Intended use:

This algorithm is intended for use to assist radiologist to detect pneumonia in chest xray screening.

Indication for use:

This algorithm is intended for use to detect pneumonia in chest xray anteroposterior and posteroanterior views on gender male and female of all ages. Pneumonia may comorbid with 13 other diseases Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural_Thickening and Pneumothorax.

Algorithm Limitation:

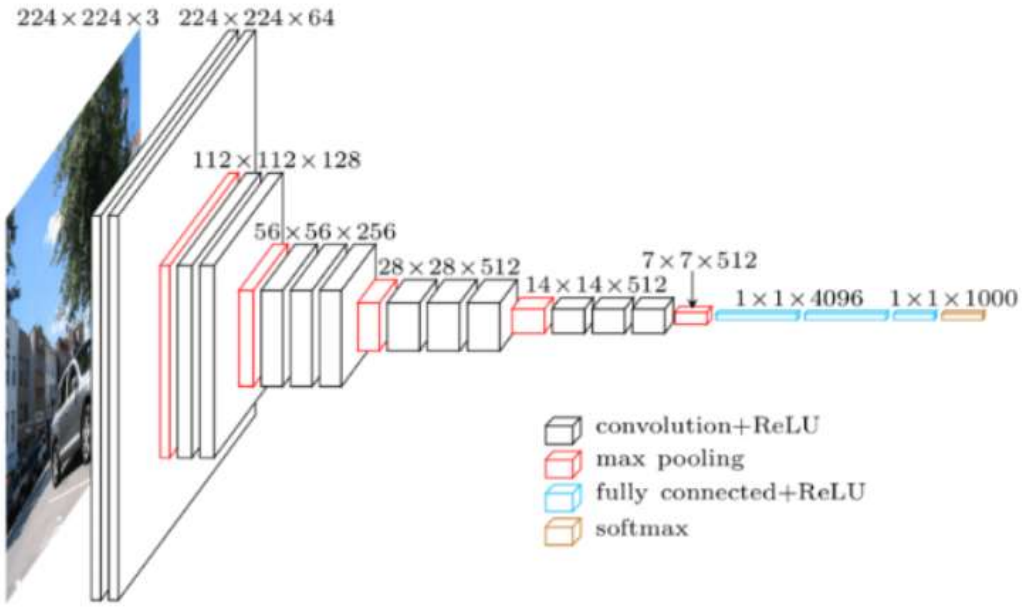
The results indicate that the presence of Effusion have a slight impact on the algorithm's sensitivity and may reduce the ability to detect pneumonia, while the presence of Infiltration and Nodule has a slight impact on specificity and may increase the number of false positive pneumonia classifications. This algorithm could also be used for workflow of re-prioritization in emergency room situation.

<u>Clinical impact of performance</u>

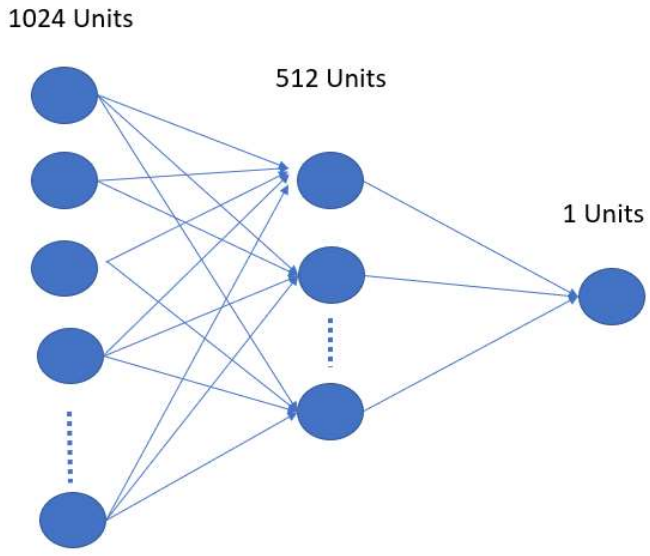
- Algorithm has False positive rate of 0.86. That is 86% of chest xray with normal conditions can be detected as pneumonia by algorithm. Final discretion will be up to radiologist to send for biopsy for confirmation.
- Algorithm has False negative rate of only 0.17. That is 17% of chest xray with penumonia can go undetected as normal. This tradeoff comes at the expense of more False positive but given higher recall 0.83 algorithm able to detect pneumonia in most of the positive cases. Final discretion will be up to radiologist to further access the accuracy of prediction carefully.
- Algorithm does not acheieve fast performance in absence of high compute GPU or cloud infrastructure. For faster inference high compute instances are required which are tyically hosted on cloud infrastructure.

Architecture diagram:

VGG16 Model archtitecture:



Classifier Fully Connected Layer:



DICOM checks

DICOM file is read using pydicom.dcmread to extract all header information. Below checks are done before preprocessing the image

- Image type is checked to ensure its Xray image and not CT Scan or MRI. This is done by checking the value of Modality attribute. For Xray images its value should be 'DX'.
- Patient Position that is view position is read from attribute PatientPosition. Its value should be either 'PA' or 'AP'.
- Body part is read from attribute BodyPartExamined. Its value should be 'CHEST' as we are examining the Chest Xray in this study.
- If all the above check are passed then image is returned else 'None' is returned.

Preprocessing steps

<u>During Inference</u>

Before runing inference on trained model, image in DICOM format needs to be preprocessed so that it can be converted into format acceptable by model to run inference.

Below steps or checks should be put in place to preprocess image.

- Image in DICOM format .dcm should be read using pydicom.dcmread method.
- Array of image pixel are then extracted using pixel_array method.
- Image is further resized to 224X224 dimension with 3 color channels.
- Image is normalized by subtracting mean and dividing it by standard deviation.
- Additional dimension is added to image as batch size.
- Finally image is then feed into model for inference.

<u>During Training</u>

- Training and Validation Images are rescaled to range between 0 and 1 by dividing each image pixel by 255. Standardizing the image size ensure faster convergence during training process.

- Images are resized to 224X224 to match the input dimension expected by VGG16 model.
- Image Augmentation technique is applied.
- Each batch of training data has 64 images whereas validation data is extracted into single batch of 572.

Architecture of the classifier

Transfer learning approach is used to train the deep learning model. Pretrained VGG16 model from Imagenet is loaded and initialized with pretrained weights. VGG16 CNN model does the feature extraction which is then feed into fully connected classifier layer.

Classifier layer comprise of three layers with 1024 units in first layer, 512 units in second layer and output layer with 1 unit to predict probability closer to 0 or 1. Relu activation function is used in hidden layer and sigmoid activation is applied to output unit to predict binary outcome. Dropout layer is added between hidden layers to reduce overfitting.

Multi-dimensional output from VGG16 is flatten before feeding it to FC layer. All layers of network are freezed except the last two layers of VGG16 and FC layers. This is done so that last few layers of VGG16 network could learn complex intricacies in xray image which will be better feature extractor when feed to FC layer.

Image Augmentation

Training and Validation images are augmented with below processing steps. These augmentations are applied keeping in mind how they could appear in the real-world setting.

- Horizontal flip is applied to training images but not vertical as we don't expect the chest xray images to be upside down.
- I applied random rotation of up to 5 degrees as we don't expect chect xrays to have wildly different orientations.
- Then I applied height and width shift by 0.05 as different view positions and image capture can cause shift.

- Zoom range is set to 0.2 as in clinical setting, we may expect xray image captured with different zoom settings.
- Shear range is set to 0.1.

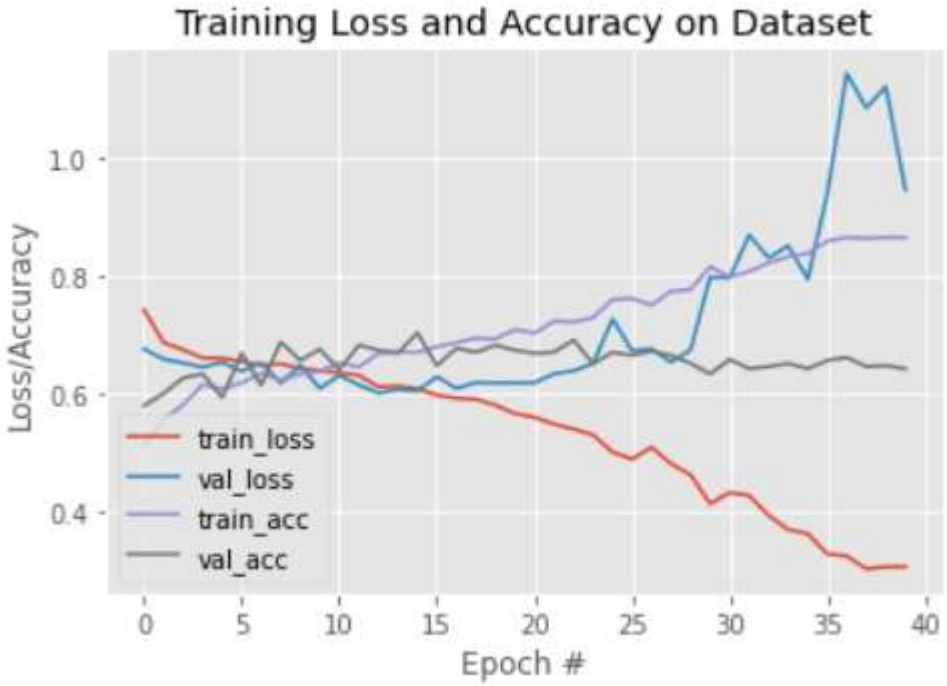
Parameters are used for training

Below parameters are used for training process.

- Optimizer – Adam optimizer with learning rate 0.0001.
- Loss – Since its binary classification task Loss function used is binary_crossentropy.
- Evaluation metric used for model is binary_accuracy and the same metric is used to monitor the performance of model while training.
- EarlyStopping with patience of 10 is set to avoid overfitting and stop the training process if binary accuracy does not improve for more than 10 epochs.
- Network is trained for 40 epochs.

Training and Validating loss

From the plot history we can observe that train loss continues to decrease with each epoch. Training Loss curve gradually goes down and it ultimately plateaus after epoch 37. Here we can safely assume that convergence has been almost reached. Validation loss gradually decreases until epoch 14 and then it gradually starts increasing depicting the overfitting scenario here. Network is getting better on training images but its not generalizing well on validation images.



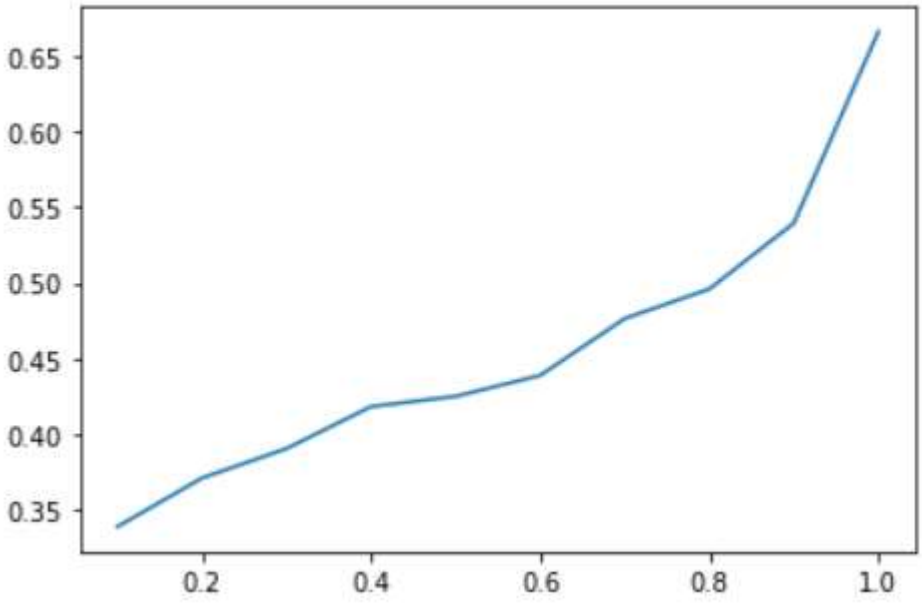
Performance statistics and threshold used in final validation

Trained algorithm achieved the AUC score of 0.82 and AP score of 0.79.
F1 score is evaluated at threshold between range 0 to 1 at the increment of 0.1 for 11 values of threshold. After this F1 score is plotted against threshold as shown below. From the curve we can observe that F1score is maximum at threshold of 0.99.

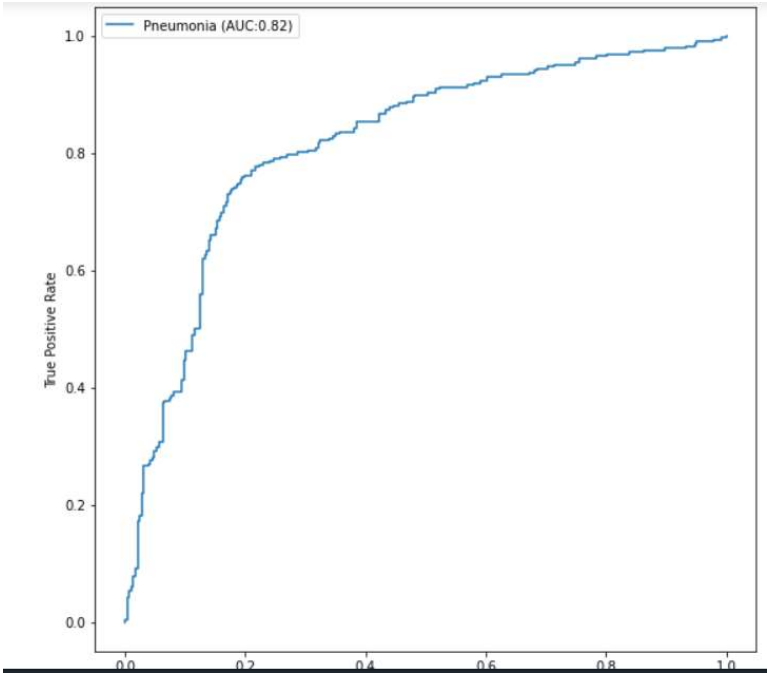
With threshold of 0.99, each of metric Precision, Recall, F1 score, Sensitivity, Specificity, False positive rate and False negative rate are evaluated. Based on these metrics we can conclude algorithm is good at detecting pneumonia in clinical settings with only 17% false negative rate (recall=0.83) at the tradeoff of lower precision and higher false positive rate of 86%.

<u>Evaluated Metrics</u>
precision: 0.4906832298136646
recall: 0.8286713286713286
f_score: 0.6163849154746424
Sensitivity: 0.8286713286713286

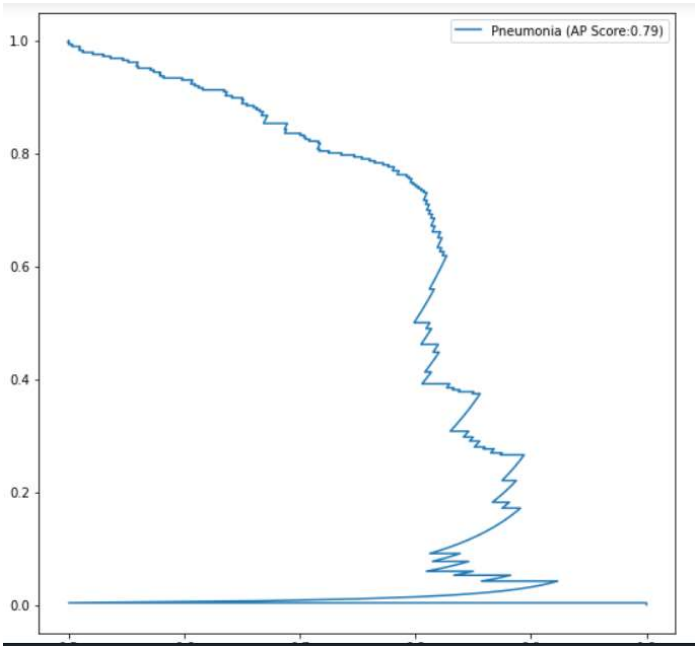
Specificity: 0.13986013986013987
False positive rate: 0.8601398601398601
False negative rate: 0.17132867132867133



F1 score on Y-axis vs threshold on X-axis



TPR vs FPR curve

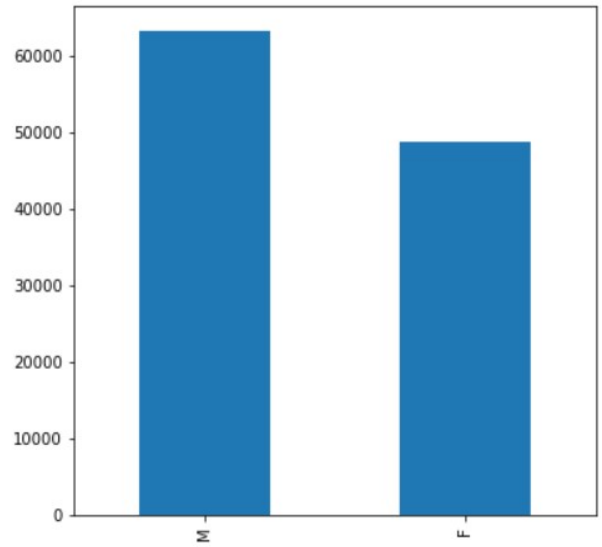


Precision vs Recall Curve

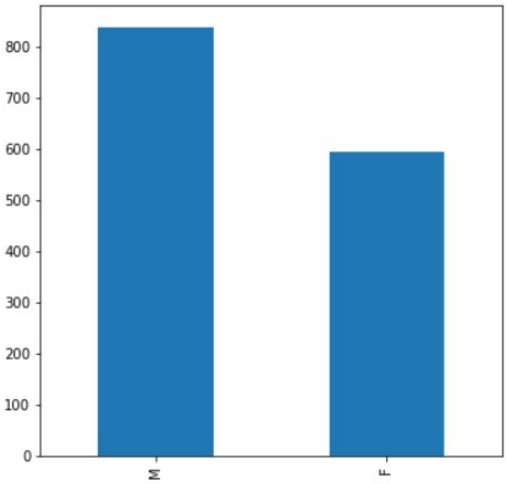
Training set

Percentage of pneumonia detected in the entire dataset is just 1.27% that is 1431 pneumonia positive images compared to 110689 images without pneumonia in NIH dataset which indicates there is a strong class imbalance in the dataset. Training data is created to include equal distribution of positive or negative pneumonia images. There are 1145 pneumonia positive images in training set and the same number of 1145 images without pneumonia are randomly sampled from 110689 images. That makes total number of images in training dataset equal to 2290.

Distribution of gender Male and Female is almost similar for images with detected Pneumonia and across whole population of data as shown in below plots. So, training and Validation set will hold similar distribution of gender.

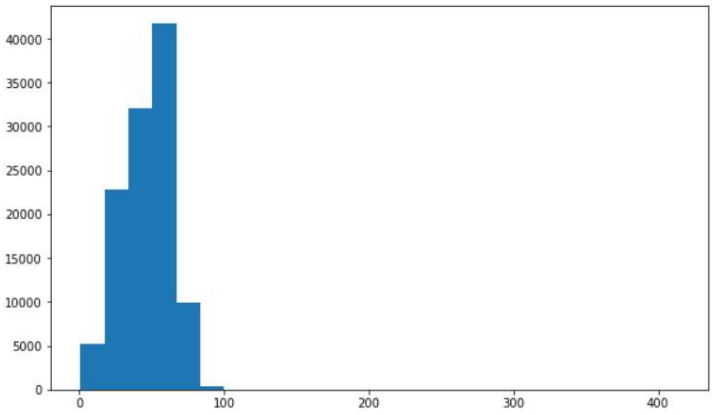


Distribution of gender across whole population

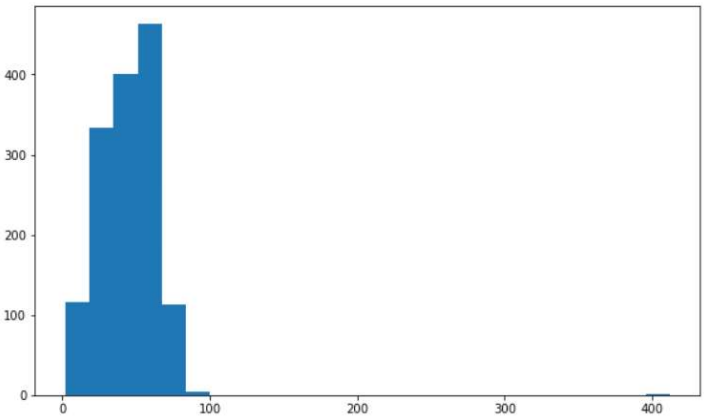


Distribution of gender with pneumonia positives.

Distribution of Age shows similarity when compared across images with Pneumonia positive and whole population as shown below. So we training and validation dataset will hold similar distribution.

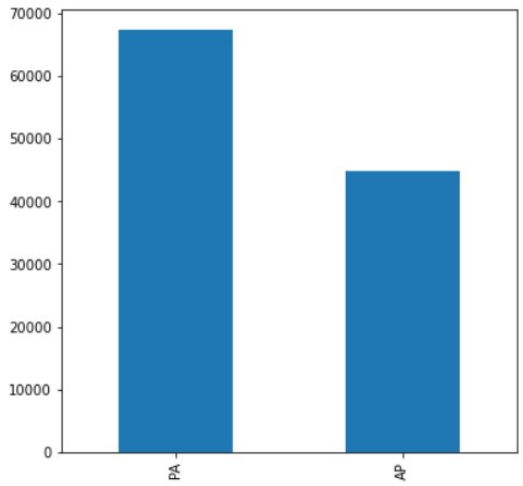


Distribution of age across whole population

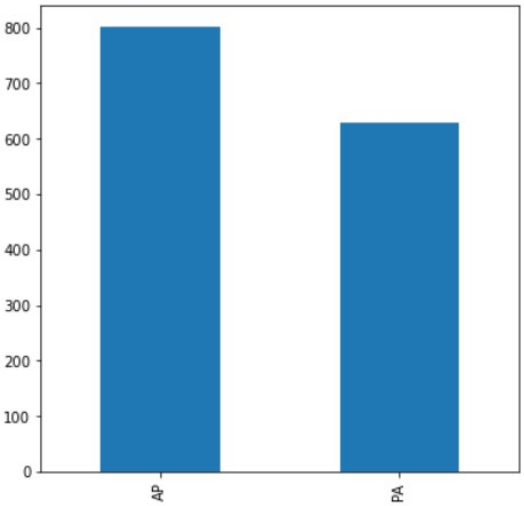


Distribution of age across images with pneumonia positives.

Distribution of view position across whole population is similar to distribution across images with pneumonia positive, so training and validation dataset will hold similar distribution.



Distribution of view position across whole population.



Distribution of view position across images with pneumonia positives.

Validating set

To prepare the Validation data I have used sklearn function train_test_split such that 80% of pneumonia positive cases are in training data and remaining 20% of positive images are in Validation dataset. So, the validation data has 286 positive images. In validation data also we want to keep the same distribution of images as it exists in training data. So negative images are randomly sampled in equal proportion which is 286 in count making total count in validation data as 572.

As discussed in prior section, distribution of age, gender and view position will be almost similar in Validation data and Training data as stratified split is used to prepared datasets with equal number of positive and negative samples.

Ground truth of the NIH dataset

This NIH Chest X-ray Dataset is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. These labels were inferred through natural language processing by mining disease classification from the associated radiological reports and are estimated to be at least 90% accurate. For the sake of this project, we treat the labels as ground truth for the purpose of classification.

Benefits:

- This process of labeling ground truth is robust enough for this case study as it achieves 90% accuracy

- Process of labeling using NLP technique is much faster compared to manual process where each image is labeled by radiologist for set of 112120 image which may take months to complete and validate.

Limitations:

- In ideal case study 100% accuracy is desired for ground truth as it validates the full authenticity of building the algorithm which predicts outcome closer to true outcome.

- Applying NLP technique requires lot of compute and memory resources which may pose constraint to hardware resource requirement.

FDA Validation Dataset

To validate the authenticity of algorithm performance on real world clinical, Production data will be requested from clinical partners considering below factors:

- Chest xray images will be collected for gender male and female of all ages.
- Patient may have pneumonia comorbid with 13 other diseases Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural_Thickening and Pneumothorax.

- This data is intended to detect pneumonia so data will be collected for body part examined “CHEST”.
- Type of image data will be in form of Xrays and will not include any MRI or CT scans.
- Collected Xrays data captured should be either anteroposterior or posteroanterior view.

Ground truth for FDA Validation Dataset

This algorithm will assist radiologist to detect suspicious pneumonia in Check Xray. Best approach to establish ground truth will be to have group of radiologists to label each image and select the ground truth based on most votes each label receives. So, the algorithm performance will be measured on this Silver Standard approach towards labeling ground truth.

Performance metric

Algorithm was evaluated and compared against the following Performance Standard.

Performance Standard:

Based on research paper CheXNet (Stanford ML Group), performance of individual radiologist was evaluated by using majority voting of the other 3 radiologists as ground truth. F1 score for each radiologist is computed as shown below.

	F1 score
Radiologist 1	0.383
Radiologist 2	0.356
Radiologist 3	0.365
Radiologist 4	0.442
Radiologist Avg	0.387
Algorithm	0.616

Our algorithm F1 score of 0.616 is significantly better than Radiologist average score of 0.387.