FDA Submission

Name: Callum Canavan

Name of device: X-ray Image Pneumonia Classifier

Algorithm Description

1. General Information

Intended Use Statement:

For assisting the radiologist(s) in the detection of pneumonia in chest x-ray.

Indications for Use:

Acting as one radiologist in a set of two or more radiologists, classifying chest x-ray images of patients aged 20-70, in both posterior-anterior (PA) and anterior-posterior (AP) positions with DX modality.

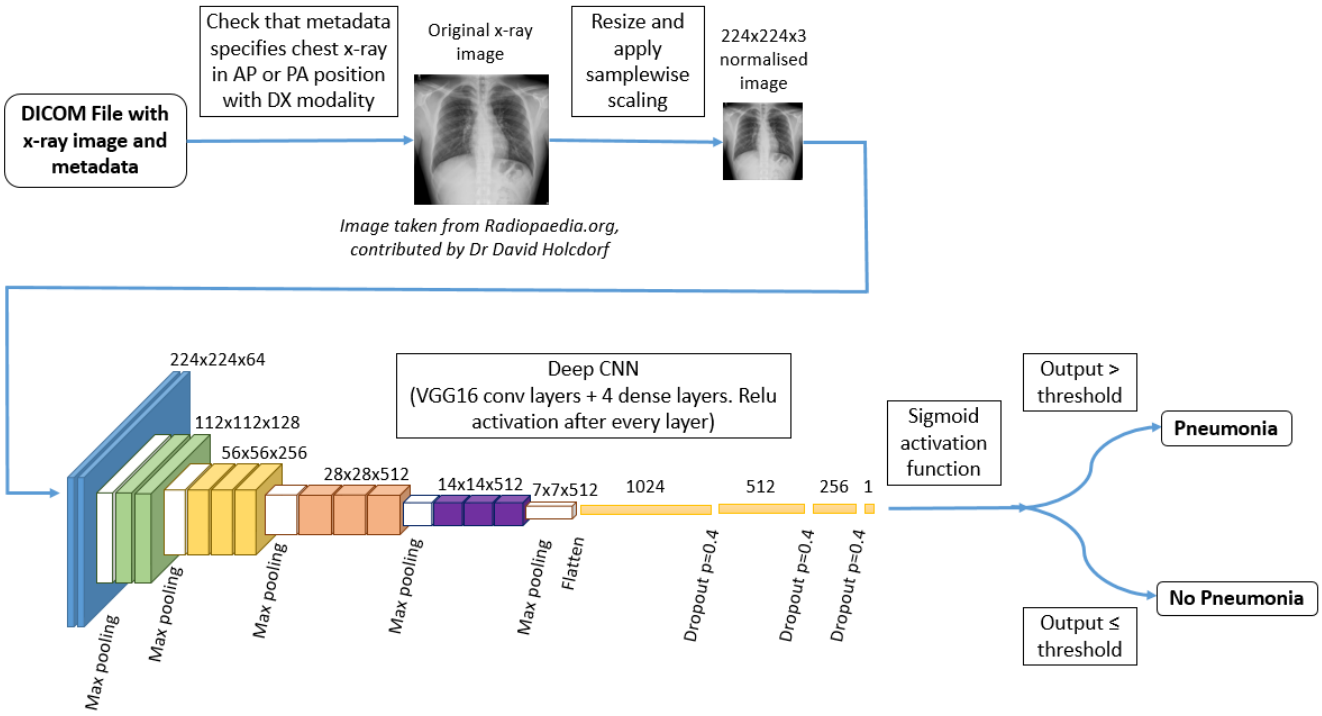
Device Limitations:

Performs more poorly than average in the presence of atelectasis (F1 of 0.366 and especially low recall of 0.341 for this population among test samples, meaning many cases of pneumonia went undiagnosed).

Clinical Impact of Performance:

In the case of a false positive, the patient may be put under unnecessary stress if treatment begins before further testing can confirm the presence of pneumonia (e.g. through sputum testing). In the case of a false negative the pneumonia may go untreated, putting the patient in serious danger. However, both of these outcomes rely on the other radiologists giving the same false classification as the model.

2. Algorithm Design and Function



DICOM Checking Steps:

The file is checked to have either AP or PA for Patient Position, Chest for Body Part Examined and DX for Modality, as these conditions were true for all images in the training data. If all three of these conditions are met, it may be passed through the model.

Preprocessing Steps:

Samples are resized into 224x224 images with 3 rgb colour channels for compatibility with the convolutional network described in the flowchart and undergo samplewise scaling (every pixel value in a sample image is shifted downward by the mean pixel value of that sample and then divided by the standard deviation of pixel values in that sample) for consistency with the training data.

CNN Architecture:

The classifying model consists of a deep convolutional neural network (VGG16) comprised of 13 convolution layers with relu activation functions and 5 max pooling layers, followed by 3 dense hidden layers with relu activation functions and a final dense output layer with one sigmoid-activated neuron (see above diagram for more details). If the output of the final layer is greater than the calibrated threshold (0.562), the image is classified as a positive case of pneumonia, and if the output is less than or equal to the threshold the image is classified as a negative case. This architecture is used because it can identify complex patterns of varying sizes within an image, some of which may indicate the presence or absence of pneumonia. The threshold was chosen to maximize F1 score (tradeoff between precision and recall) in the test dataset.

3. Algorithm Training

Parameters:

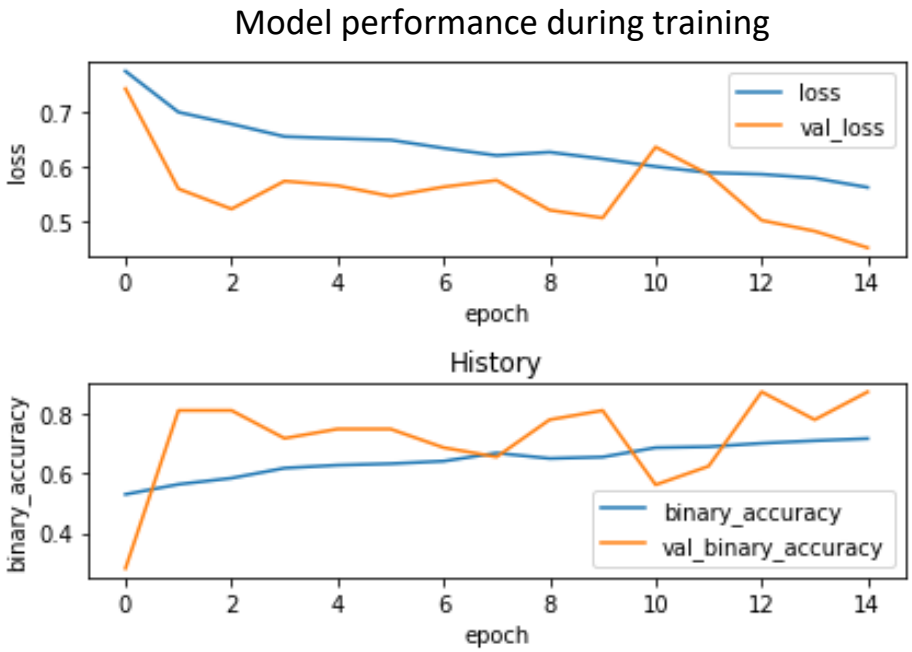
Sample images used for model training were resized into 224x224 images with 3 rgb colour channels (for compatibility with the VGG16 convolutional network structure) and samplewise scaled (to prevent vanishing/exploding gradients). For data augmentation during training they were also randomly horizontally flipped (probability 0.5), height-shifted, width-shifted, rotated, sheared and zoomed. The scales for height-shifting, width shifting, shearing and zooming were each chosen from a uniform distribution in the range [0, 0.1] and the angle of rotation in degrees was chosen from a uniform distribution in the range [-10, 10] for each image in each epoch.

The images were forward-propagated through the model in batches of 64, chosen for a tradeoff between training speed and computational efficiency. The batch size for validating the model after each epoch and testing the model after training was chosen to be 32 since training speed was not a factor here.

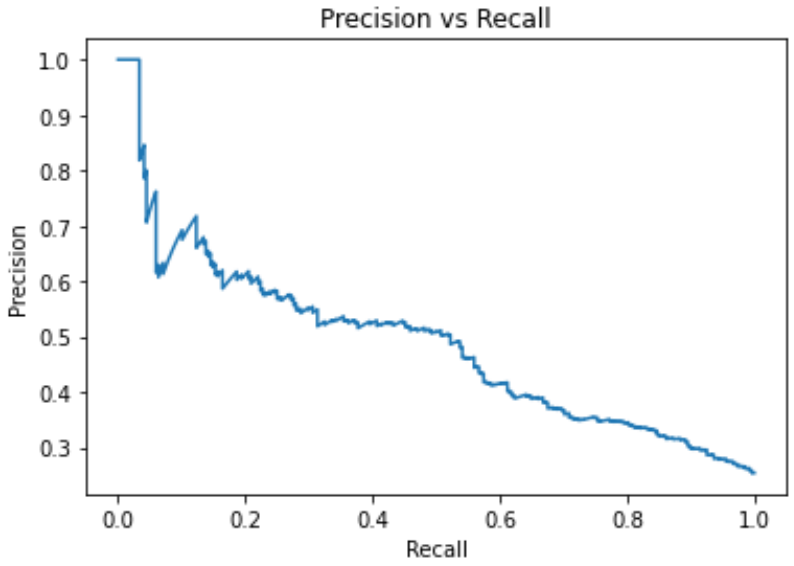
Adam (with default $\beta_1 = 0.9$ and $\beta_2 = 0.999$) was chosen as the optimizing algorithm during model training due to its combination of momentum and RMSProp properties making it a historically good default for complex image classification problems. A learning rate of 1e-4 was used after trying several rates in the [1e-3, 1e-5] log range and finding this learning rate to give fastest convergence (within around 12 epochs) and lowest loss on the validation set.

The pretrained convolutional layers of this model were taken from VGG16, a model originally trained on ImageNet to classify images across 1000 categories. All convolutional layers except the last one had their weights frozen during training, meaning the last convolutional layer was the only one which was fine-tuned. The dense layers at the end of VGG16 were removed and replaced with 4 new dense layers of sizes 1024, 512, 256 and 1, with the last layer representing the output. The first 3 dense layers are relu activated while

the last is sigmoid activated. These were subsequently trained on the training set with random dropout (rate 0.4) after each hidden layer for regularization.



The training loss (binary cross entropy of classifications, also known as log loss) decreased steadily over all training epochs. The validation loss was also found to decrease but did so more erratically, and seemed to converge by the 15th epoch (training the model for more epochs after this led to overfitting).



Final Threshold and Explanation:

The above figure shows the tradeoff between precision and recall on the test set. The threshold (0.562) between positive and negative cases given the final output of the model was chosen to optimise algorithm's F1 score (0.513), the harmonic mean of precision and recall. This gave a precision of 0.460 and recall of 0.560 on the test set.

4. Databases

Description of Training Dataset:

The training dataset consisted of 2310 images with a 1:1 ratio of positive to negative pneumonia cases. The gender distribution was slightly skewed towards males and patient ages were mostly within the range 20-70. View positions were approximately equally distributed between PA and AP positions, with a slight skew towards PA. The two most frequently co-occurring diseases with pneumonia were infiltration and edema, occurring in 41.7% and 24.1% of positive pneumonia cases, respectively.

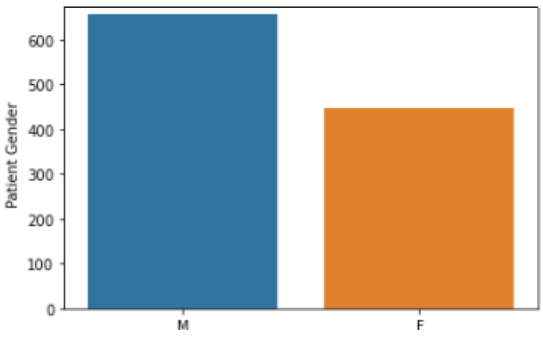
Description of Validation Dataset:

The training dataset consisted of 1104 images with a 1:3 ratio of positive to negative pneumonia cases. The gender distribution was skewed towards males (more so than in the training data) and patient ages were mostly within the range 20-70. View positions were slightly skewed towards the PA position over AP. The two most frequently co-occurring diseases with pneumonia were infiltration and edema, occurring in 44.6% and 22.5% of positive pneumonia cases, respectively.

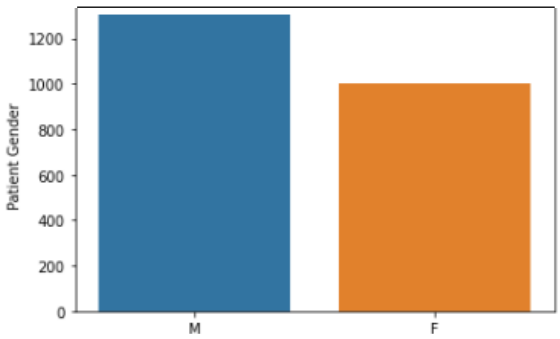
No image or patient ID was common to both the training and validation datasets, to prevent data leakage and give a more accurate reflection of the algorithm's performance in a real scenario (it is unlikely that any given patient in a real hospital setting will have had a previous x-ray of theirs used in our training dataset).

Demographic distributions are shown below for each dataset.

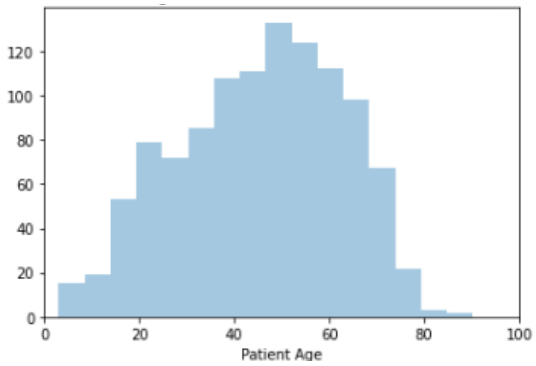
Gender distribution (validation set)



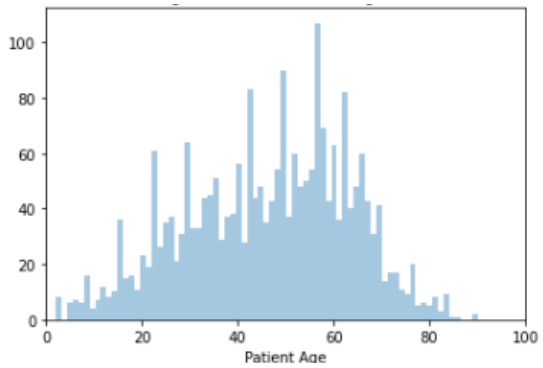
Gender distribution (training set)



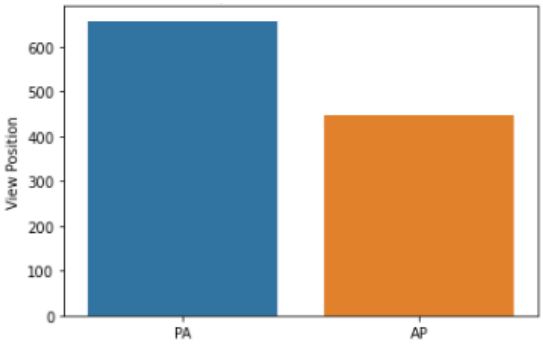
Age distribution (validation set)



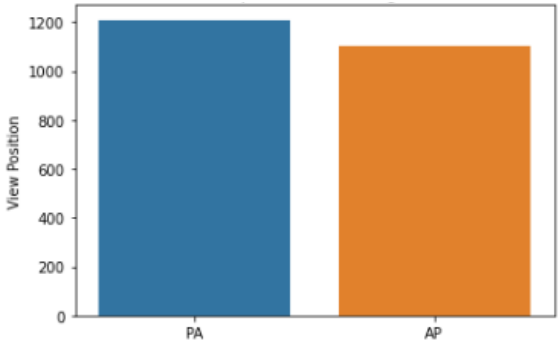
Age distribution (training set)



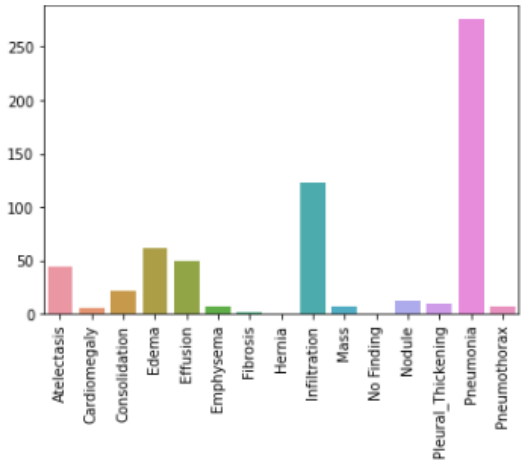
View position (validation set)



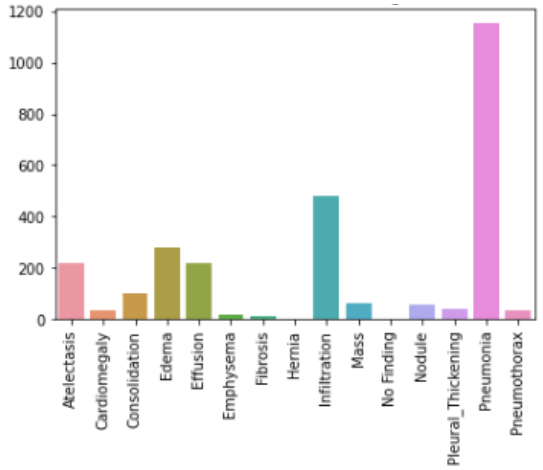
View position (training set)



Co-occurrence (validation set)



Co-occurrence (training set)



5. Ground Truth

Data used for training and validating this model were obtained from the NIH Chest X-ray Dataset. 14 disease labels (including pneumonia) specifying the presence of each disease were created using natural language processing to mine diagnoses from the radiological report associated with each image. The benefit of this method is that labels of 112,000 images were obtained quite easily and with a consistent methodology. The limitation of this method is that some labels may be incorrect with no way to check (without access to the original radiological reports), although label accuracy is estimated to be over 90%. More details on this process can be found in [1].

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

Patient population should comprise of males and females in approximately equal proportion distributed between the ages of 20 and 70 who have tested negatively for atelectasis. Data should comprise of x-rays taken of the chest in the AP and PA positions in approximately equal proportion with DX modality. Positive pneumonia cases should make up around 20-30% of the dataset.

Ground Truth Acquisition Methodology:

Ground truth labels should be obtained with silver standard classification: having 3 radiologists classify each x-ray image for the presence of pneumonia and aggregate them using a majority voting method.

Algorithm Performance Standard:

The algorithm's F1 score should be at least as high as that of the average radiologist. A suitable value for this would be 0.387 as found in this study [2].