FDA SUBMISSION

Your Name: INDRANUJ GANGAN

Name of your Device: PXRAY-DGX

Algorithm Description

1] INTENDED USE: The device is intended to be used as an assisting pneumonia diagnostic software tool for radiologists

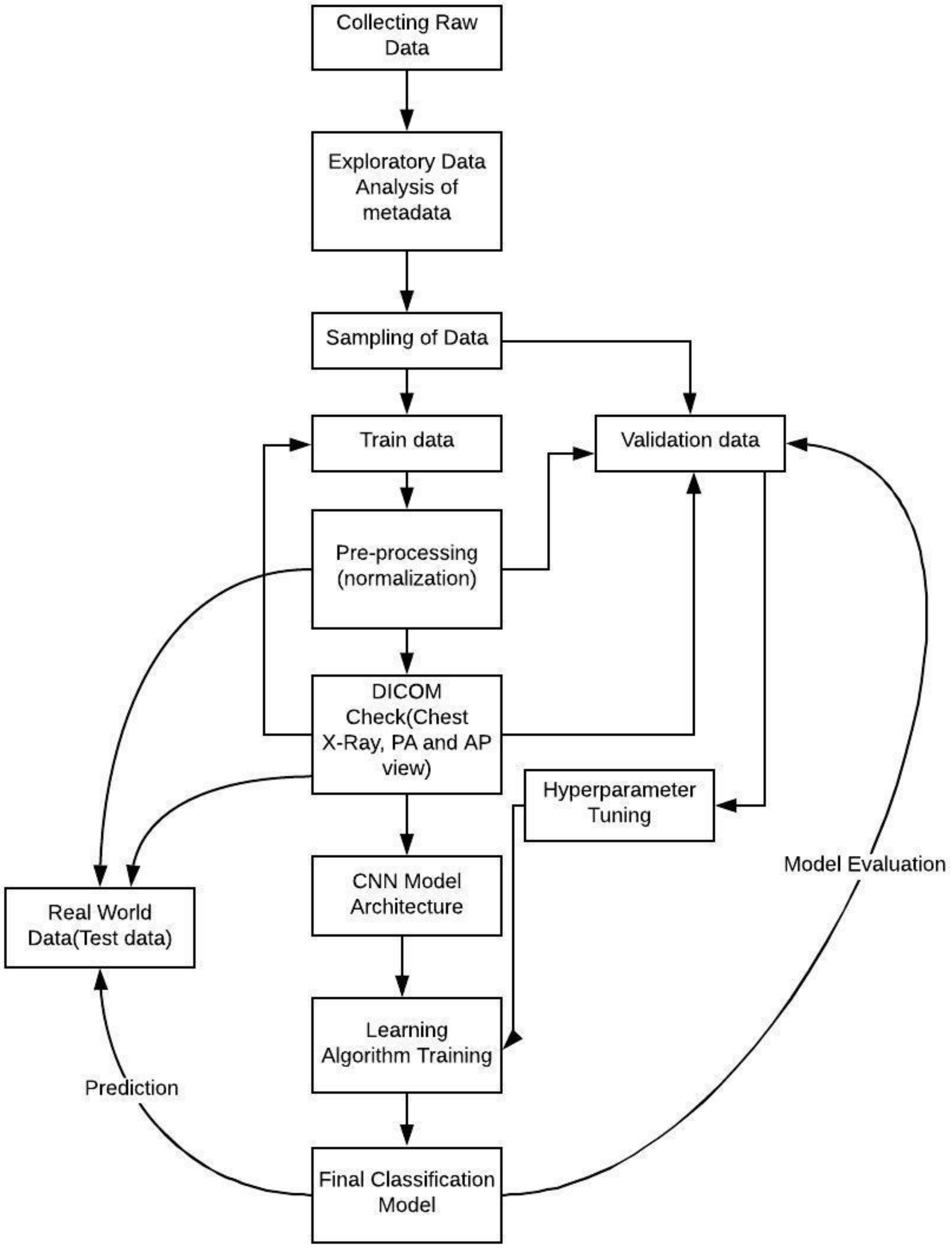
2] INDICATIONS OF USE: The device will be used to help/assist radiologists in diagnosing pneumonia using chest X-Ray with higher assurance. The device is suitable for population of 10-80 years of age. It will be used for both male and female

3] There is high co-occurrence of infiltration and edema (both have similar intensity profiles with pneumonia) with pneumonia and hence might be a algorithmic limitation while detecting.

4] Clinical Impact: The device will have significant impact on reducing false positive and false negatives and hence help enhance the diagnostic process. Impact of false positive is that patient will have to spend unnecessary amount on treatment and also there can be side effects of medication while in case of false negative it might lead to death of patient if diagnosis is delayed.

Algorithm Design and Function

1] Flowchart:


Collecting Raw Data
Exploratory Data Analysis of metadata
Sampling of Data
Train data
Validation data
Pre-processing (normalization)
DICOM Check(Chest X-Ray, PA and AP view)
Hyperparameter Tuning
Model Evaluation
Real World Data(Test data)
CNN Model Architecture
Learning Algorithm Training
Prediction
Final Classification Model

2] DICOM Check conditions: It is applied to train, validation and real world data as shown by arrows in flowchart

- PA and AP view position
- Body part examined: Chest
- Modality: X-Ray('DX')

3] Pre-processing steps: The images are normalized and resized to (1,224,224,3). It is applied to train, validation and real-world data as shown by arrows in flowchart

4] Performance standards: Model with least validation loss is selected. Later a threshold value is finalized for which F1 score is maximum from F1 score-Threshold graph as shown below:

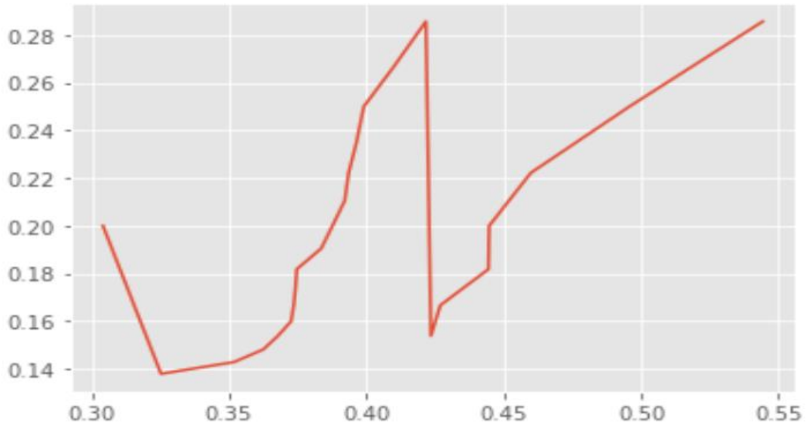


FIG 1: F1 score-Threshold graph

The finalized threshold is 0.42 at which F1 score is 0.28

5] Algorithm steps

- Collected NIH Chest X-Ray data consisting of 112,112 images from 30805 patients. The images used were in png format. The dataset had labels for 14 diseases.
- Implemented train-test split with train data having 50% pneumonia and 50% Non-pneumonia cases so that training can be done on balanced data. The test set had 20% pneumonia and 80% Non-pneumonia cases to reflect real world setting.
- The images are normalized
- The weights of the network are initialized with weights from a model pretrained on ImageNet (VGG16 model) and only last layer of it is trained by adding different layers.
- The network is trained end-to-end using Adam. We train the model using minibatches of size 16. We use an initial learning rate of 0.0001 that is decayed by a factor of 10 each time the validation loss plateaus after an epoch, and pick the model with the lowest validation loss.

- The model selected gave F1 score of 0.28 at threshold of 0.42 with AUC of around 0.59

- CNN Architecture for the best model:

```
new_model = Sequential()

# Add the convolutional part of the VGG16 model from above.
new_model.add(vgg_model)

# Flatten the output of the VGG16 model because it is from a
# convolutional layer.
new_model.add(Flatten())

# Add a dropout-layer which may prevent overfitting and
# improve generalization ability to unseen data e.g. the test-set.
new_model.add(Dropout(0.25, seed=1))

# Add a dense (aka. fully-connected) layer.
# This is for combining features that the VGG16 model has
# recognized in the image.
new_model.add(Dense(1024, activation='relu'))

# Add a dropout-layer which may prevent overfitting and
# improve generalization ability to unseen data e.g. the test-set.
new_model.add(Dropout(0.25, seed=1))

# Add a dense (aka. fully-connected) layer.
# This is for combining features that the VGG16 model has
# recognized in the image.
new_model.add(Dense(512, activation='relu'))

# Add a dropout-layer which may prevent overfitting and
# improve generalization ability to unseen data e.g. the test-set.
new_model.add(Dropout(0.25, seed=1))

# Add a dense (aka. fully-connected) layer.
# This is for combining features that the VGG16 model has
# recognized in the image.
new_model.add(Dense(256, activation='relu'))

# Add a dense (aka. fully-connected) layer.
# Change the activation function to sigmoid
# so output of the last layer is in the range of [0,1]
new_model.add(Dense(1, activation='sigmoid'))
```

Algorithm Training

1] Image Augmentation Parameters: rescale=1. / 255.0,
horizontal_flip = **True**,
height_shift_range= 0.1,
width_shift_range=0.1,
rotation_range=20,
shear_range = 0.1,
zoom_range=0.1)

2]Batch size: Train data batch size = 20; Validation data batch size = 32

3] Optimizer: Adam

4] Learning rate = 0.0001 and decay of 0.1

5] All layers except the last layer of VGG16 models were frozen. New 4 dense layers were added with dropout of 0.25. To this a sigmoid activation layer was added.

6] The model selected gave F1 score of 0.28 at threshold of 0.42 with AUC of around 0.59 and the precision and recall at this threshold were quite balanced.

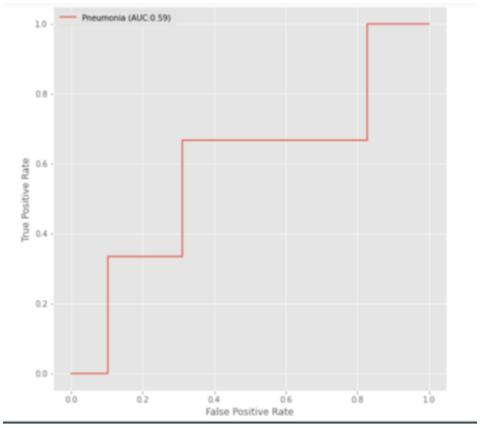


FIG 2: ROC CURVE

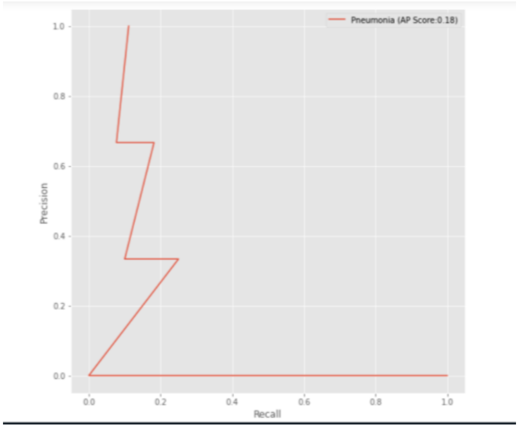


FIG 3: PRECISION-RECALL CURVE

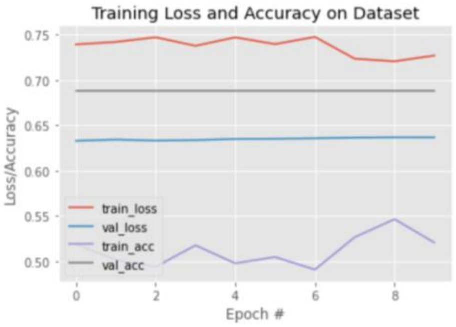


FIG 4: Train loss, Validation loss, Train accuracy , Validation accuracy plot

Databases

1] Training and Validation: We use the ChestX-ray14 dataset released by Wang et al. (2017) which contains 112,120 frontal-view X-ray images of 30,805 unique patients. Wang et al. (2017) annotate each image with up to 14 different thoracic pathology labels using automatic extraction methods on radiology reports. We label images that have pneumonia as one of the annotated pathologies as positive examples and label all other images as negative examples. There were total 1431 images of pneumonia. For the pneumonia detection task, we randomly split the dataset into 80% training data and 20% validation data stratified on pneumonia (target variable) and then balanced the training data it to get 1:1 ration of positive and negative cases (1008 images) by down sampling the negative class. For validation data 1:4 ratio of positive to negative cases (630 images) was obtained.
For training and validation dataset creation chest as examined part and modality ‘DX’(X-Ray) was considered.

2] Test(Real World data): It consists of 6 images derived from DICOM files. It is passes through DICOM check to see if all conditions are satisfied (View position: AP or PA view, body part examined: chest, Modality: ‘DX’). It undergoes pre-processing steps used during training

3] Co-occurrence with diseases: Infiltration, Edema, Atelectasis, Effusion

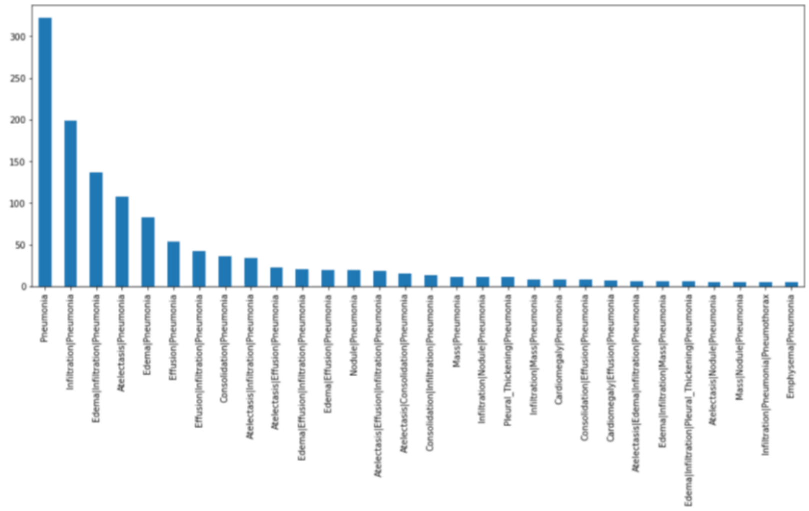


FIG 5: Co-occurrence of pneumonia with other diseases

4] Patient age distribution:

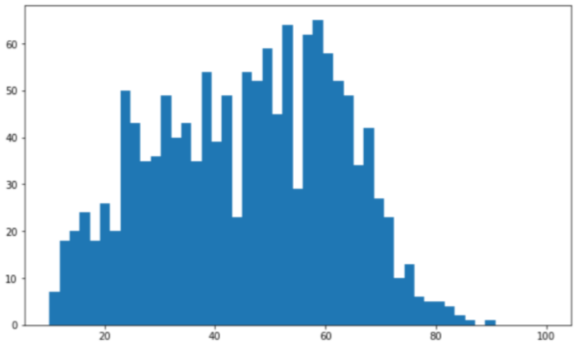


FIG 6: Age distribution for pneumonia patients

5] Pateint Gender distribution:

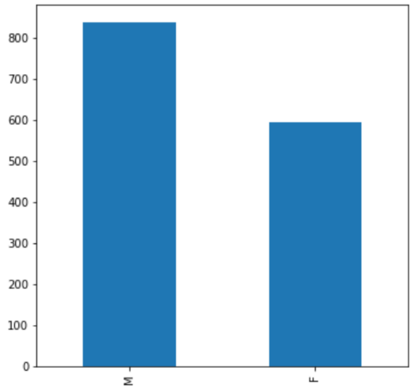


FIG 7: Gender distribution for pneumonia patients

Ground Truth

To create these labels, the authors 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 original radiology reports are not publicly available but you can find more details on the labeling process in this Open Access paper: "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." (*Wang et al.*)

Limitation: In case of pneumonia ultimate gold standard is biopsy results for better performance. In case of NLP text mining NLP algorithm can face problems which retrieving information have complex vocabulary ,for example: ‘The image is unlikely to have pneumonia’. NLP algorithm might fail to pick negation conveyed by word unlikely

Benefit: NLP text mining is a faster and convenient approach of annotating to train deep learning models as getting biopsy results of patients In cohort is next to impossible

FDA Validation Plan

A] General Information

1] Age Range: 10-80 years

2] Sex: Male and Female both

3] Type of imaging modality: X-Ray

4] Body part imaged: CHEST

5] Prevalence of disease of interest: 20%

6] Any other diseases that should be included *or* excluded as comorbidities in the population: Infiltration and Edema can be excluded because of its high co-occurrence with pneumonia. These both have similar intensity profile with that of pneumonia.

B] Ground Truth Acquisition Methodology: In this particular use case a silver standard approach will be used where in multiple radiologists will annotate test set images and then mode will be taken for final annotation. This method is faster compared to gold standard method like biopsy which takes very long and patient needs to be put on medications as soon as possible.

C] Algorithm Performance Standard: F1 score is the performance metric considered because it takes into consideration both precision and recall and hence optimize false positives and false negatives. It is harmonic mean of precision and recall. The F1 score is used to measure a test's accuracy, and it balances the use of precision and recall to do it. The F score can provide a more realistic measure of a test's performance by using both precision and recall**.** Please refer the following paper: https://arxiv.org/pdf/1711.05225.pdf