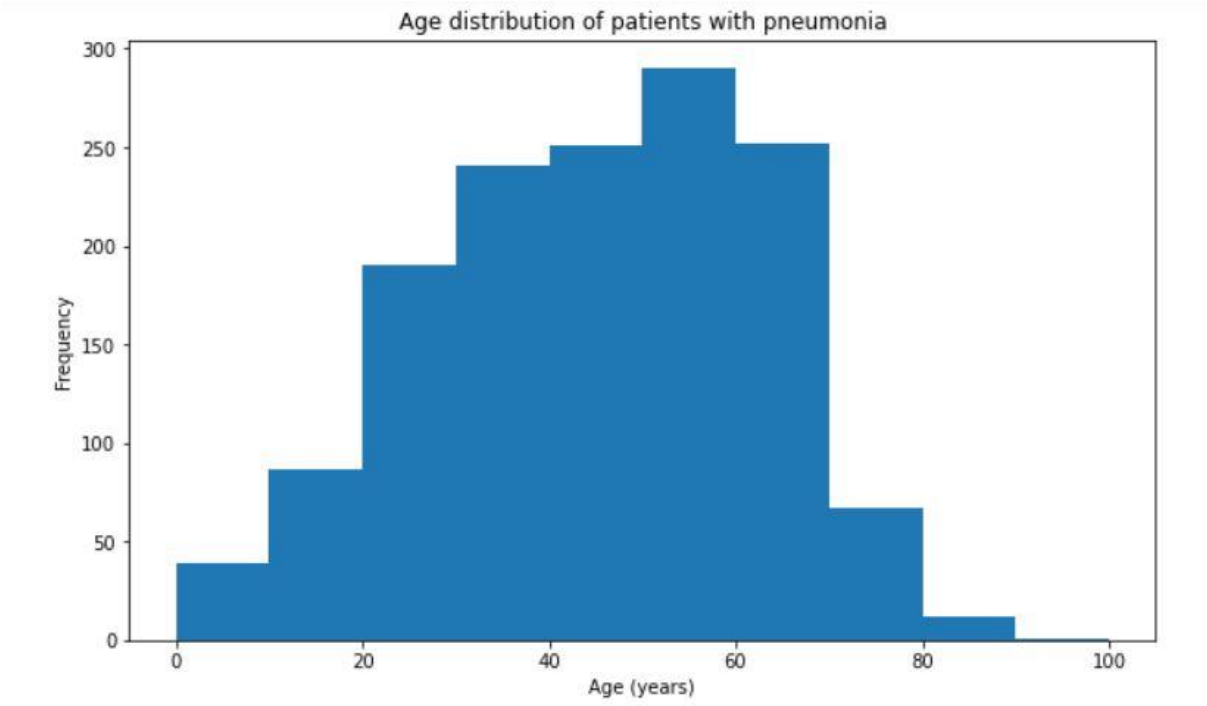

Figure 11 - Age distribution of pneumonia patients in overall dataset

We can see from the above data that any algorithm for pneumonia detection would be most relevant to patients between 30 and 70 years of age, as this is the range for which maximum data is available.

The gender distribution of pneumonia patients is shown below:

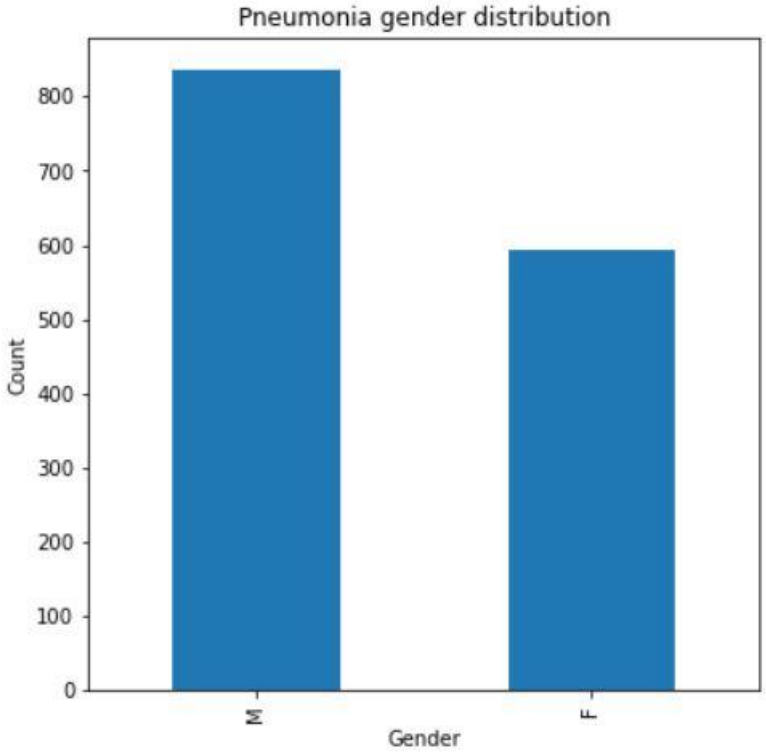

Figure 12 - Gender distribution in overall dataset

The following graph shows the diseases comorbid with pneumonia:

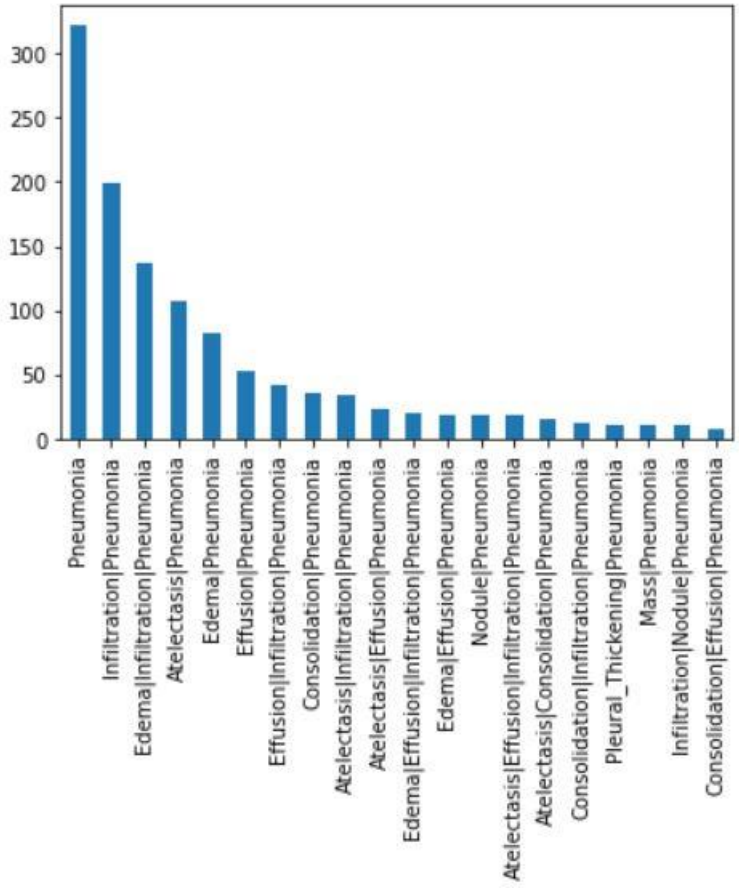

Figure 13 - Diseases comorbid with pneumonia in overall dataset

The following graph shows the distribution of the number of diseases for patients:

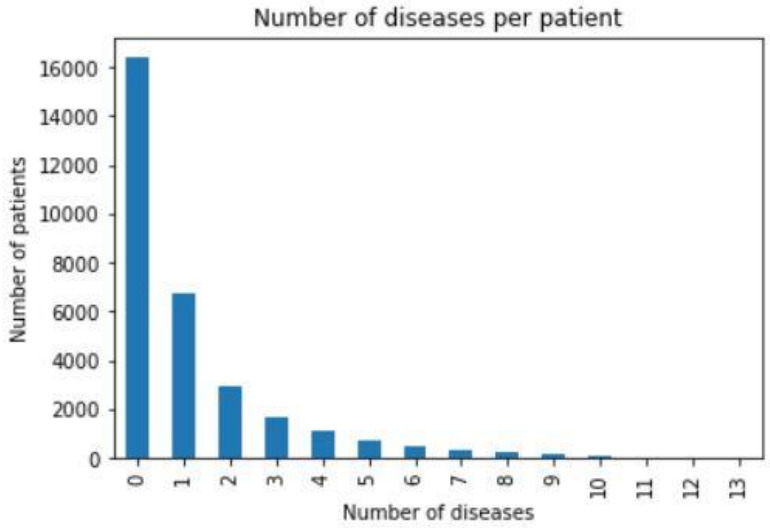

Figure 14 - Diseases per patient in overall dataset

Training Dataset:

The training dataset was selected as follows:

1. Select 80% of the original dataset, with the stratification based on the pneumonia class. In other words, this first splitting step would select 80% of the overall records in the dataset into the training set, where 1.276% of this training set are pneumonia cases.
2. The training dataset was then balanced to obtain a 50/50 prevalence of pneumonia positive and pneumonia negative cases. This was achieved by the down-sampling or discarding of a large number of negative cases, so that there were equal numbers of positive and negative cases.

The final number of records in the training dataset are **2288**, with 1144 being pneumonia positive and 1144 being negative.

The distribution of age, gender and view position is similar to that found in the original dataset, and is shown below:

Male: 1335

Female: 953

View position “PA”: 1179

View position “AP”: 1109

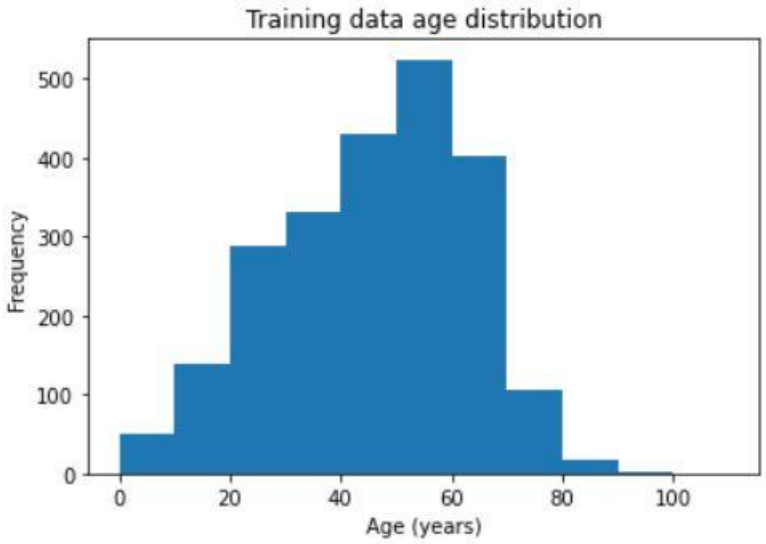

Figure 15 - Age distribution in training data

We can again see from the above histogram that any algorithm for pneumonia detection would be most relevant to patients between 30 and 70 years of age, as this is the range for which maximum training data is available.

The distribution in the age range 30-70 years is as follows:

Male: 975

Female: 765

View Position "PA": 905

View Position "AP": 796

View Position "PA" within Males: 518

View Position "AP" within Males: 475

View Position "PA" within Females: 387

View Position "AP" within Females: 321

The following graph shows the diseases comorbid with pneumonia in the training set:

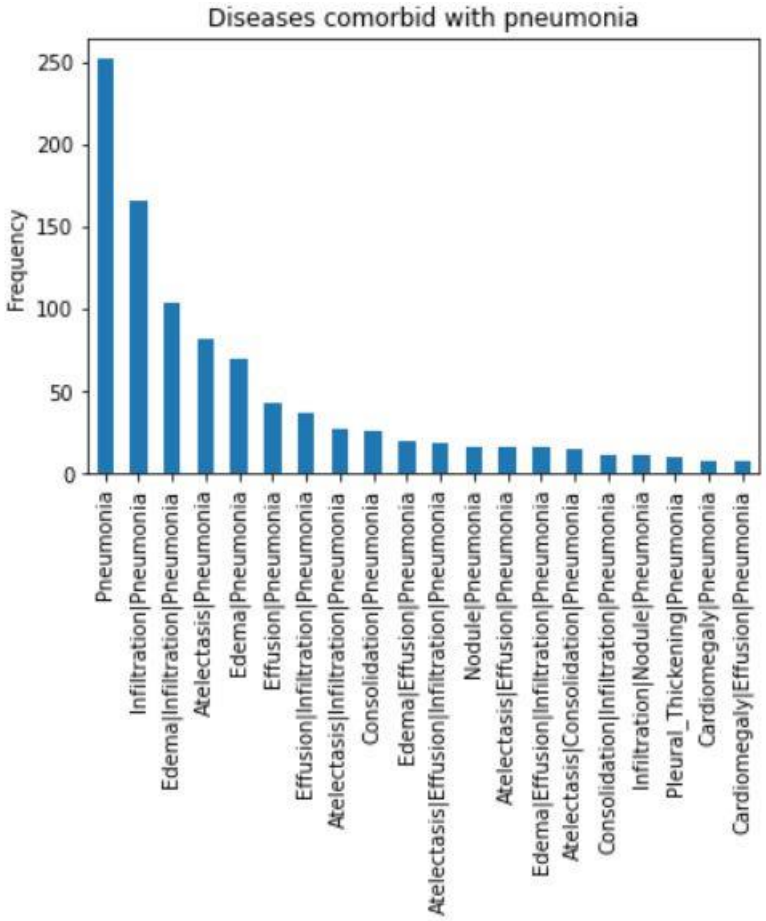

Figure 16 - Diseases comorbid with pneumonia in training data

We can see that the disease prevalence is similar to that found in the overall dataset.

Validation Dataset:

The validation dataset was selected as follows:

1. Select 20% of the original dataset, with the stratification based on the pneumonia class. In other words, this first splitting step would select 20% of the overall records in the dataset into the training set, where 1.276% of this training set are pneumonia cases.
2. The validation dataset must reflect real world prevalence of disease, however it must also provide a realistic number of positive and negative cases so that the performance of the algorithm can be accurately measured and fine-tuned over both pneumonia positive as well as pneumonia negative cases. For this reason, the validation data was resampled after the first step so as to contain pneumonia: non-pneumonia cases in the ration 1:5. This was achieved by the down-sampling or discarding of a number of negative cases, so that the ratio of positive to negative cases was maintained at 1:5

The final number of records in the validation dataset was **1716**, with 286 being pneumonia positive and 1430 being negative.

The distribution of age, gender and view position is similar to that found in the original dataset, and is shown below:

Male: 977

Female: 739

View position “PA”: 976

View position “AP”: 740

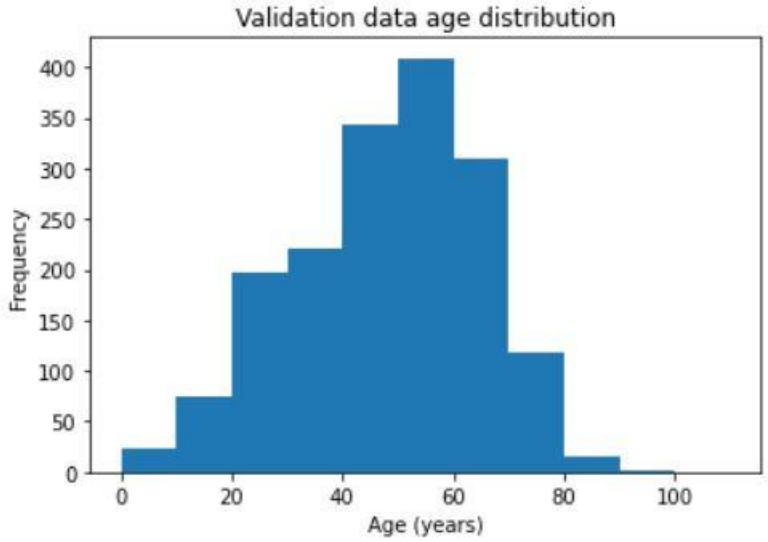

Figure 17 - Age distribution in validation data

We can again see from the above histogram that the algorithm for pneumonia detection can be adequately tuned for patients between 30 and 70 years of age, as this is the range for which maximum validation data is available.

The distribution in the age range 30-70 years is as follows:

Male: 716

Female: 584

View Position “PA”: 760

View Position “AP”: 540

View Position “PA” within Males: 399

View Position “AP” within Males: 317

View Position “PA” within Females: 361

View Position “AP” within Females: 223

The following graph shows the diseases comorbid with pneumonia in the validation set:

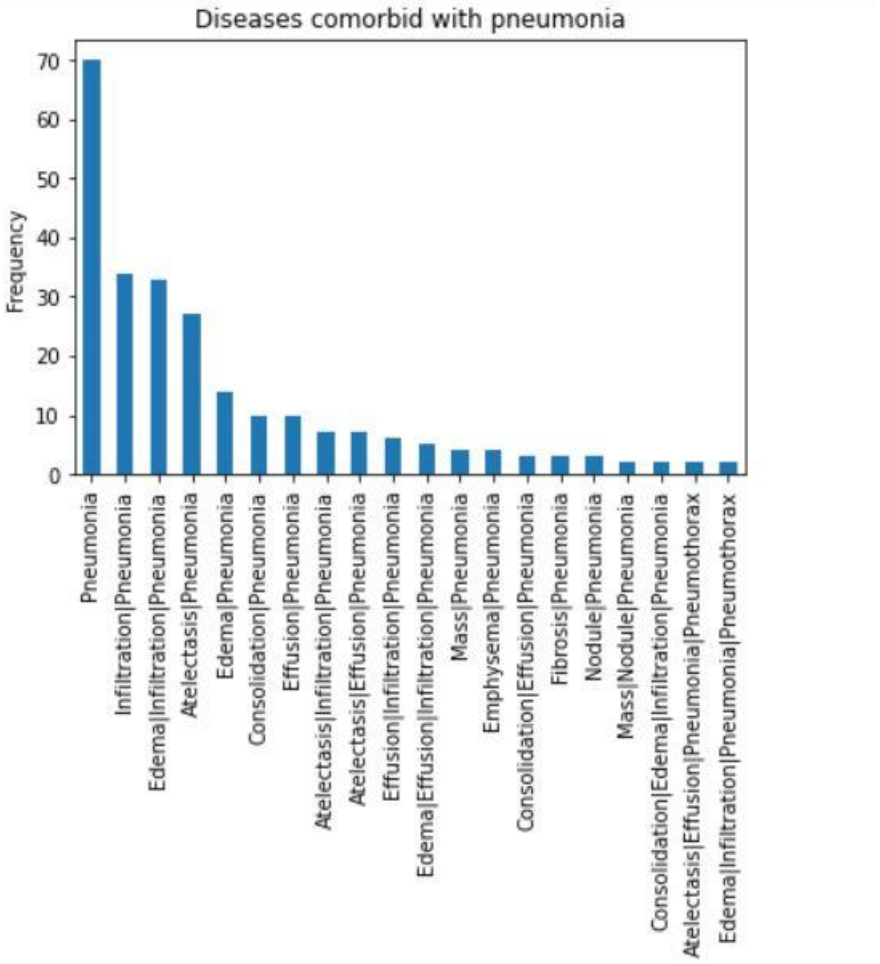

Figure 18 - Diseases comorbid with pneumonia in validation data

We can see that the disease prevalence is similar to that found in the overall dataset.

Ground Truth

The data for this project was labelled by an automated process that extracted the patient's findings from their radiology reports using Natural Language Processing (NLP) techniques. Specifically, weakly supervised multi-label classification was used. The approach involved using text mining to find the disease concepts in certain pre-defined sections of standard radiology reports using two machine learning tools: DNorm and MetaMap. The disease concepts were translated to disease labels with the aid of a board-certified radiologist. Then, negation detection was applied, using a hand-crafted set of negation rules based on syntactic dependency information, to detect one or more diseases mentioned but within a negation, or "not present" context. If a radiology report did not fulfil any of the disease presence rules, it was marked as "No Finding". This method of automatic labelling was itself validated using human annotator interpretation of radiology reports which was used as the gold standard, and resulted in a precision score of 0.66, a recall of 0.93 and an overall F1-score of 0.77 for pneumonia.

The benefits of using this type of ground truth labelling are:

- It provides complete transparency in terms of the logic used to detect diseases from radiologist reports
- It can be scaled with minimal effort to label a large number of chest X-rays
- It represents the best benchmarked effort yet to automatically label chest X-rays based on radiology reports

The disadvantages of using this type of ground truth labelling are:

- It is far from ideal in that the most accurate labelling can only be provided by human annotators who will be able to understand the context and pick up disease labels from radiology reports.
- The inaccuracy in the ground truth leads our deep learning algorithm to make inaccurate assessments in its own labelling of the chest X-rays, and optimise for a significant amount of false "truth". The inaccuracy therefore gets compounded and our algorithm will always remain less accurate than the ground truth.
- The ground truth methodology already scores low on pneumonia detection due to its small prevalence in the overall dataset. This drags down the performance of our pneumonia detection algorithm, and for this reason, our pneumonia detection algorithm may have very specific and limited use in a clinical setting.

The ground truth labelling algorithm is described in this paper: ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.

Authors: Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers Department of Radiology and Imaging Sciences, Clinical Center, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892

FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

The external validation set we would request from our clinical partner to validate our pneumonia detection algorithm would include patients with the following description:

1. Age range: 30-70 years, with between 10%-14% of the data to include patients from 0-30 years, and 70 years+.
2. Sex: Balanced for Male and Female (50%-50%)
3. Type of imaging modality: Chest CT X-ray in the "AP" or "PA" view position. Balanced for each view position.
4. Body part imaged: CHEST, RIBCAGE or similar
5. Prevalence of pneumonia: In the range of 2% to 5%
6. Include: Pneumonia, Infiltration, Atelectasis, Cardiomegaly, Mass, Nodule, Effusion, Edema, Emphysema, Pneumothorax, Normal (or No Finding)
7. Exclude: Any comorbidities other than the above

Ground Truth Acquisition Methodology:

The ground truth for the ideal FDA validation dataset should be created taking a majority vote on each chest X-ray to classify it as pneumonia/not-pneumonia by a panel of 3 to 5 senior board-certified radiologists – the more radiologists the better. It can be hard to detect pneumonia from radiology reports as the differences between pneumonia and other lung diseases can be minor on a given chest X-ray. For this reason, a "silver" standard is better than a "gold" standard based on a single radiologist read. The performance of the pneumonia detection algorithm would be compared to this silver standard.

Algorithm Performance Standard:

The performance of this pneumonia detection algorithm should be based on its **AUROC (Area under the ROC curve) and F1 score** metrics calculated over the validation dataset. As described earlier, the F1 score is the harmonic mean between precision and recall and is calculated as:

F1 score = (Precision + Recall)/(2 * Precision * Recall)

It balances the precision and recall and does not let get one get too low in relation to the other. In this way, it balances the false positives with the false negatives. The AUROC curve shows how much better the algorithm can do at different decision thresholds than a random guess.

According to the literature (CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, Pranav Rajpurkar et al. 2017; ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, Wang et al. 2017), the AUROC benchmark for pneumonia detection on the same dataset as the one the algorithm trained on, ranges between 0.633 and 0.768. Our algorithm on the internal validation dataset achieved 0.69. Given this data, a “good” AUROC on the FDA validation dataset would be between 0.65 and 0.768. Note the higher AUROC was achieved with a more complex CNN architecture requiring with higher GPU and infrastructure requirements.

According to the same benchmarking data as above, the F1 score benchmark for pneumonia detection on the same dataset as the one the above algorithm was trained on, ranges between 0.387 (the average score based on four practicing academic radiologist reads) and 0.435 (for CheXNet, a DenseNet based AI algorithm). Our algorithm on the internal validation dataset achieved 0.387. Given that our algorithm will serve to assist radiologists in detecting pneumonia, a “good” F1 score on the FDA validation dataset would be between 0.387 or above.