Pneumonia Detection from Chest X-rays – FDA Submission

Dhar Rawal dharrawal@gmail.com 281-995-6423 02/23/2022

Udacity – AI for Healthcare Program

Intended Use Statement

This machine learning algorithm is designed for assisting the radiologist in screening for pneumonia on a chest X-ray. It is intended for professional use.

Indications for Use

- Screening chest X-rays with view positions 'PA' or 'AP' only

Limitations

This algorithm has an AUC of 0.7 and a maximum F1 score of 0.07. Given the relatively low AUC and F1 scores, this algorithm is not a reliable indicator for a positive diagnosis and should be used only as an aid to screening and for ruling out the presence of pneumonia.

- A false positive might result in the patient having to undergo additional tests/biopsies with potentially high costs and possibly unnecessary treatments. In case of a misdiagnosis, the patient may end up getting treatments unrelated to the true condition, which could result in further deterioration of their health.
- A false negative could result in the patient not getting treatment on time and put their life at risk.

Below conditions occurred frequently in conjunction with Pneumonia in the training X-rays and could cause a misdiagnosis:

Conditions	Comorbidity with Pneumonia
Infiltration	42%
Edema	24%
Effusion	19%
Atelectasis	18%

Radiologists must be careful to check for the presence of these other conditions OR a false positive in case of a positive finding.

Model and Classifier Architecture

Layer	Description
VGG16	Truncated at block5_pool layer. All except last convolutional layer are frozen
Flatten	Flatten the output from VGG16
Dropout	0.5
Dense	256 units, relu
Dense	1 unit, sigmoid

DICOM checks

Before sending the data to the model for prediction, the following DICOM fields are checked:

- BodyPartExamined – must be CHEST
- Modality – must be DX (Digital radiography X-ray)
- PatientPosition – must be PA or AP (Anterior/Posterior or Posterior/Anterior)

Image preprocessing

All images are passed through the Keras ImageDataGenerator and throught the preprocess_input function for VGG 16. This converts the images from RGB to BGR and then centers each color channel's values around zero by subtracting the mean. No scaling is performed.

Data augmentation

Additional images are added to the training set by augmenting supplied images as follows:

- Images may be horizontally flipped
- Height and width are randomly stretched by no more than 10%
- Images are rotated randomly by no more than 10%
- Images may be randomly sheared by no more than 10%
- Images may be randomly magnified by no more than 10%

Note that no augmentation is performed on the validation set

Training parameters

Training used the following parameters

- Adam optimizer
- Learning rate of 1e-4
- Loss metric is binary crossentropy
- 10 epochs
- Training batch size is 32
- Weights are saved after an epoch if validation loss is minimum
- Validation batch size is 128 and validation data is randomly generated for each epoch

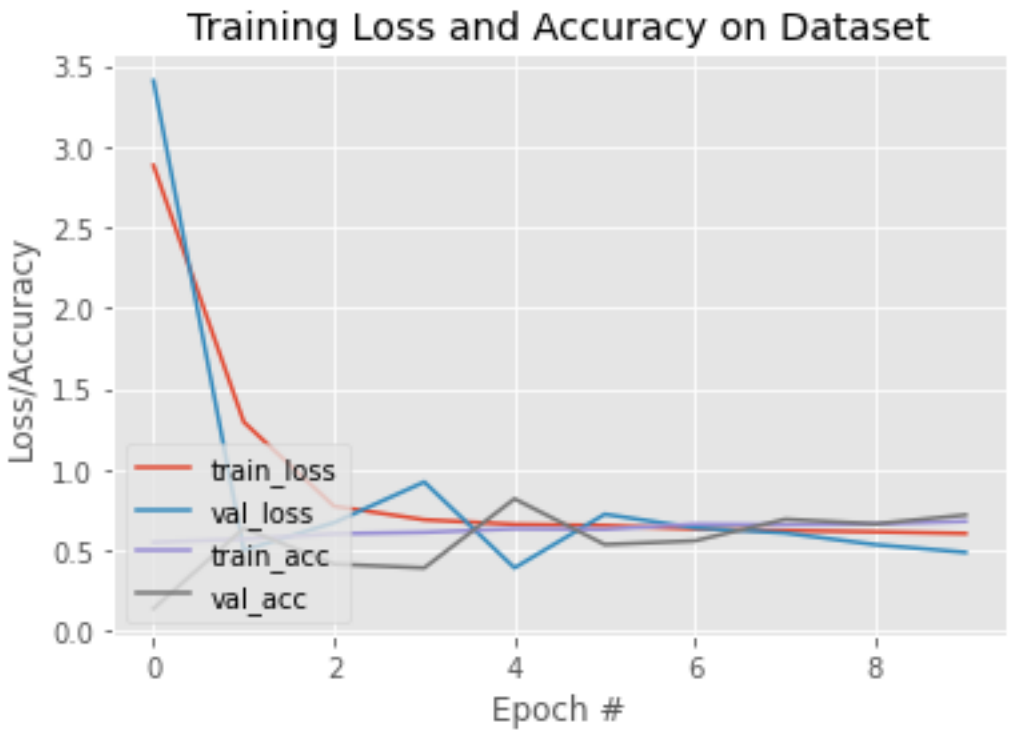
Training and validation loss behavior

As can be seen from the graph and printout below, validation loss rapidly decreases in the first 2 epochs and then bounces until epoch 5 after which it seems to stabilize

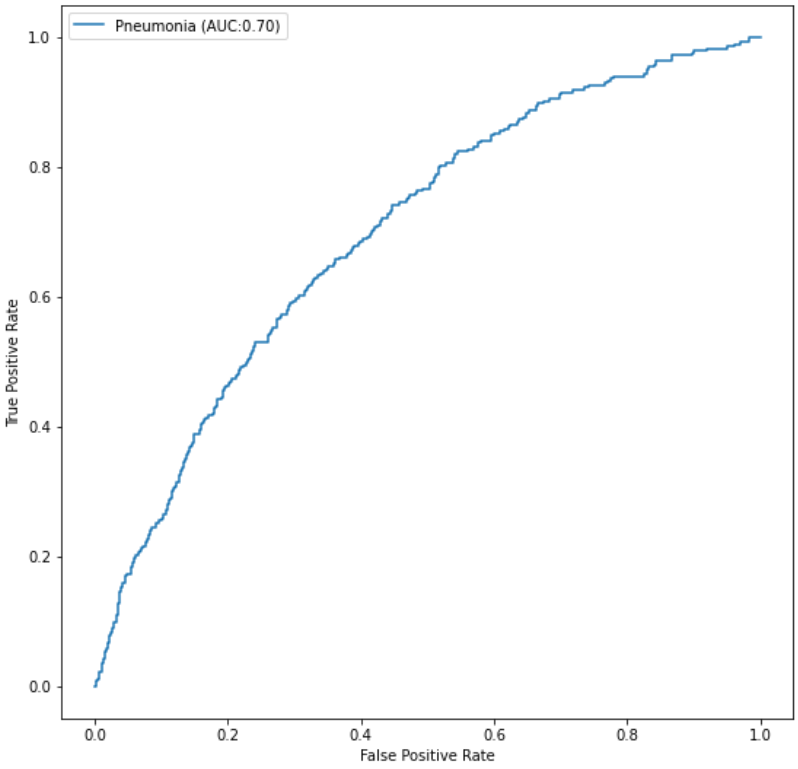
```
553467904/553467096 [==============================] - 8s 0us/step
Found 22408 validated image filenames belonging to 2 classes.
Epoch 1/10
72/72 [==============================] - 450s 6s/step - loss: 2.8812 - binary
_accuracy: 0.5493 - val_loss: 3.4059 - val_binary_accuracy: 0.1389
```

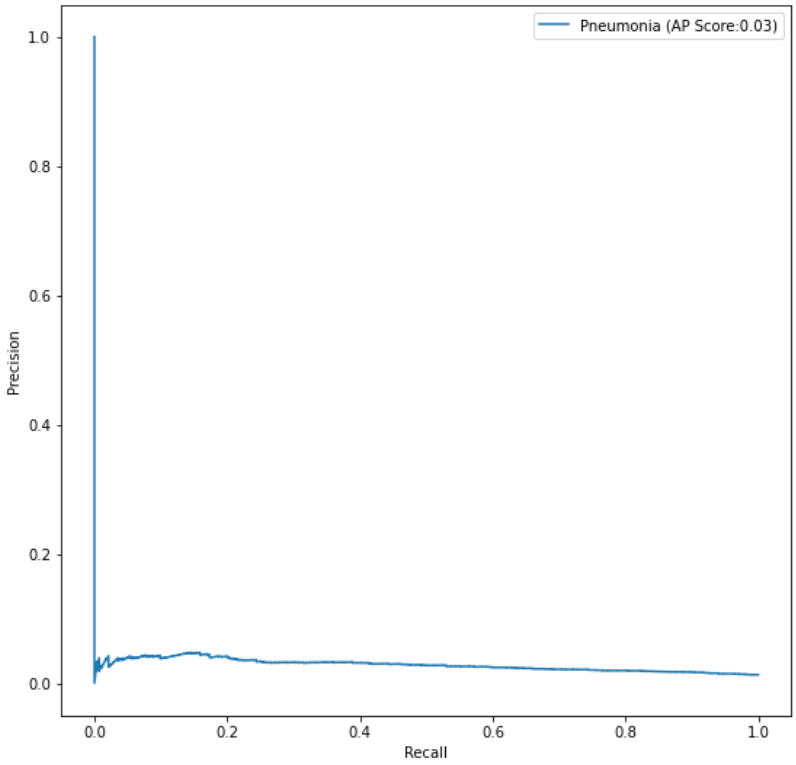
```
Epoch 00001: val_loss improved from inf to 3.40589, saving model to xray_clas
s_my_model.best.hdf5
Epoch 2/10
72/72 [==============================] - 393s 5s/step - loss: 1.2994 - binary
_accuracy: 0.5690 - val_loss: 0.4999 - val_binary_accuracy: 0.6356

Epoch 00002: val_loss improved from 3.40589 to 0.49994, saving model to xray_
class_my_model.best.hdf5
Epoch 3/10
72/72 [==============================] - 389s 5s/step - loss: 0.7736 - binary
_accuracy: 0.6017 - val_loss: 0.6709 - val_binary_accuracy: 0.4144

Epoch 00003: val_loss did not improve from 0.49994
Epoch 4/10
72/72 [==============================] - 394s 5s/step - loss: 0.6884 - binary
_accuracy: 0.6109 - val_loss: 0.9226 - val_binary_accuracy: 0.3906

Epoch 00004: val_loss did not improve from 0.49994
Epoch 5/10
72/72 [==============================] - 387s 5s/step - loss: 0.6599 - binary
_accuracy: 0.6328 - val_loss: 0.3909 - val_binary_accuracy: 0.8184

Epoch 00005: val_loss improved from 0.49994 to 0.39090, saving model to xray_
class_my_model.best.hdf5
Epoch 6/10
72/72 [==============================] - 375s 5s/step - loss: 0.6535 - binary
_accuracy: 0.6293 - val_loss: 0.7231 - val_binary_accuracy: 0.5354

Epoch 00006: val_loss did not improve from 0.39090
Epoch 7/10
72/72 [==============================] - 390s 5s/step - loss: 0.6334 - binary
_accuracy: 0.6616 - val_loss: 0.6388 - val_binary_accuracy: 0.5584

Epoch 00007: val_loss did not improve from 0.39090
Epoch 8/10
72/72 [==============================] - 386s 5s/step - loss: 0.6226 - binary
_accuracy: 0.6581 - val_loss: 0.6066 - val_binary_accuracy: 0.6918

Epoch 00008: val_loss did not improve from 0.39090
Epoch 9/10
72/72 [==============================] - 385s 5s/step - loss: 0.6142 - binary
_accuracy: 0.6620 - val_loss: 0.5346 - val_binary_accuracy: 0.6632

Epoch 00009: val_loss did not improve from 0.39090
```

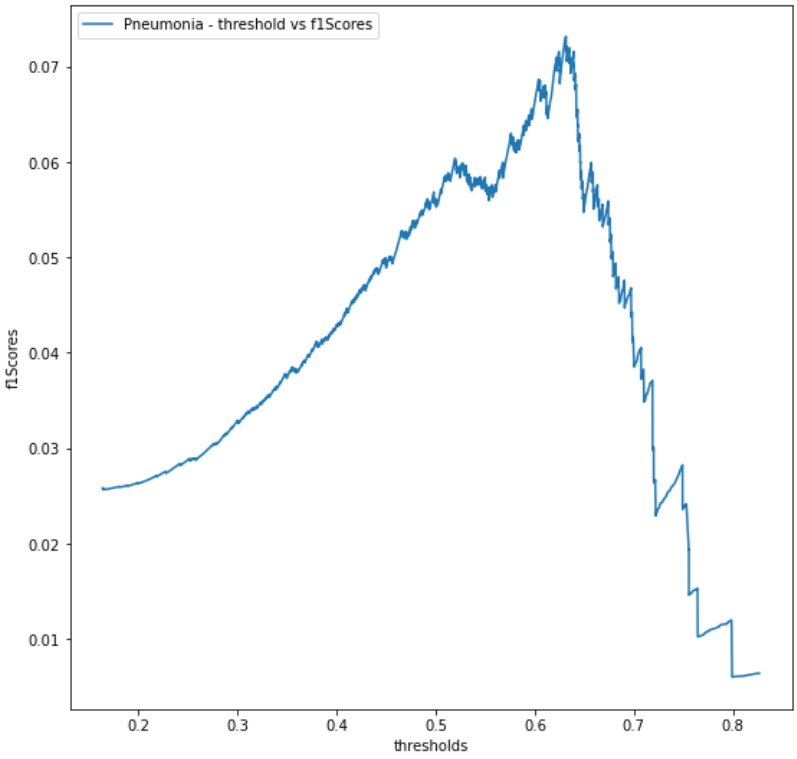
```
Epoch 10/10
72/72 [==============================] - 388s 5s/step - loss: 0.6044 - binary
_accuracy: 0.6790 - val_loss: 0.4869 - val_binary_accuracy: 0.7184

Epoch 00010: val_loss did not improve from 0.39090
```



Performance statistics and Threshold







Based on the maximum f1 score of 0.07, a threshold of 0.63 was selected to separate the negative and positive findings

Training and Validation Sets

The training dataset was created as follows:

- All data was randomly split such that 80% of the images went into the training set and the remaining 20% were put into the validation set. The random split was done in such a way that the % of pneumonia images were the same in both sets
- The negative pneumonia images in the training set were then randomly culled such that the number of positive and negative pneumonia images was the same
- The negative pneumonia images in the validation set were similarly adjusted to ensure that the % of positive cases matched the % of positive findings in the full dataset

Dataset Labeling – Benefits and limitations

The ground truth of this NIH dataset was created based on existing radiology reports accompanying the images.

The benefit of this labeling is that it is easily available.

The drawback is that the radiologist could have made an incorrect diagnosis and therefore the labeling could have been incorrect.

Ideal Dataset and Ground Truth

A more foolproof ground truth could be based on biopsy reports, but this is expensive to generate and generally may only be available in a subsample of the positive findings. BUT…, this would be ideal.

We would also want all the images culled for the presence of foreign objects. Ideally, images should also be cropped so that only the lungs are visible. All images should be screened to ensure correct orientation and scaling

Ideal Performance Metric and Metric Value

Based on a literature search, it is clear that our model is far from ideal and there is tremendous scope for improvement. I would not submit this to the FDA until our model achieved AUC and F1 scores of 99% or greater with a low standard deviation

1. https://pubmed.ncbi.nlm.nih.gov/34492046/#:~:text=Chest%20X%2Dray%20imaging%20is,using%20chest%20X%2Dray%20images.

   The proposed method achieved accuracy rates of 98.81% and 86.85% and sensitivity rates of 98.80% and 87.02% on the Kermany and RSNA datasets, respectively. Statistical analyses on the datasets using McNemar's and ANOVA tests showed the robustness of the approach. Average F1 scores were 99% and average AUC was 98% with standard deviations of 0.61 and 0.68 respectively

Comparison of state of the art methods

Dataset	Method	Acc(%)	Pre(%)	Rec(%)	F1(%)	AUC(%)
Kermany	Mahmud et al. [39]	98.10	98.00	98.50	98.30	-
	Zubair et al. [13]	96.60	97.20	98.10	97.65	-
	Stephen et al. [20]	93.73	-	-	-	-
	Sharma et al. [19]	90.68	-	-	-	-
	Liang et al. [11]	90.50	89.10	96.70	92.70	-
	Proposed Method	**98.81**	**98.82**	**98.80**	**98.79**	**98.35**
RSNA	Antin et al. [40]	-	-	-	-	61.00
	Zhou et al. [41]	79.70	-	-	80.00	-
	Yao et al. [42]	-	-	-	-	71.30
	Rajpukar et al. [14]					76.80
	Proposed Method	**86.85**	**86.89**	**87.02**	**86.95**	**86.85**

https://doi.org/10.1371/journal.pone.0256630.t007

2. https://www.sciencedirect.com/science/article/pii/S0010482520302250?via%3Dihub

Table 1. Performance comparison of the proposed method with other state-of-the-art approaches in non-COVID pneumonia detection.

Task	Methods	Accuracy (%)	AUC score (%)	Precision (%)	Recall (%)	Specificity (%)	F1 score (%)
Normal/ Pneumonia	Proposed	**98.1**	**99.4**	**98.0**	98.5	**97.9**	**98.3**
	Residual	91.2	96.4	90.7	95.9	84.1	93.4
	Inception	88.7	92.6	88.9	94.1	80.2	91.1
	VGG-19	87.2	90.7	85.6	91.1	77.9	89.3

Task	Methods	Accuracy (%)	AUC score (%)	Precision (%)	Recall (%)	Specificity (%)	F1 score (%)
	[21]	95.7	99.0	95.1	98.3	91.5	96.7
	[22]	92.8	96.8	–	93.2	90.1	–
	[23]	96.4	99.3	93.3	**99.6**	–	–
Viral/ Bacterial Pneumonia	Proposed	**95.1**	**97.6**	**94.9**	96.1	**94.3**	**95.5**
	Residual	89.5	92.4	88.3	96.9	78.1	92.4
	Inception	85.8	90.6	84.5	93.8	72.1	88.9
	VGG-19	83.2	88.5	81.1	91.3	71.7	86.6
	[21]	93.6	96.2	92.0	**98.4**	86.0	95.1
	[22]	90.7	94.0	–	88.6	90.9	–
Normal/ Viral/ Bacterial/ Pneumonia	Proposed	**91.7**	**94.1**	**92.9**	**92.1**	93.6	**92.6**
	Residual	86.3	88.5	86.3	88.5	93.5	87.4
	Inception	81.1	84.6	75.4	84.9	86.2	78.9
	VGG-19	79.8	83.1	74.5	82.9	83.4	77.9
	[21]	**91.7**	93.8	91.7	90.5	**95.8**	91.1

3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7345724/

The final proposed weighted classifier model is able to achieve a test accuracy of 98.43% and an AUC score of 99.76 on the unseen data from the Guangzhou Women and Children's Medical Center pneumonia dataset

Architecture	Accuracy	Precision	Recall	F1 Score	AUC Score
ResNet18	97.29	97.03	98.25	97.63	99.46
DenseNet121	98.00	97.53	99.00	98.26	99.65
InceptionV3	97.00	97.02	97.75	97.39	99.49
Xception	96.57	95.85	98.25	97.03	99.59
MobileNetV2	96.71	96.08	98.25	97.15	99.52
Weighted Classifier	98.43	98.26	99.00	98.63	99.76

4. https://stanfordmlgroup.github.io/projects/chexnet/

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)