FDA Submission

Your Name: Janice Block
Name of your Device: Pneumonia Highlighter

Algorithm Description

1.General Information

Intended Use Statement: The Pneumonia Highlighter offers automated identification of chest roentgenograms likely to be consistent with a diagnosis of pneumonia. It is intended as an aid for the practicing clinical radiologist.

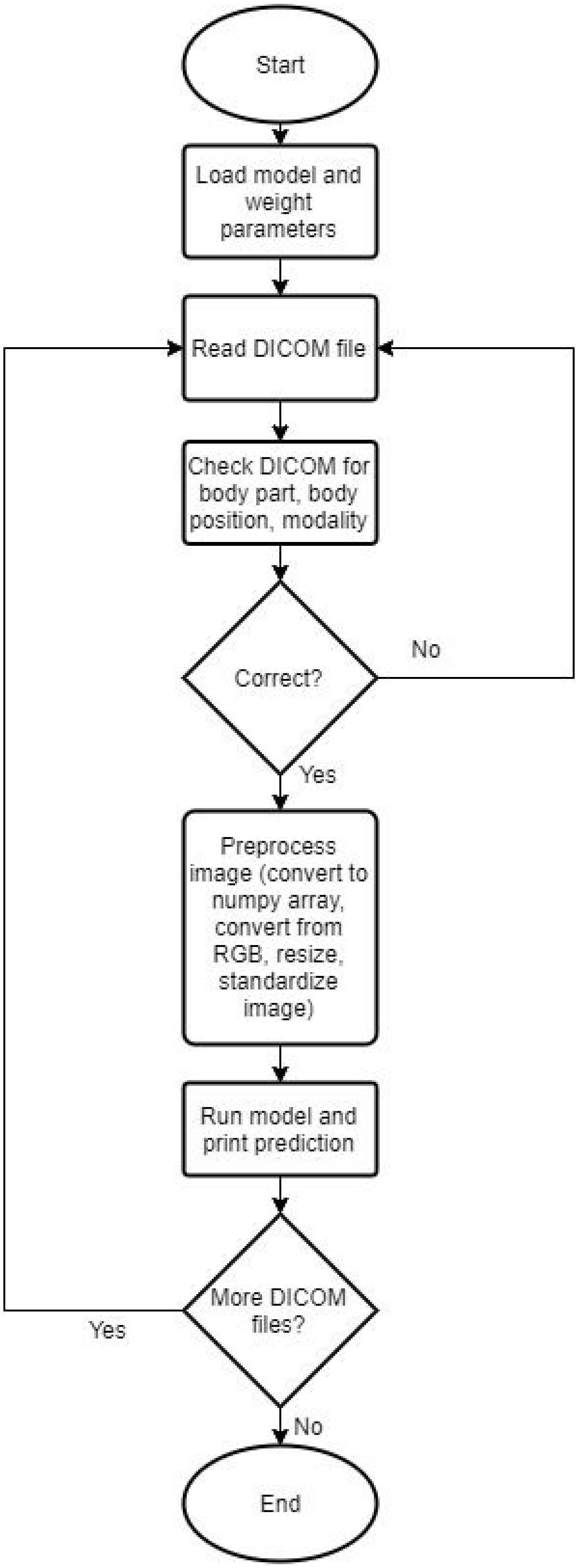
Indications for Use: The Pneumonia Highlighter offers automated identification and tagging of chest roentgenograms likely to be consistent with a diagnosis of pneumonia, thus allowing for rapid screening, identification, and potential prioritization of higher risk films in the queue.

Device Limitations: The Pneumonia Highlighter serves to aid the practicing clinical radiologist but is not a substitute for a radiologist. The device has the following limitations:
1) The Pneumonia Highlighter has been trained and tested on frontal radiographs alone. In many cases the lateral radiograph is also required: such as identifying lingular lobe pneumonia in some people, or such as when distinguishing between pneumonia and effusion, or such as when distinguishing between consolidation and effusion.
2) The Pneumonia Highlighter does not have access to medical history -- a factor which can be critical in distinguishing between pneumonia and other diagnoses. For example, a radiologist might tag a film as “consolidation” or “infiltrate” but imply pneumonia. Consolidations and infiltrates do not necessarily imply pneumonia, and a medical history such as an acute infectious illness with fever and cough may clarify the picture.
3)While the Pneumonia Highlighter is compatible with prioritization software, it does not itself contain software for prioritizing high risk chest roentgenograms. Prioritization software must be supplied separately.
4)The Pneumonia Highlighter reads roentgenograms in DICOM format but not necessarily in other formats. It is appropriate for chest roentgenograms but not for other imaging studies — for example, it is not able to identify pneumonia in an abdominal roentgenogram in which the chest is also partially viewable.

Clinical Impact of Performance: A radiologist can opt to read positively classified films first, prior to reading negative films. This is expected to result in more rapid identification and treatment of infected patients.

The algorithm has been designed to detect as many true positives as possible (because true positives require immediate treatment), despite the fact that this design may also increase the number of false positives. It is more important to correctly identify all true positives than it is to correctly label all negatives. The danger of a positive film classified as negative is that a patient might not receive timely treatment of illness. An error in which a negative film is classified as positive is less of a danger, since no treatment delay is involved. In most cases errors classifying negative films are detected by the expert radiologists and/or primary care physicians.

2. Algorithm Design and Function


Start
Load model and weight parameters
Read DICOM file
Check DICOM for body part, body position, modality
Correct?
No
Yes
Preprocess image (convert to numpy array, convert from RGB, resize, standardize image)
Run model and print prediction
More DICOM files?
Yes
No
End

DICOM Checking Steps:
After reading in the DICOM file, each DICOM file is screened for proper positioning (AP or PA view), proper body part (chest), and proper modality (digital radiography). If the file does not conform to the above specifications for one one or more of these three fields, the process is aborted for that DICOM file and the next DICOM file is read. Otherwise, the DICOM file is converted into a pixel array for processing.

Preprocessing Steps:
The pixel array is converted into a numpy array, converted from RGB (so that the input will be small), and resized to the shape required for the model, i.e. (batch size, dimension[0], dimension[1], channels). Images are then standardized by subtracting the mean and dividing by the standard deviation so that all images are comparable by the algorithm.

CNN architecture:
The CNN architecture is that of VGG16 plus four added layers.
VGG16:The conv1 layer is of a fixed size (224,224) RGB image. The image is passed through a stack of convolutional layers, followed by three fully connected layers. All hidden layers within VGG16 are equipped with rectification (ReLU) nonlinearity.
Added layers:
Flatten() - reshapes the tensor so that it is the shape equal to the number of elements contained in the tensor.
Dropout (0.5) - increases sparsity of connections between layers to reduce overfitting
Dense (1024, activation = ‘relu’) - a nonlinear function layer
Dense (1, activation = ‘sigmoid’) - last layer; returns a value between 0 and 1.

3. Algorithm Training

Parameters:

Batch size is specified as 70. (It was found that larger batch sizes overfit, and smaller batch sizes produced poorer predictions).
LaBatches of images are transformed as follows:

Rescaling:
rescale=1. / 255.0
Images are rescaled from RGB for smaller input values; works better with keras deep learning.

Position and image dimension adjustment for better comparisons between images:

```
horizontal_flip = True,
vertical_flip = False,
height_shift_range = 0.1,
width_shift_range = 0.1,
rotation_range = 20,
```

Shear transformation for desired geometry:

shear_range = 0.1,

Zoom transformation to zoom in on image desired:

zoom_range= 0.1

Optimizer learning rate:

Adam(lr=1e-4). The Adam algorithm optimizes the learning rate to 0.0001. (Slower learning rates resulted in overfitting, and more rapid learning rates resulted in poorer predictions.)

Loss function:

Binary cross entropy, a loss function for binary classification tasks used for determining the decision boundary of classification. With binary cross entropy, classification is between two classes.

Tuning of layers of preexisting VGG16 architecture:

For the VGG16 layers, layers [0:17] are frozen. The last two layers are trainable, meaning that they are the layers utilized for fine-tuning of the model.

Layers added to the preexisting architecture:

Flatten() - reshapes the tensor so that it is the shape equal to the number of elements contained in the tensor.
Dropout (0.5) - increases sparsity of connections between layers to reduce overfitting
Dense (1024, activation = ‘relu’) - a nonlinear function layer
Dense (1, activation = ‘sigmoid’) - last layer; returns a value between 0 and 1, for purposes of classification. We classify by selecting a threshold within this spectrum from 0 to 1.

Performance visualization curves (Fig.1, Fig.2):

Fig. 1

Accuracy of the validation set was deemed to be less important than ongoing decreases in loss with each epoch. As can be seen in Fig.1, overfitting occurred rapidly after the first epoch: training accuracy increased and train loss decreased over each epoch, whereas validation accuracy decreased and validation loss increased over later epochs.

It was possible to remove this overfitting by reducing batch size and/or increasing learning rate, but only at significant cost to overall predictive ability of the model. The algorithm maintains the best weights obtained even when further epochs show worsening model fits for the validation set.

Fig. 2

The AUC curve plots the number of true positives as a function of the number of false positives. It is a measure of reliability of positive model predictions, i.e. how many images labeled by the algorithm as “pneumonia” are true positives for pneumonia, and how many images labeled by the algorithm as “pneumonia” are false positives? AUC values range between 0 and 1, and an AUC value of 0.5 implies that an algorithm’s positive predictions are no better than that of random selection.

As can be seen from Fig 2, the AUC value for this algorithm is 0.65. The algorithm performance for positive predictions is better than random, but many false positives remain.

Fig.1

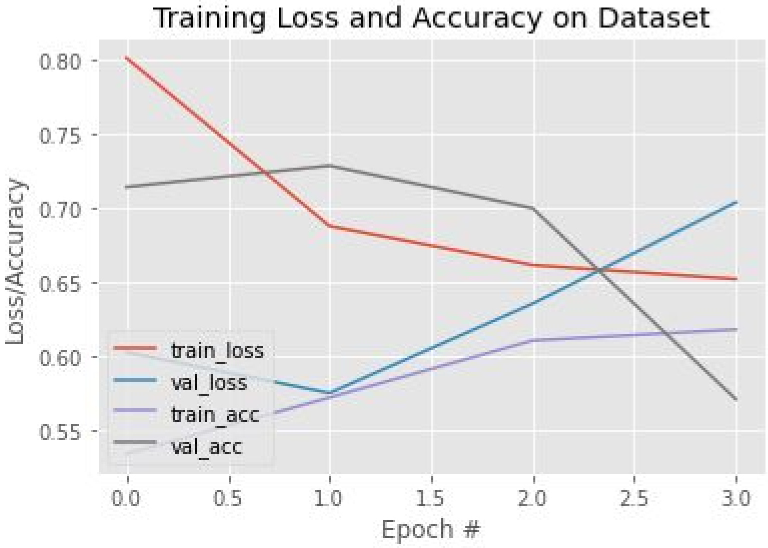
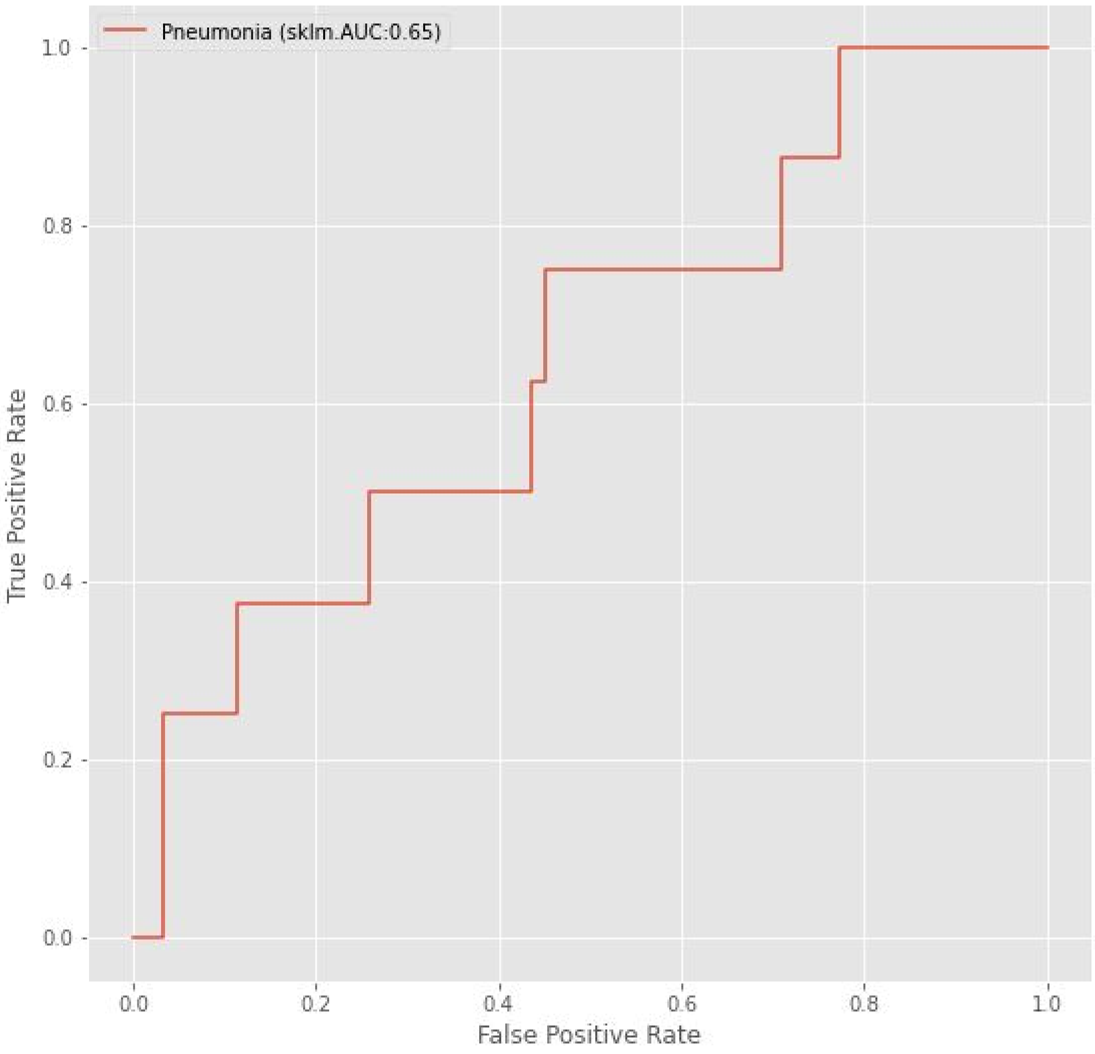


Fig.2



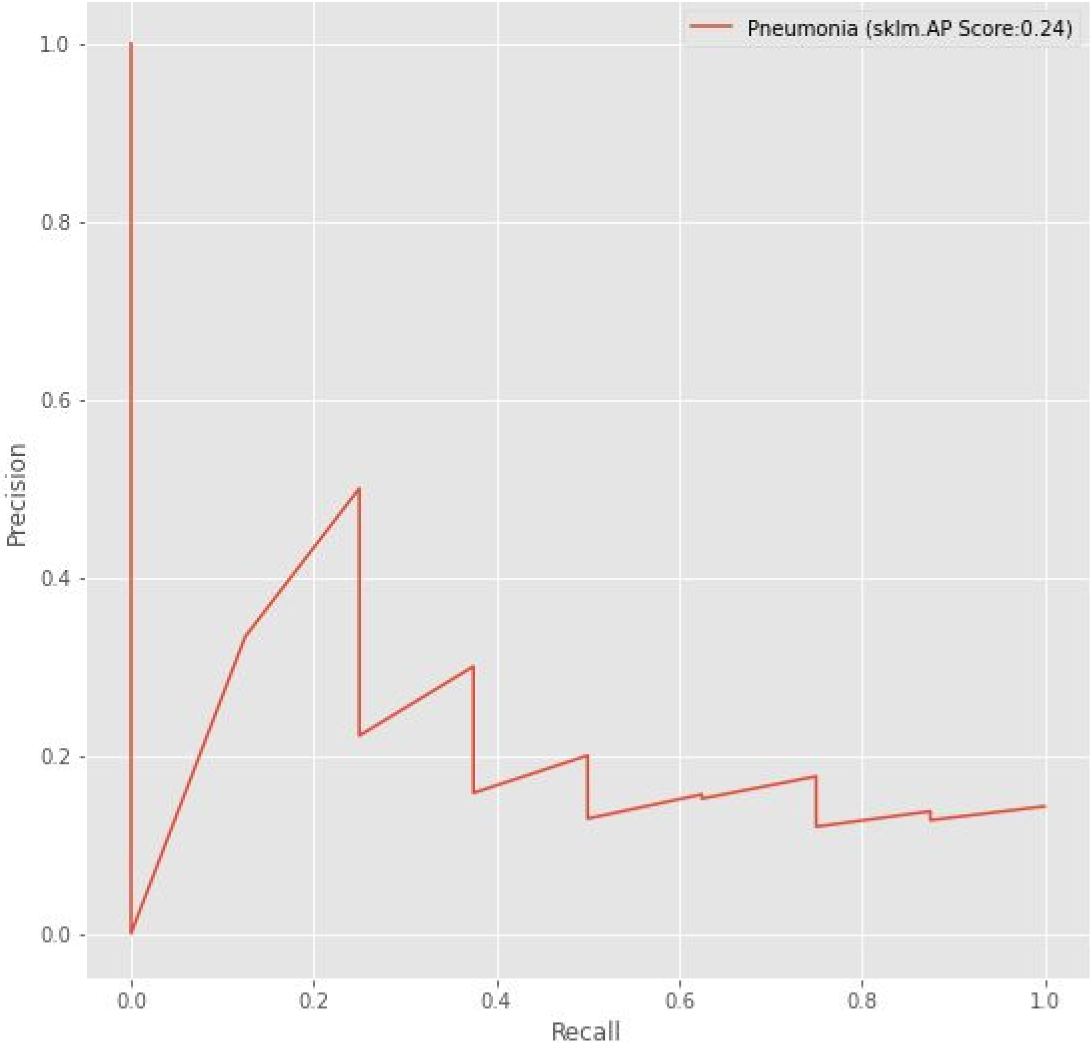
Precision Recall Curve:
The precision recall curve (Fig 3) illustrates the balance between precision and recall for every threshold between 0 and 1. As can be seen from the curve, thresholds around 0.2 offer better precision at the expense of recall, whereas thresholds beyond 0.5 offer better recall at the expense of precision.

This algorithm is intended to serve as an aid to the radiologist by tagging images that are more likely to be positive for a diagnosis of pneumonia. Thus we would like to select a threshold that will be less likely to miss potentially positive results (i.e. we want good recall), even if by doing so we are more likely to tag negative images as positive (a worse precision). We may thus sacrifice some precision for the sake of recall.

However, we do not want to sacrifice precision any more than necessary: an algorithm that tags nearly all images as positive would be of little use to the radiologist.

0.38 was selected as the optimal threshold. Out of a sample of 70 images, a threshold of 0.38 produced 33 false positives and only 1 false negative. False negatives are more concerning than false positives because they imply that a patient might not receive necessary treatment for pneumonia. Thus, despite the false positives, 0.38 appears to be a reasonable threshold.

Fig.3



4. Databases

The database contains 112,120 chest roentgenograms, of which 12.8% were tagged with a diagnosis of pneumonia. Fig. 4 illustrates the distribution of diagnoses across the database sample. Fig. 5 illustrates the frequency of concurrent diagnoses when one of the diagnoses is pneumonia. As can be seen, a diagnosis of pneumonia often overlaps with other diagnoses. (For clarification, see section on “Ground Truth,” below).

Most of the images had no diagnostic labels, i.e. they were labeled “No Finding.” Of the positive diagnoses shown in Fig. 4, pneumonia is not the most common diagnosis in the database. Note that the database distribution is not reflective of relative incidence of each diagnosis in the population as a whole. For example: in this database, there are more chest roentgenograms consistent with diagnosis of mass than pneumonia.

Age and gender may be confounding variables that can confuse the results of a classification algorithm like this one. Thus, age and gender sample distributions in images labelled pneumonia should mirror age and gender distributions within the database as a whole. Figs. 6 and 7 demonstrate age frequencies across the total population and across those with pneumonia, respectively; and Figs. 8 and 9 illustrate gender frequencies across the total population and across those with pneumonia, respectively. As can be seen, age and gender distributions for images labelled “pneumonia” resemble age and gender distributions of the database as a whole.

Fig.4

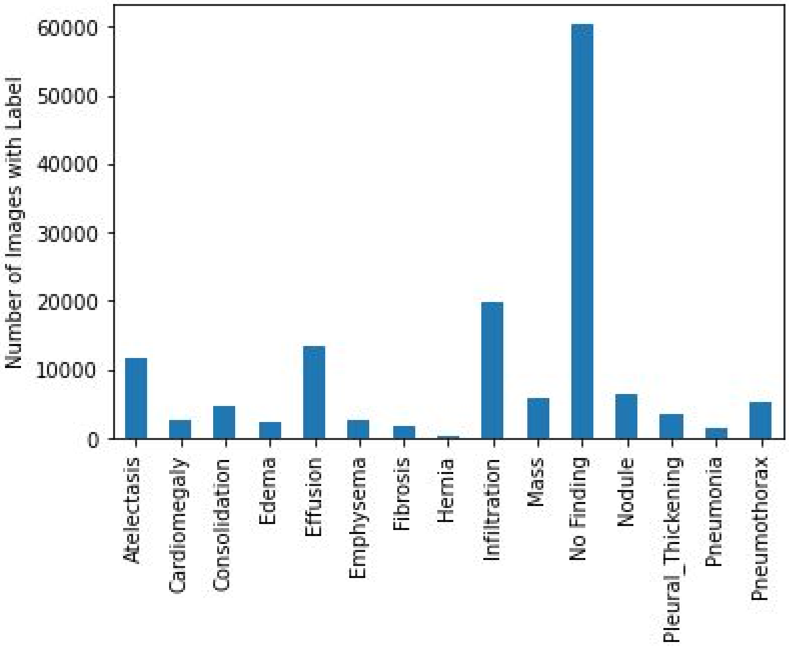


Fig.5

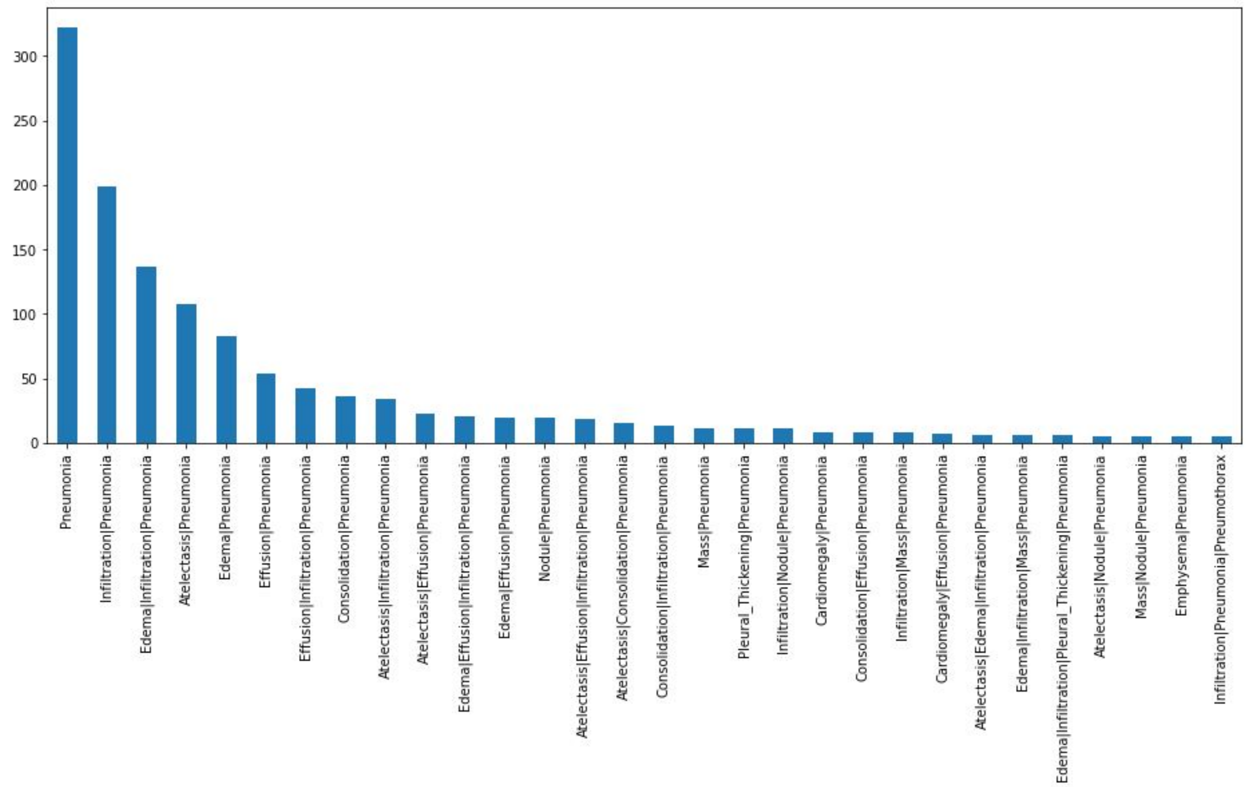

300
250
200
150
100
50
0
Pneumonia
Infiltration|Pneumonia
Edema|Infiltration|Pneumonia
Atelectasis|Pneumonia
Edema|Pneumonia
Effusion|Pneumonia
Effusion|Infiltration|Pneumonia
Consolidation|Pneumonia
Atelectasis|Infiltration|Pneumonia
Atelectasis|Effusion|Pneumonia
Edema|Effusion|Infiltration|Pneumonia
Edema|Effusion|Pneumonia
Nodule|Pneumonia
Atelectasis|Effusion|Infiltration|Pneumonia
Atelectasis|Consolidation|Pneumonia
Consolidation|Infiltration|Pneumonia
Mass|Pneumonia
Pleural_Thickening|Pneumonia
Infiltration|Nodule|Pneumonia
Cardiomegaly|Pneumonia
Consolidation|Effusion|Pneumonia
Infiltration|Mass|Pneumonia
Cardiomegaly|Effusion|Pneumonia
Atelectasis|Edema|Infiltration|Pneumonia
Edema|Infiltration|Mass|Pneumonia
Edema|Infiltration|Pleural_Thickening|Pneumonia
Atelectasis|Nodule|Pneumonia
Mass|Nodule|Pneumonia
Emphysema|Pneumonia
Infiltration|Pneumonia|Pneumothorax

Fig. 6

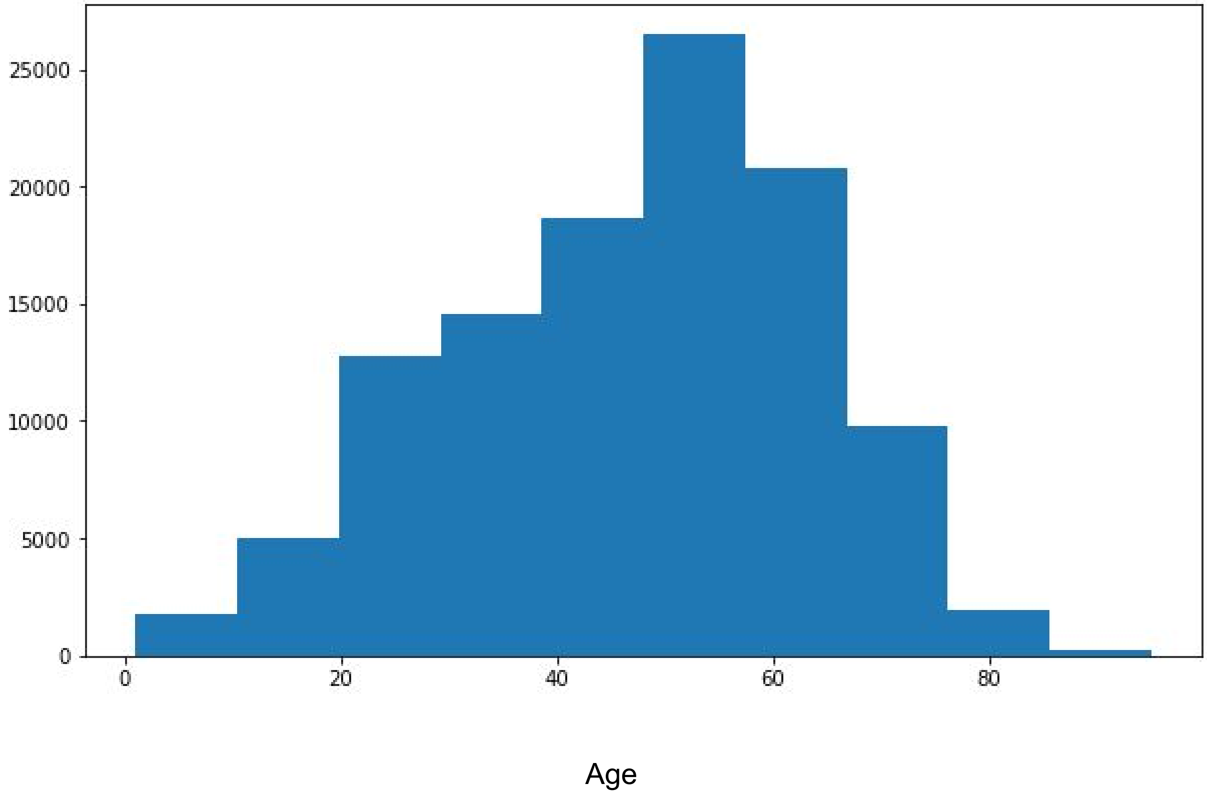

25000
20000
15000
10000
5000
0
0
20
40
60
80
Age

Fig. 7

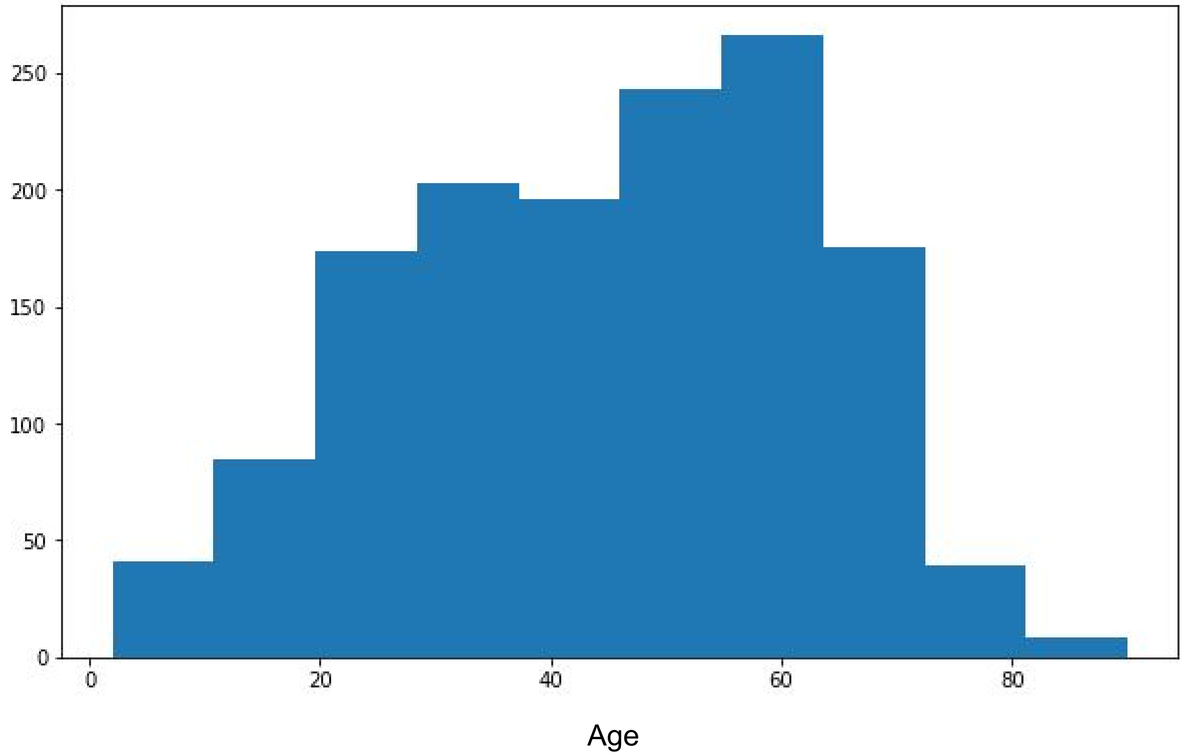

250
200
150
100
50
0
0
20
40
60
80
Age

Fig. 8

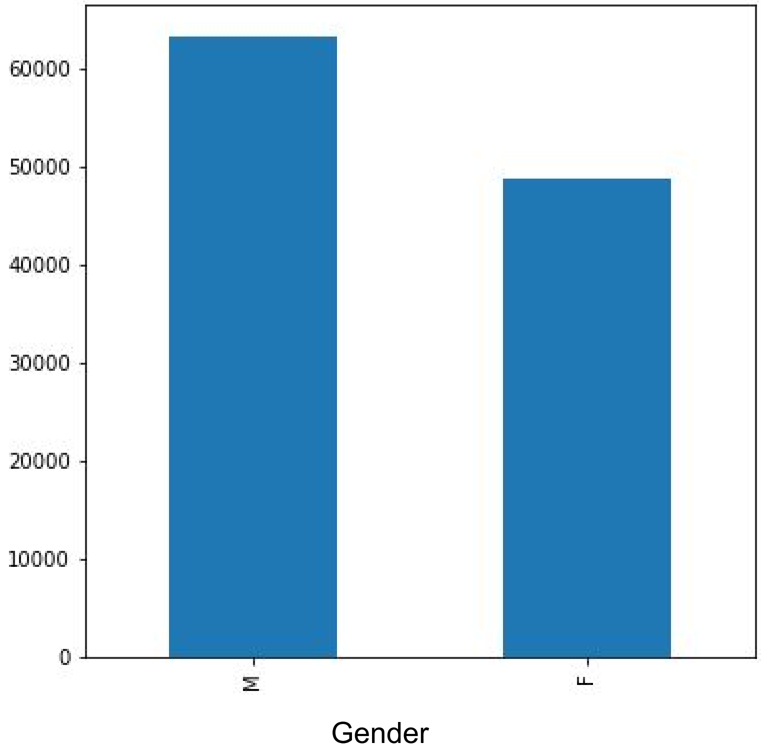
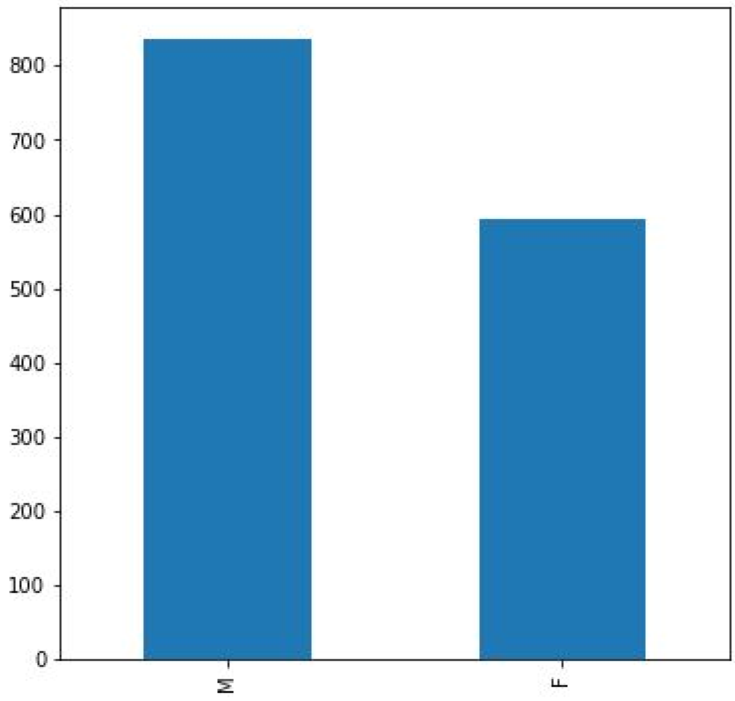

60000
50000
40000
30000
20000
10000
0
M
F
Gender

Fig. 9



Gender

Train test split: The data was divided into a train set comprising 80% of the data and a validation set comprising 20% of the data. Train and test sets were stratified such that distributions of various categories contained within the data (see Fig. 4) would be similar across both train and test sets. The validation set contained 286 images labeled positive for pneumonia and 1144 images labeled negative for pneumonia.

Description of Training Dataset: For the training set, images were randomly selected such that 50% were positive for pneumonia and 50% were negative for pneumonia. This selection of equal distributions allows for sufficient learning of both positive and negative images.

Description of Validation Dataset: For the validation set, it is desirable to select a distribution that mirrors that of the database as a whole somewhat more closely. Thus, for the validation set, images were randomly selected such that 20% were positive for pneumonia and 80% were negative for pneumonia.

5. Ground Truth

Diagnostic labels have been obtained using the ground truth of an expert radiology reading.

Note that the algorithm has been trained to distinguish between “Pneumonia” and those images that have not been tagged with a diagnosis of pneumonia. In this database (see Fig. 4), atelectasis, consolidation, and infiltration have been tagged as *different* diagnoses from that of pneumonia even though they are often used to indicate pneumonia or are often confused with pneumonia. Effusion, too, may be confused with pneumonia.

In practice, consolidation and infiltration are descriptive rather than diagnostic words, and these words are *often* used to describe pneumonia. A consolidation, for example, might be due to pus (i.e. pneumonia), sequestration, malignancy, or other causes. Not in all cases does the radiologist use the word “pneumonia” -- he may, for example, tag a film as “consolidation,” even though it is clear to both the radiologist and the primary care physician that the term “consolidation” refers to pneumonia. In practice the radiologist is often able to make this distinction using information not contained within the DICOM file. If the patient has a fever of 40 degrees C, shaking chills, and a cough, the “consolidation” is more likely to refer to pneumonia, even though the label does not say so; whereas if the patient has been asymptomatic since birth, the consolidation may refer to a sequestration, even though the label does not say so.

Similarly, radiologists often use the term “infiltrate” when they mean pneumonia. In such cases, a radiologist may never use the word “pneumonia” at all. However, an infiltrate does not always imply pneumonia: it might also be associated with tuberculosis, yeast (cryptococcus), amyloidosis, or other causes. The radiologist and primary care physician will be more likely to consider “infiltrate” to imply pneumonia if it is seen in the context of fever, cough, and an acute illness. As can be seen, the term “infiltrate” is the most common tag in this database. Thus this overlap -- between diagnoses of “pneumonia” and “infiltrate” -- is a significant problem!

Atelectasis is a common diagnosis, too, and pneumonia can look very similar on a chest roentgenogram, and in clinical practice these diagnoses are often confused. Atelectasis may also be present along with pneumonia. Effusion is similar in appearance to a viral pneumonia presentation.

In summary, the ground truth for the x-rays in the database, as noted, is a diagnosis by an expert radiologist. But radiologists often consider information that is not available in a DICOM file (medical history, etc.); and while radiologists do sometimes combine diagnoses explicitly (Fig.5), many radiologists do not always detail their assumptions in the written file.
For example, when a radiologist thinks “pneumonia” but writes “infiltrate” or “consolidation” instead of “pneumonia”, the algorithm reading the DICOM file sees only “infiltrate” or “consolidation” and assumes that the image has been tagged as negative for pneumonia. Erroneous learning is the result.

All of these diagnostic ambiguities inherent to the ground truth serve to reduce measured algorithm performance.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
A dataset will be collected using 500 frontal chest roentgenogram images as digital radiographs stored in DICOM files. The data will be collected from 500 adult patients, half of whom are male and half of whom are female. The FDA validation set will contain between 4% and 30% films

labelled positive for pneumonia as per the ground truth (see below) and the rest labelled negative for pneumonia.

While the FDA validation set need not mirror the population demographics exactly, it should provide ample opportunity for testing the algorithm on both men and women; and it should reflect the fact that most chest films in the population are not positive for pneumonia.

Regarding age, it is important to keep in mind that pneumonia may appear different in young children. A “round pneumonia” on chest xray is not common in adults but is common in children. Neonatal pneumonia often presents an interstitial appearance, whereas adult pneumonia appears more often as a focal infiltrate.

The Pneumonia Highlighter algorithm was trained using the NIH Chest xray dataset, in which the vast majority of adult chest roentgenograms were obtained from adults. (See Fig. 6 and Fig. 7). Thus it seems prudent to limit the FDA validation set to adults only.

Ground Truth Acquisition Methodology:
Each of the roentgenograms will be read by 4 radiologists and also by the Pneumonia Highlighter algorithm. In each case, the ground truth (positive or negative for pneumonia) will be established by the majority of radiologists.

The four radiologists will not have access to the patient’s medical history and will be limited to reporting pneumonia or not pneumonia. They will not report other findings or diagnoses. If a chest xray appears to be a pneumonia, then the radiologists will be instructed to report it as such. (Even though the radiologist might otherwise not be willing to report it as such on account of the absence of medical or clinical history. Thus, a consolidation or infiltrate with clinical suspicion of pneumonia should be read as pneumonia.)

After the radiologists have read the xrays and after the ground truth has been established for each xray, the dataset will be reviewed for proportion of films positive for pneumonia vs. proportion of films negative for pneumonia. If the number of films positive for pneumonia is greater than 30% of the total, films positive for pneumonia will be removed randomly until no more than 30% are positive for pneumonia. If the number of films positive for pneumonia is less than 4% of the total, films negative for pneumonia will be removed randomly until there are no less than 4% positive for pneumonia.

Algorithm Performance Standard:
Algorithm performance will be evaluated using recall and precision. These metrics obtained by applying the Pneumonia Highlighter to the FDA validation data will be compared to recall and precision obtained by each of the four radiologists on the same FDA validation data.