FDA Submission

Your Name: María del Rocío Bernal Zamorano

Name of your Device: PneuNet

Algorithm Description

1. General Information

Intended Use Statement: This algorithm is intended for use as a pneumonia early detection tool for screening studies of a broad population of men and women up to 100 years old. The algorithm classifies chest x-ray images according to the suspicion or not of the presence of pneumonia, so the radiologist can focus on the positive cases first and do a more detailed check and further studies on these cases.

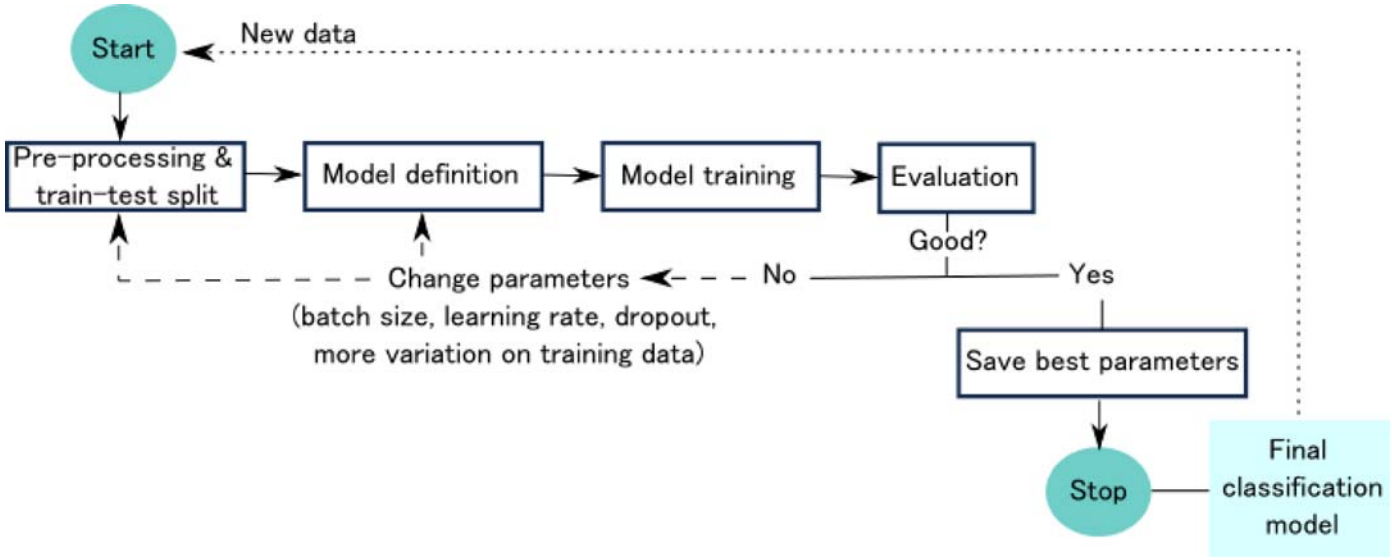
Indications for Use: This algorithm can be used on both women and men within an age range from 0 to 100 for whom a chest x-ray has been taken due to pneumonia suspicion. It can be use for both patient positions during image acquision, AP and PA. It can be used in the clinic to avoid specialist's burnout so they can analyze the positive cases first.

Device Limitations: Pneumonia may occur in a combination of other diseases. Its most-common comorbidities are infiltration, edema, effusion, and atelectasis. In the case of presence of infiltration and/or edema diseases in a chest x-ray, there is a limitation of this algorithm. The algorithm may perform very poorly on the accurate detection of pneumonia in the presence of infiltration and/or edema because it is not able to differentiate between them. This would affect the algorithm accuracy and performance so it is recommended to avoid images with infiltration and/or edema.

Clinical Impact of Performance: A false positive in this device would mean that the algorithm has detected pneumonia while the patient is healthy. A false negative would mean that a patient with pneumonia has been classified as healthy, which could be lethal for the patient (although a radiologist will look at all the images so it could be avoided). It is therefore preferred in this case to have a low number of false negatives. That means a high recall (low false negatives) is preferred over a high precision (low false positives). A high sensitive (high recall) test is preferred for screening studies since it is very reliable when the result is negative (low false negatives). Therefore the algorithm is rarely misdiagnosing people who have the disease. Although we are not taking false positives into account so we are still labeling a lot of negative cases as positive, this is better than missing cases with pneumonia.

2. Algorithm Design and Function

Algorithm Flowchart



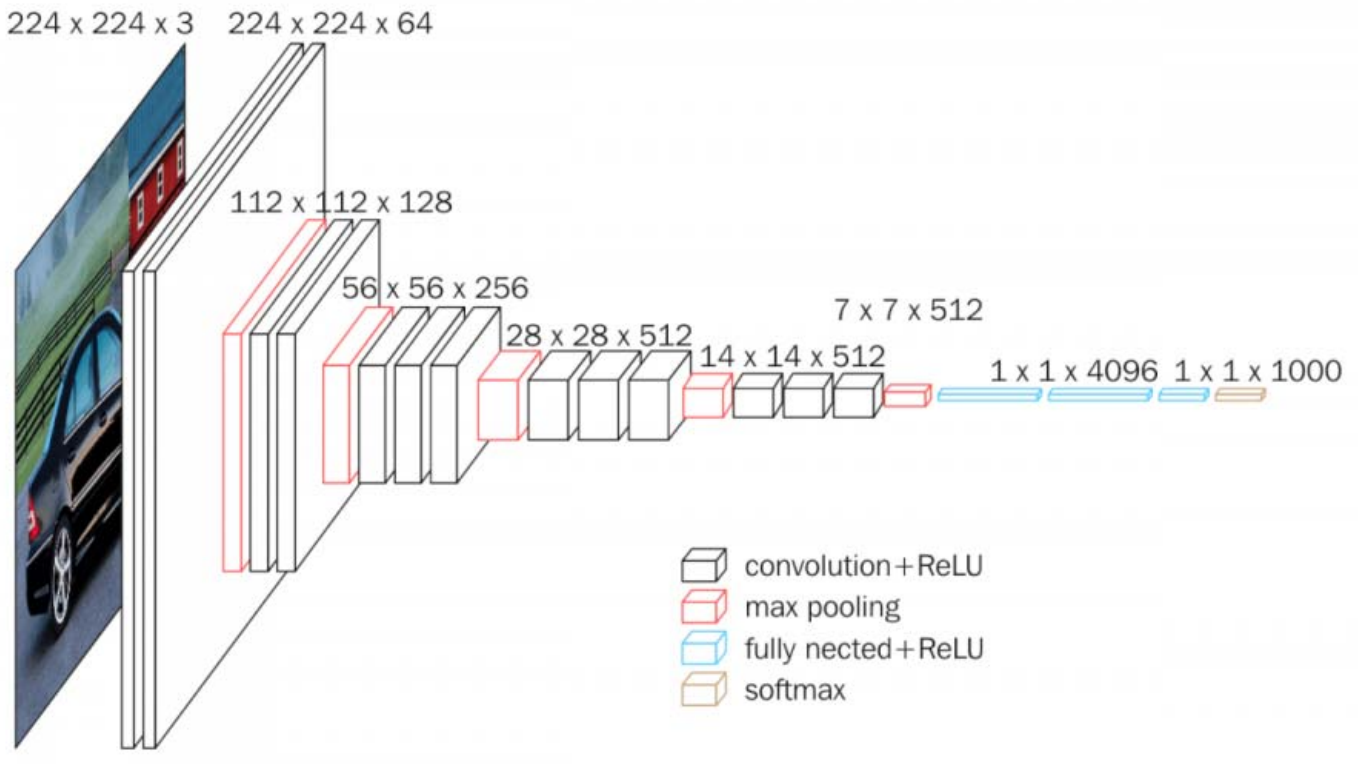
DICOM Checking Steps: It will be check that in the DICOM appear the image modality as DX, the patient position as AP or PA, and the body part examied as CHEST. If that is not the case, a message will be displayed.

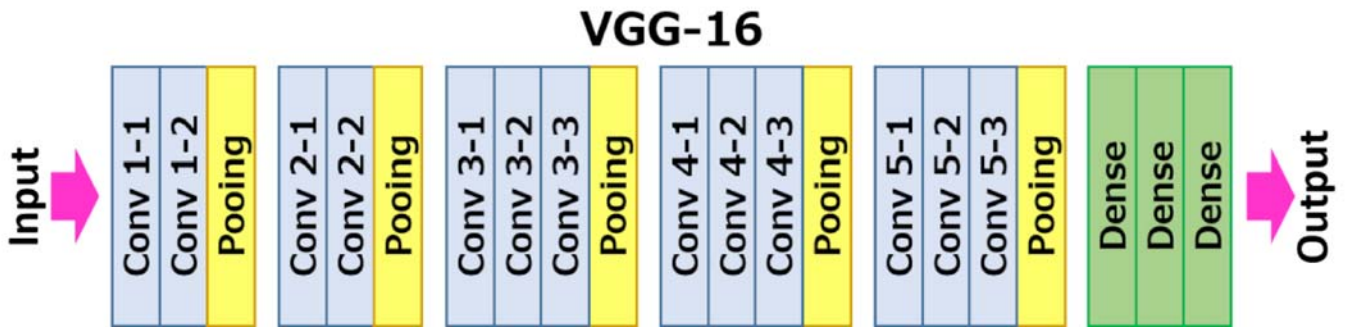
Preprocessing Steps: Image normalization, specifically standarization, is done to all images prior to using them for training. That is substracting the mean from each pixel and dividing by the image's standard deviation. Making all the intensity values fall within a small range that is close to zero helps the weights on the convolutional filters stay under control. Later, image resize is required, since CNNs have an input layer that specifies the size of the image they can process - in this case (for a VGG16 model) the image size is (1,224,224,3). After that, we clean the data by selecting only people whose age is lower than 100. Finally, we perform image augmentation of the training data (not for validation) to enlarge the dataset by creating different versions of the original data.

CNN Architecture: CNN architecture is inspired by the organization and functionality of the visual cortex and designed to mimic the connectivity pattern of neurons within the human brain. The neurons within a CNN are split into a three-dimensional structure, with each set of neurons analyzing a small region or feature of the image. In other words, each group of neurons specializes in identifying one part of the image. CNNs use the predictions from the layers to produce a final output that presents a vector of probability scores to represent the likelihood that a specific feature belongs to a certain class. A CNN is composed of several kinds of layers:

- Convolutional layer: creates a feature map to predict the class probabilities for each feature by applying a filter that scans the whole image, few pixels at a time.
- Pooling layer (downsampling): scales down the amount of information the convolutional layer generated for each feature and maintains the most essential information (the process of the convolutional and pooling layers usually repeats several times).
- Fully connected input layer: “flattens” the outputs generated by previous layers to turn them into a single vector that can be used as an input for the next layer.
- Fully connected layer: applies weights over the input generated by the feature analysis to predict an accurate label.
- Fully connected output layer: generates the final probabilities to determine a class for the image.

In particular, we use a VGG16 convolutional neural network model, proposed by K. Simonyan and A. Zisserman. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. The architecture of this model is shown below:



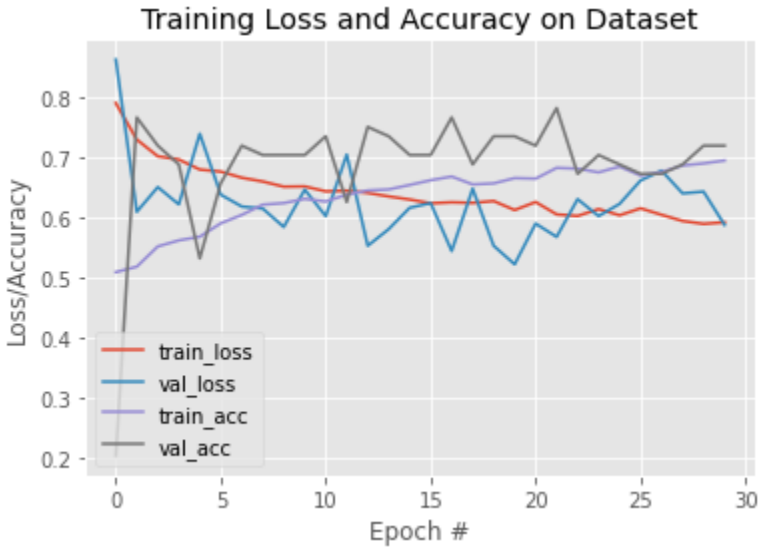


3. Algorithm Training

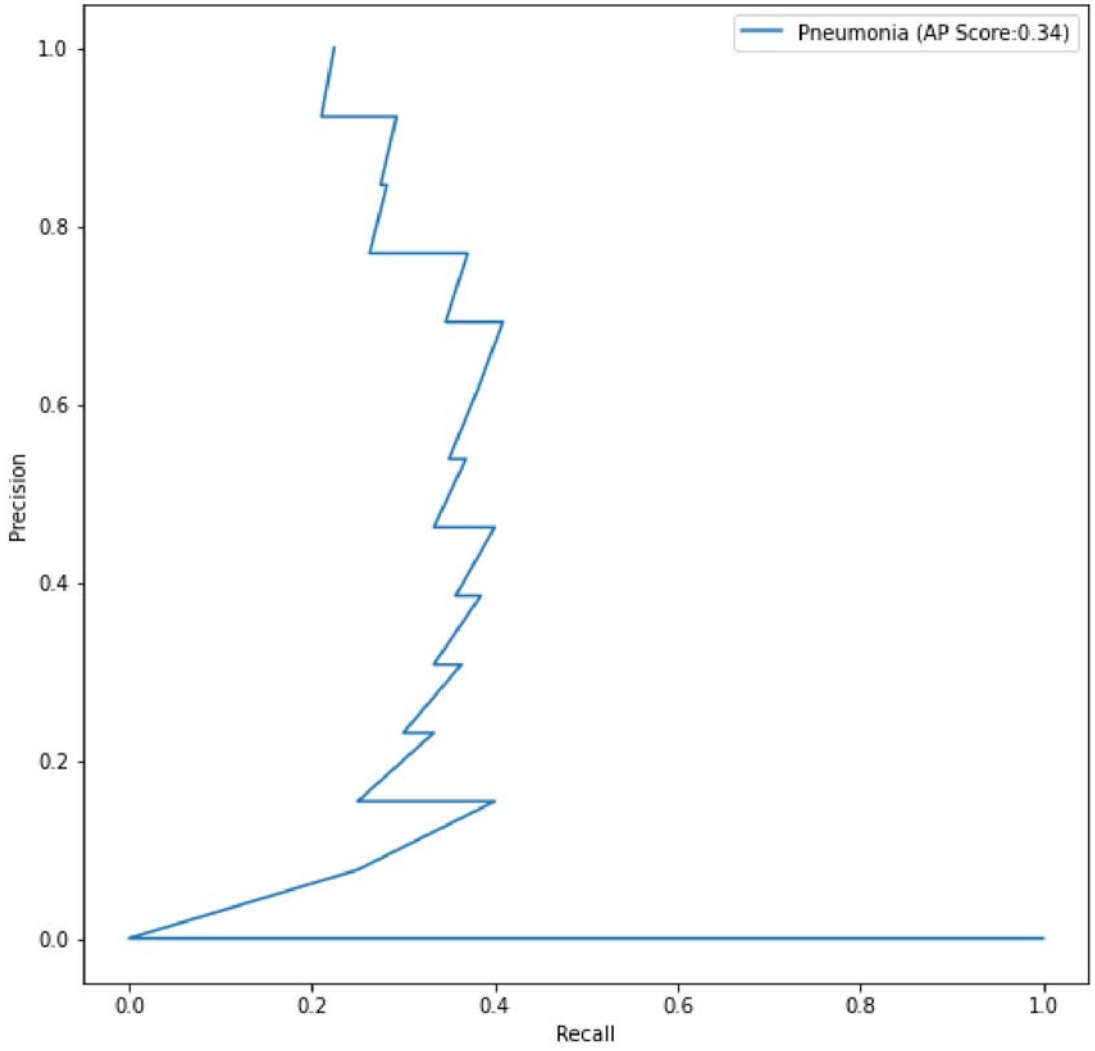
Parameters:

- **Types of augmentation used during training:** rescale=1. / 255.0, horizontal_flip = True, vertical_flip = False, height_shift_range= 0.3, width_shift_range= 0.3, rotation_range= 30, shear_range = 0.3, zoom_range=0.3
- **Batch size:** 64
- **Optimizer learning rate:** Adam(lr=1e-4)
- **Layers of pre-existing architecture that were frozen:** all convolutional layers except the last one - [0:17] layers.
- **Layers of pre-existing architecture that were fine-tuned:** block5_conv3 and block5_pool - 2 layers
- **Layers added to pre-existing architecture:** Flatten(), Dropout(0.5), Dense(1024, activation='relu'), Dropout(0.5), Dense(512, activation='relu'), Dropout(0.5), Dense(256, activation='relu'), Dense(1, activation='sigmoid') - 8 layers.

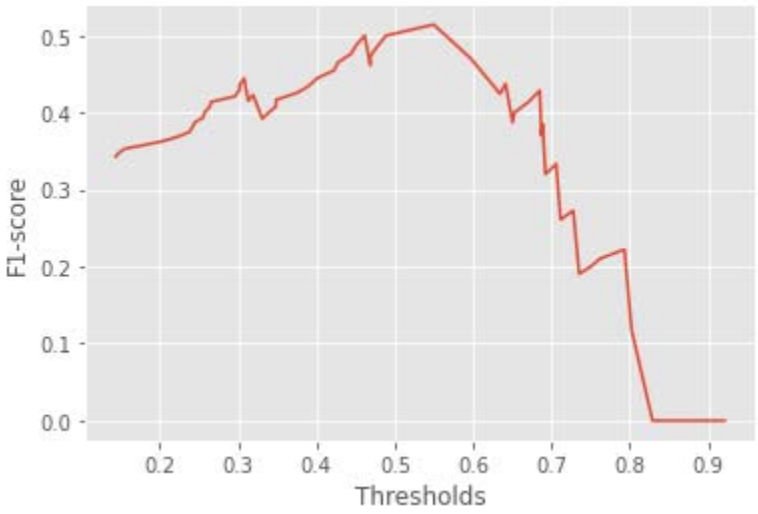
Algorithm training performance

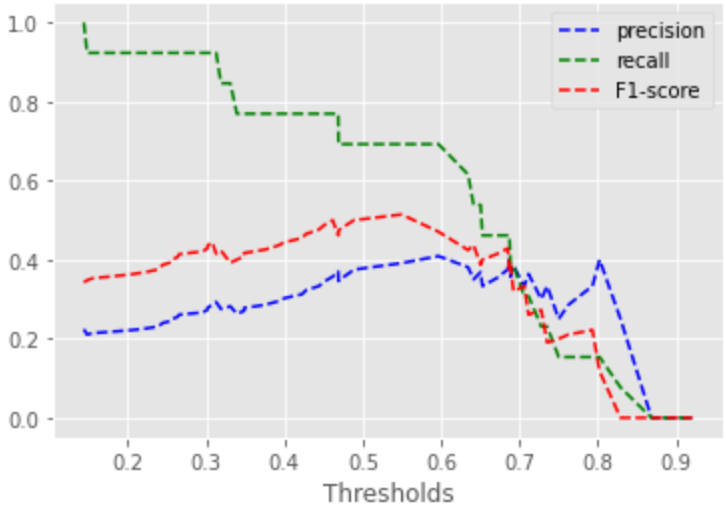


Precision - Recall curve



Final Threshold and Explanation: The final threshold is the one that maximizes the F1-score (0.51), that is a threshold of 0.55. The corresponded precision and recall are 0.39 and 0.69 respectively. The reason to chose the F1-score to evaluate the performance of the model is because, for binary classification problems, it combines both precision and recall. It is maximized when precision and recall are balanced, and it allows us to better measure a test's accuracy when there are class imbalances.



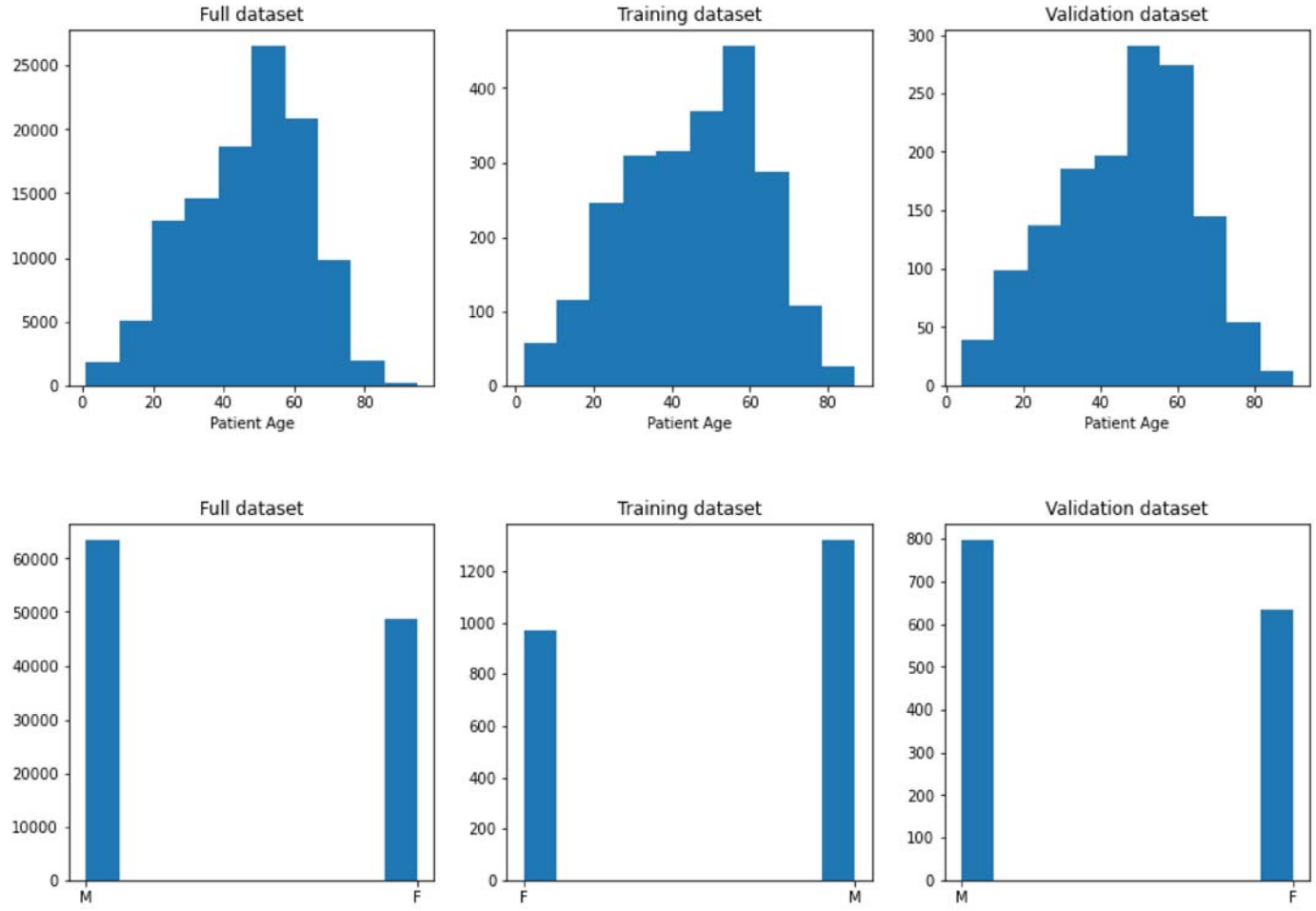


4. Databases

Description of Training Dataset: The training dataset contains 80% of the total positive cases and it is balanced (50% positive and 50% negative). This dataset is augmented to enlarge the dataset and increase the variation of the training data.

Description of Validation Dataset: The validation dataset contains 20% of the total positive cases. For the validation dataset, we need to have an inbalanced dataset to reflect the original distribution, however the original distribution is too skewed towards the negative class (pneumonia has a relative frequency of 1.3% as shown in the EDA). So in order to keep the validation set unbalanced and still not too skewed, we will use a a 20-80% proportion (20% positive and 80% negative) for pneumonia to reflect the imbalance.

The graphs below show that the age and gender distribution in both training and validation datasets are similar than those of the original dataset.



5. Ground Truth

The ground truth of a dataset represents the set of labels that determine which class each data element belongs to. The method by which they're created can have a significant impact on the overall accuracy of the algorithm and also its reliability in a clinical setting. The dataset provided for this project was curated by the NIH specifically to address the problem of a lack of large x-ray datasets with ground truth labels to be used in the creation of disease detection algorithms. The disease labels were created using Natural Language Processing (NLP) to mine the associated radiological reports (Wang et al.). Radiologists tend to write abstract and complex logical reasoning sentences, therefore a variety of NLP techniques were adopted for detecting the pathology keywords and removal of negation and uncertainty. Each radiological report was either linked with one or more keywords or marked with 'Normal' as the background category.

The advantages of using this method to create the dataset are the saving of time and money. Since it is not necessary to go through each radiologist report manually, it is a fast method. It is also less costly than hiring a group of radiologists to label the images. The disadvantages, on the other hand, are that the access to a NLP tool is required, or otherwise it is required to build one. Also, that these labels may not be 100 percent accurate. However, in this case, although there could be some erroneous labels, the NLP labeling accuracy is estimated to be >90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset: Women and men from 0 to 100 years old with no prior history of infiltration and/or edema. The data should be DX images of chest in PA or AP position.

Ground Truth Acquisition Methodology: Radiologist labels are expected to be used to extract diagnosis from. This could be obtained by a NLP tool or by a silver standard that involves hiring several radiologists to each make their own diagnosis of an image and determining the final diagnosis by a voting system. A NLP tool is preferred to save money and time.

Algorithm Performance Standard: As shown in some papers, the F1 score can be used as evidence of good performance on a device. The common values of F1 scores for some of the models shown in these papers are the following: 0.387 - radiologist average (Rajpurkar et al.), 0.435 - CheXNet (Rajpurkar et al.), 0.71 - MetaMap (Wang et al.), 0.77 - ChestX-ray8 (Wang et al.), 0.94 (Kaushik et al.). The average value is 0.65, which is higher than our F1 score (0.51). However, we can see that our value is higher than the one obtained by the radiologist average and also higher than the one obtained for the CheXNet model.

Bibliography

- Wang et al.: "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases", 2017, arXiv:1705.02315.
- Rajpurkar et al.: "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning", 2017, arXiv:1711.05225.
- Kaushik et al.: "Pneumonia Detection Using Convolutional Neural Networks (CNNs)", 2020, DOI: 10.1007/978-981-15-3369-3_36.