# FDA Submission

**Your Name:** Khalil Rejiba

**Name of your Device:** Pneumonia detector in chest radiographs

# Algorithm Description

## 1. General Information

**Intended Use Statement:**

Assisting radiologists in detecting Pneumonia in Chest X-Rays

**Indications for Use:**

Applicable for patients ranging from 0 to 100 years of age. Applicable for both males and females.

The radiograph used should show the chest of the patient placed in the posterioanterior configuration (PA) or the anteroposterior configuration (AP).

If a patient is suspected to have Pneumonia, a chest X-ray is ordered by the doctor. The algorithm can be run before the radiologist sees the image. The result is shown to the radiologist along with the X-ray. They can further analyze it and confirm whether the prediction is right or wrong.

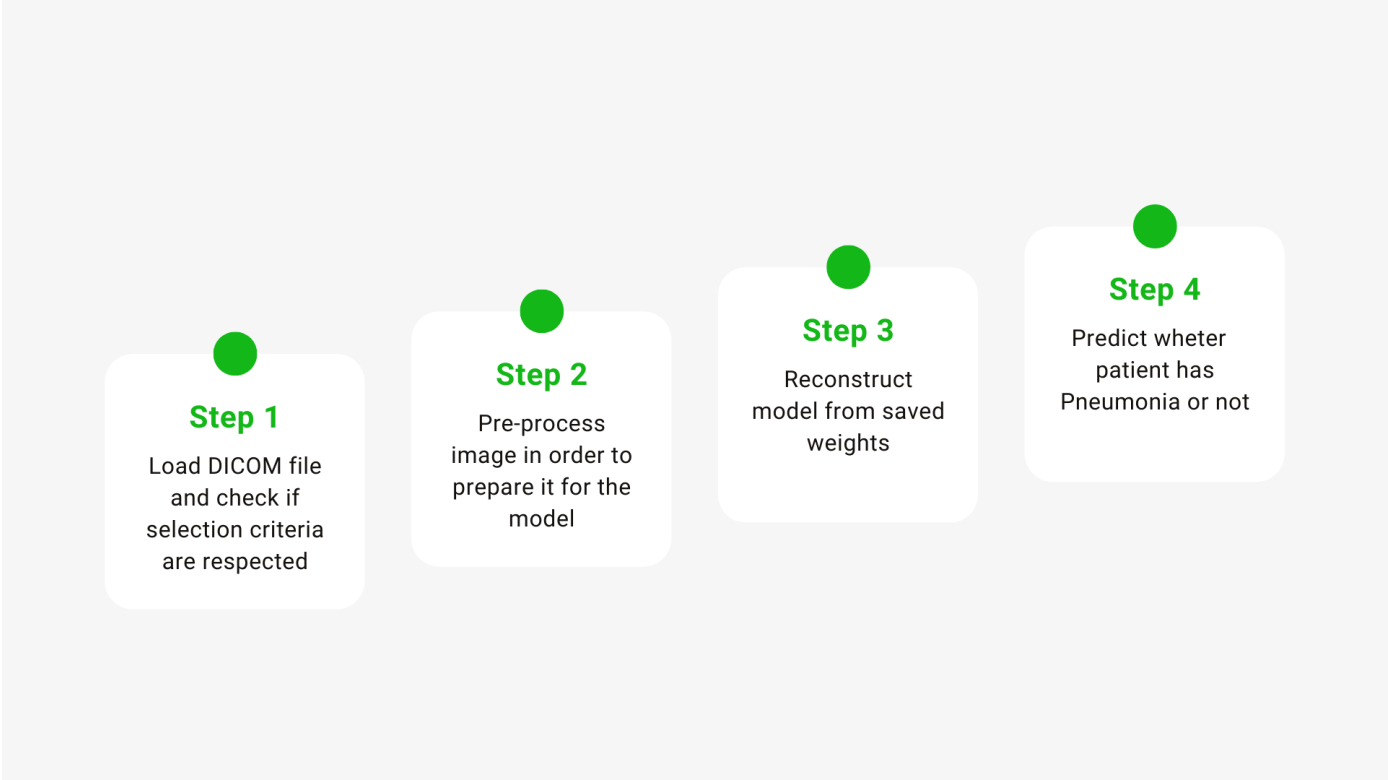
The images are extracted from DICOM files and then preprocessed.

**Device Limitations:**

- In the dataset used to train the model, Pneumonia is often observed with Infiltration and Edema. More caution is required when patients are also suspected to have **Infiltration** and/or **Edema**.
- You cannot use **Lateral views** in this algorithm.
- The **time** needed to make the prediction might vary depending on the device used. Inference of one image takes almost two seconds when using a CPU.

**Clinical Impact of Performance:**
The algortithm can be used to help push serious cases of Pneunomia to the top of the queue of the radiographs to be seen by the radiologist. This can help save patients as Pneunomonia can be life-threatening.
The algorithm results in a high number of false positives and a low number of false negatives. Having a high number of false positives is acceptable as the prediction is always verified by a radiologist afterwards. Therefore, its means more work for the radiologist to check a false positive and rule out Pneumonia if that's the case. Having a low number of false positives is something we want as cases of Pneumonia are less likely to be missed.

## 2. Algorithm Design and Function




**DICOM Checking Steps:**

- Body part: Chest
- View: PA or AP
- Modality: X-ray

**Preprocessing Steps:**

- Normalization
- Resizing
- Slight blurring to reduce foreign objects impact on prediction
- Prepare image to be used by vgg16 model using keras.applications.resnet.preprocess_input
- Reshaping to NHWC format

**CNN Architecture:** The model used is a Convolutional Neural Network. The first part is the VGG16 (https://arxiv.org/abs/1409.1556) network which is trained on the `ImageNet` dataset. We include the first 19 layers, i.e. up to layer `block5_pool` . Then, we added the layers shown in the following table. The networks contains: 15,340,833 trainable parameters and 12,354,880 non-trainable parameters.

| Layer (type) | Output Shape | Param # |
|---|---|---|
| vgg-pretrained (Model) | (None, 7, 7, 512) | 14714688 |
| flatten_1 (Flatten) | (None, 25088) | 0 |
| dense_1 (Dense) | (None, 512) | 12845568 |
| dropout_1 (Dropout) | (None, 512) | 0 |
| dense_2 (Dense) | (None, 256) | 131328 |
| dropout_2 (Dropout) | (None, 256) | 0 |
| dense_3 (Dense) | (None, 16) | 4112 |
| dropout_3 (Dropout) | (None, 16) | 0 |
| dense_4 (Dense) | (None, 1) | 17 |

## 3. Algorithm Training

**Parameters:**
We used an Adam optimizer with early stopping.

- Types of augmentation used during training:
  We used `ImageDataGenerator` from the Keras library with the following parameters:

  > horizontal_flip = True
  > vertical_flip = False
  > height_shift_range = .1
  > width_shift_range = .1
  > rotation_range = 10
  > zoom_range =.1

- Batch size: 64
- Optimizer learning rate: 1e-5
- Layers of pre-existing architecture that were frozen:

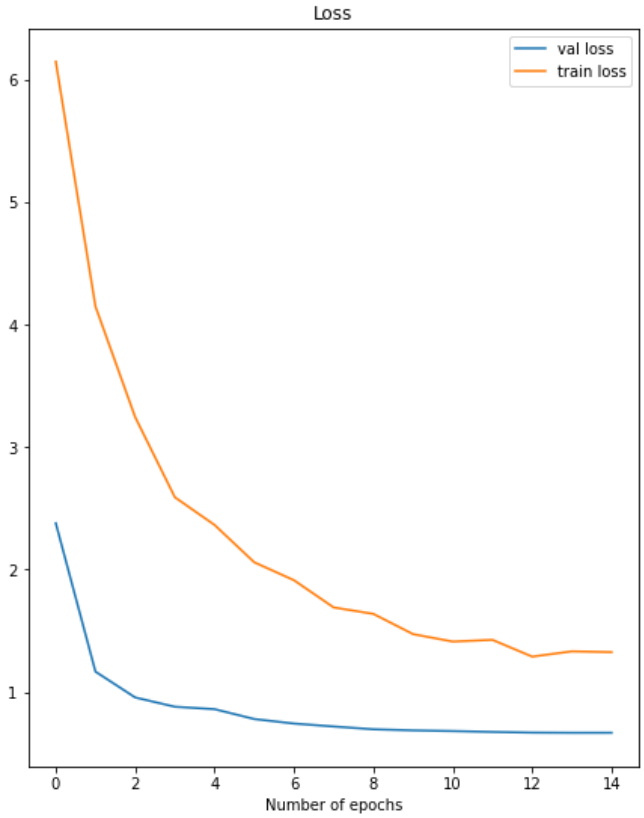
  > Layers 0-16 were frozen.

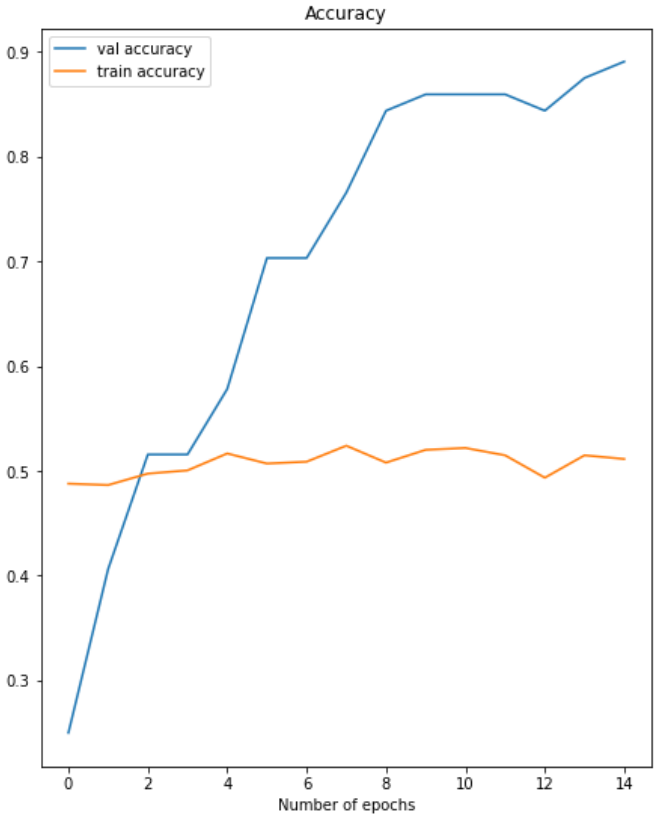
- Layers of pre-existing architecture that were fine-tuned:

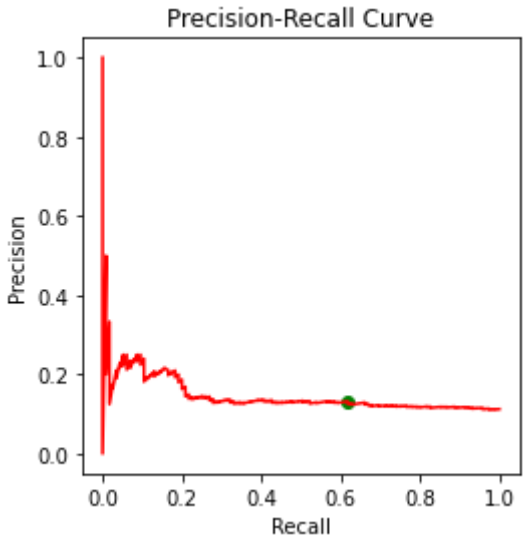
  > The last convolutional layer (17) was trained. The last pooling layer (18) doesn't contain any trainable parameters.

- Layers added to pre-existing architecture:

  > Refer to CNN Architecture for more details.






**Final Threshold and Explanation:**

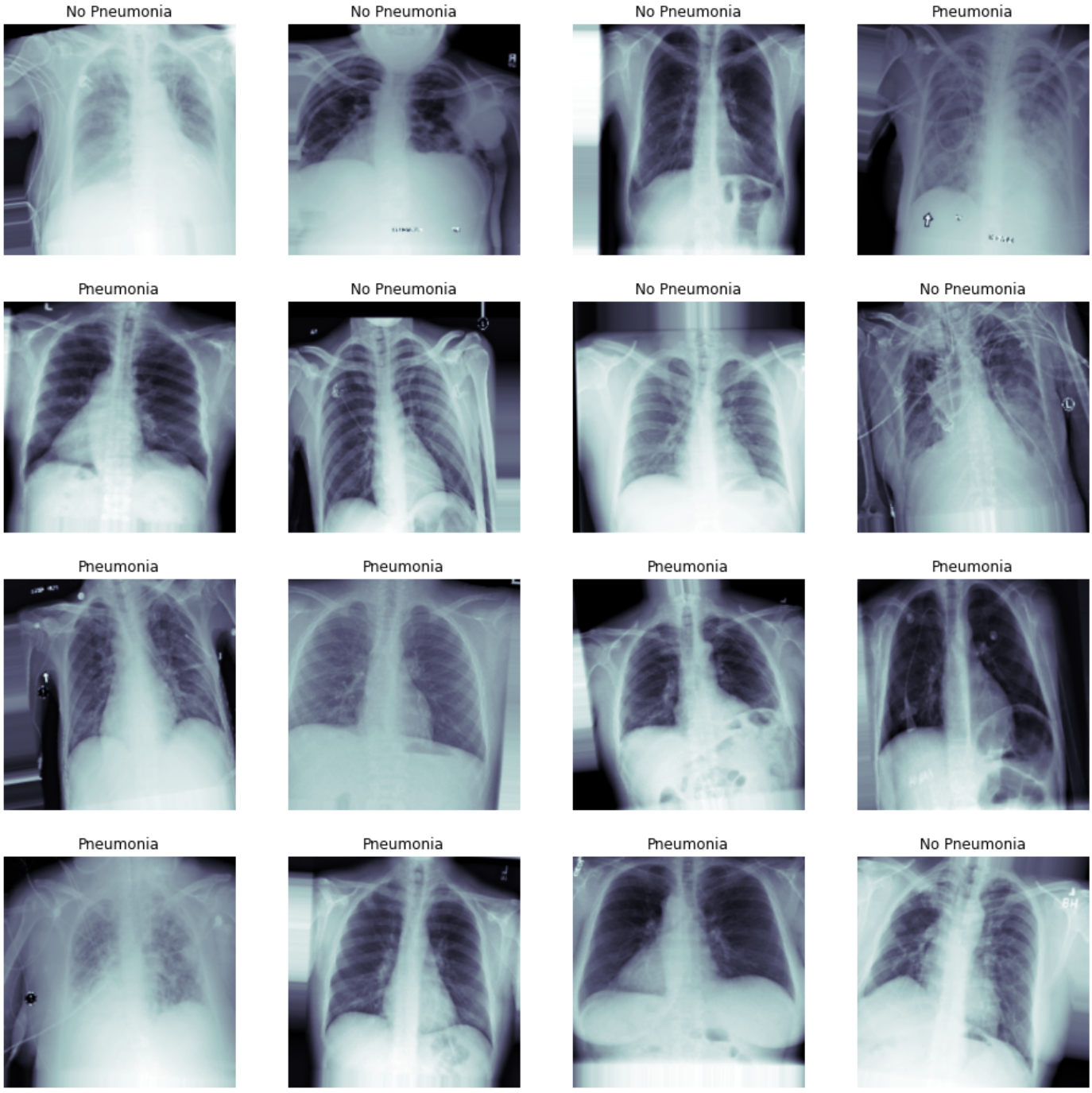
We chose to maximize the F1-score which gives equal importance to precision and recall.
In the literature, the average radiologist has an F1-score of 0.387. → Rajpurkar2017 (https://arxiv.org/abs/1711.05225)

Maximum F1-score: 0.218
Corresponding Recall: 0.617
Corresponding Precision: 0.132
Corresponding Threshold: 0.483

The F1-score of our algorithm is lower than the average score of a radiologist. False negatives are the most problematic in the clinical case we're interested in. Therefore, we accepted a lower F1-score as long as the corresponding recall was sufficiently high.

## 4. Databases

Both training and validation sets are extracted from the *NIH Chest X-ray Dataset*. The prevalence of Pneumonia in this dataset is 1.3%. Healthy subjects represent 53.8% of the people in this study. The other 44.9% corresspond to people who either have Pneumonia and another disease at the same time or simply have a disease other than Pneumonia. The other diseases present are: Mass, Atelectasis, Pleural Thickening, Edema, Effusion, Pneumothorax, Herni, Infiltration, Consolidation, Emphysema, Cardiomegaly, Nodule and Fibrosis. Around 23% of patients with Pneumonia don't have comorbidities and around 14% of patients with Pneumonia also have Infiltration but no other disease. Around 10% of patients with Pneumonia have Infiltration and Edema but no other disease.
**Description of Training Dataset:**
The training dataset contains 2290 samples of which 50% are Pneumonia cases. The age distribution is similar to that in the original dataset, e.g. 95% of the subjects are between the ages of 18 and 80.
We used data augmentation to increase the number of images our model uses to learn.
The following figure shows some samples from the training set after augmentation.




**Description of Validation Dataset:**
The validation dataset contains 2860 samples of which 10% are Pneumonia cases. The prevalence of Pneumonia (10%) was estimated for an ER setting and it is an overestimation of the prevalence of Pneumonia in the original dataset (1.3%) which doesn't necessarily reflect an ER setting. The age distribution is similar to that in the original dataset, e.g. 95% of the subjects are between the ages of 18 and 80.

## 5. Ground Truth

The target labels are obtained from radiology reports through Natural Language Processing(NLP). Labelling accuracy is estimated to be >90%. This is expected because NLP models can sometimes provide wrong labels when the phrase is ambiguous. For chest X-ray and lung diseases, the gold standard is a biopsy, which is difficult to obtain in some cases along with the image data. That's why a labelling by several radiologists is sufficient.

## 6. FDA Validation Plan

**Patient Population Description for FDA Validation Dataset:**
FDA Validation Dataset should contain all age groups and genders. The age distribution should be so that 95% of the subjects are between the ages of 18 and 80. The gender distribution should be 56.5% males and 43.5% females. The views should only be PA or AP. Having people with Infiltration and/or Edema can give false results. This validation dataset should contian 10% of Pneumonia cases.

**Ground Truth Acquisition Methodology:**
Ideally, biopsy results would make the best labels. If not available, a NLP model extracts the labels from radiology reports. A voting system which takes into account the experience of the radiologist could be used to make the final decision.

**Algorithm Performance Standard:**
The F1-score should be more than 0.2 with recall larger than 0.5. In Rajpurkar2017 (https://arxiv.org/abs/1711.05225), the authors proposed a 121-layer CNN named CheXNet was proposed to detect Pneumonia. It reached an F1-Score of 0.435, which is higher than our algorithm. We accepted a lower F1-score as long as the corresponding recall was sufficiently high as explained in Algorithm Training.