Identifying Cardiomegaly in Chest X-Ray Images

Neeharika Sinha

June 1st 2020

Table of Contents

1. Introduction

The challenge in this project is to build an algorithm to identify through the chest X-ray images, whether a patient is suffering from Cardiomegaly. Cardiomegaly is one of the 14 Common Thorax Disease Categories, namely; Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural_Thickening or Hernia. We need an extremely accurate algorithm as this will lead to peoples lives.

The reason for selecting Cardiomegaly is because it is one of a non-physician seemed tractable anomaly, which is to look for a big heart in a patient. It will be a good quick and reliable confirmation of the model of a layman.

Images obtained by MRI machines, CT scanners, and X-rays, are some of essential medical imaging techniques to allow clinicians to identify any abnormalities in the human body. Chest X-rays are the most common type, because of its lower dose of radiation, lower cost and it needs only less than a minute to take an image. These images often contain large amounts of complex informations that can be strenuous and time consuming for doctors or clinical practitioner to asses.

Machine learning and deep learning algorithms offer the opportunity to streamline pathologists' decision-making, allowing them to review detailed data with improved accuracy and fewer errors.The FDA recently cleared an AI algorithm that can detect distal radius fractures and provide clinical decision support at the point of care.

In this project the Machine learning and the deep learning algorithms are build specifically to identify the "Cardiomegaly" infection. Although the model is decent enough to be used for predicting Cardiomegaly in Chest Xray data, however, it is not a good approximation of a real clinical environment where it remains a rare condition. The model is thus better suited for informative or additional information and not at all well suited for a screening-style use.

2. Data Acquisition and Wrangling

The data acquired here for the analysis is extracted from the clinical PACS database at National Institutes of Health Clinical Center and consists of ~60% of all frontal chest x-rays in the hospital. The dataset comprises of 112,120 frontal-view X-ray images of 30,805 unique patients with the text-mined fourteen disease image labels. There are some multi-labels CXR images too, mined from the associated radiological reports using natural language processing as seen in this this IPython notebook.

The data set consist of a csv file "Data_Entry_2017.csv" with features identified as "Image Index', 'Finding Labels', 'Follow-up #','Patient ID', 'Patient Age','Patient Gender','View Position'. Another set of data consists of Chest Xray images corresponding to "Image Index". The csv file was analyzed to pull out some useful analogies.

2.1 The Age and Patient Gender

Then "Data_Entry_2017.csv" file was analyzed to see various features and correlation between them. We found some outliers in the Age values, which was removed to do some real correlation. The age 117 years was considered to be the threshold, as it is considered to be the maximum accepted value. The Figure 1 below show the plot before and after data wrangling.

The "Data_Entry_2017.csv" file was recently (April 20202) updated to "Data_Entry_2017_v2020.csv" to correct errors in the follow-up numbers and ages of some patients. Figure 1 C now shows that the outliers were removed in the new updated file at NIH website.

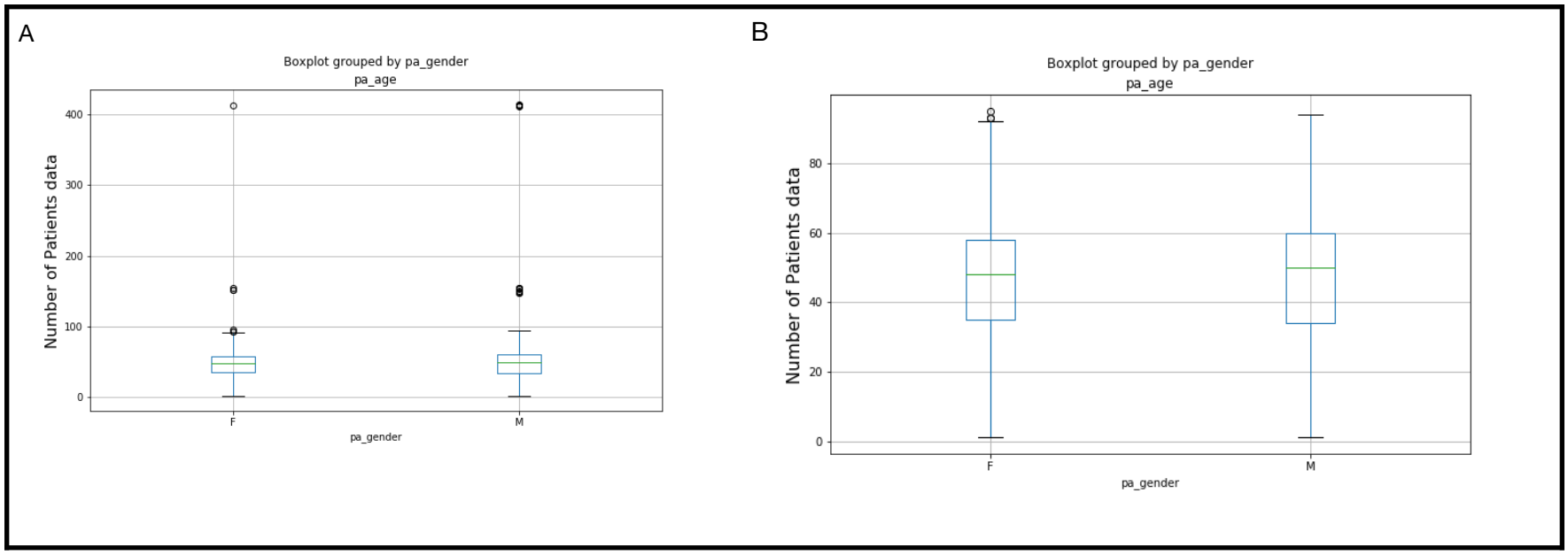

Figure 1. The box plot exhibit the outliers in the feature “Age” . Considering the reliable age to be 117 years of a human , the plot B gives the plot after removing the outliers.

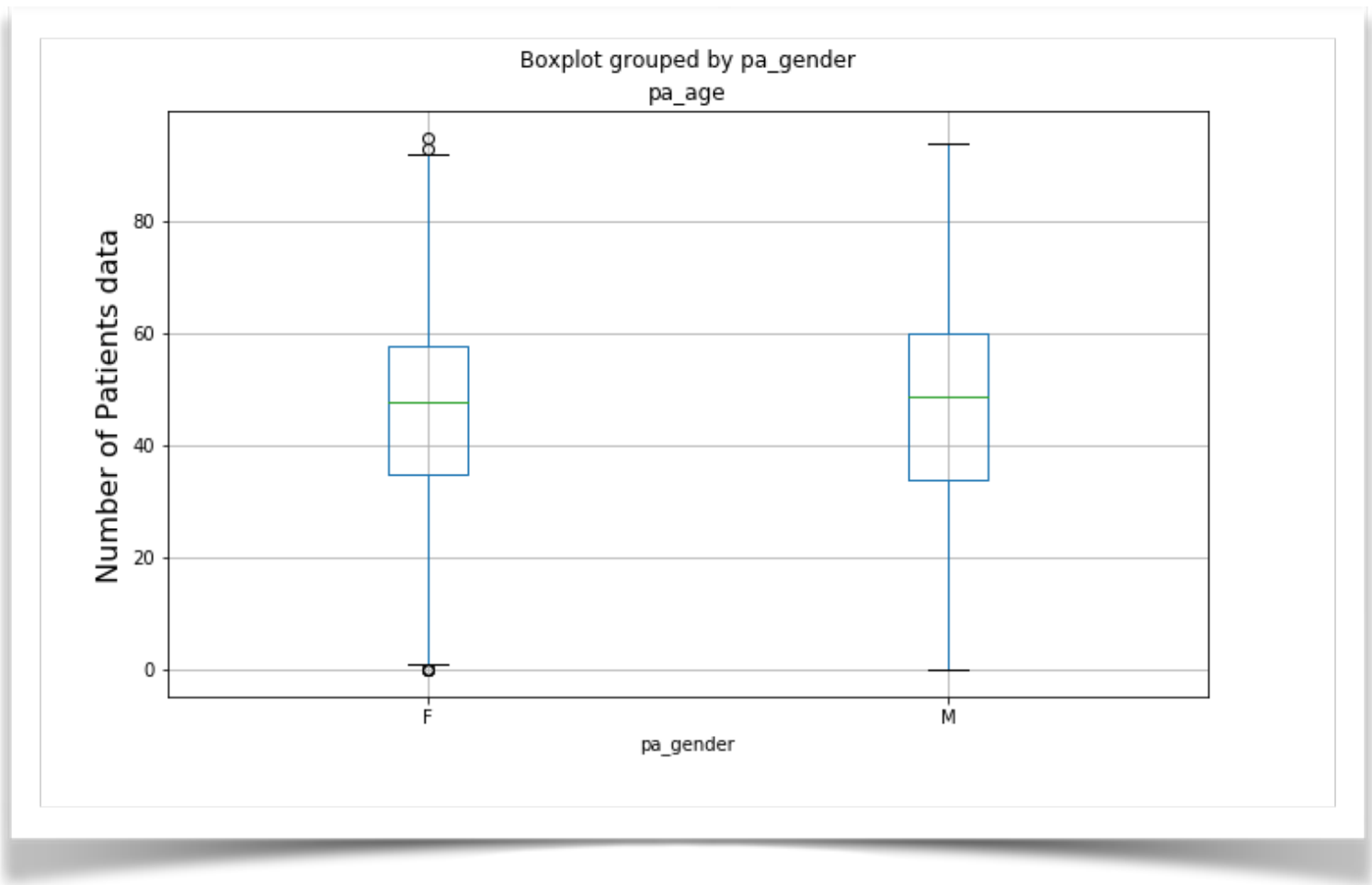

Figure 1 C. The update plot of patients with age from file (Data_Entry_2017_v2020.csv)

The Figure 2 shows the Frequency distribution of Patient Gender in CXR data set. Point to be noted here is that this not only includes the CXR images of infectious patient but also after the treatment labelled as ‘No Findings’ in the “Data_Entry_2017.csv" data set , as can be seen from this IPython notebook. At least we can make a hypothesis that the marked infectious diseases are more related to female as compared to men.

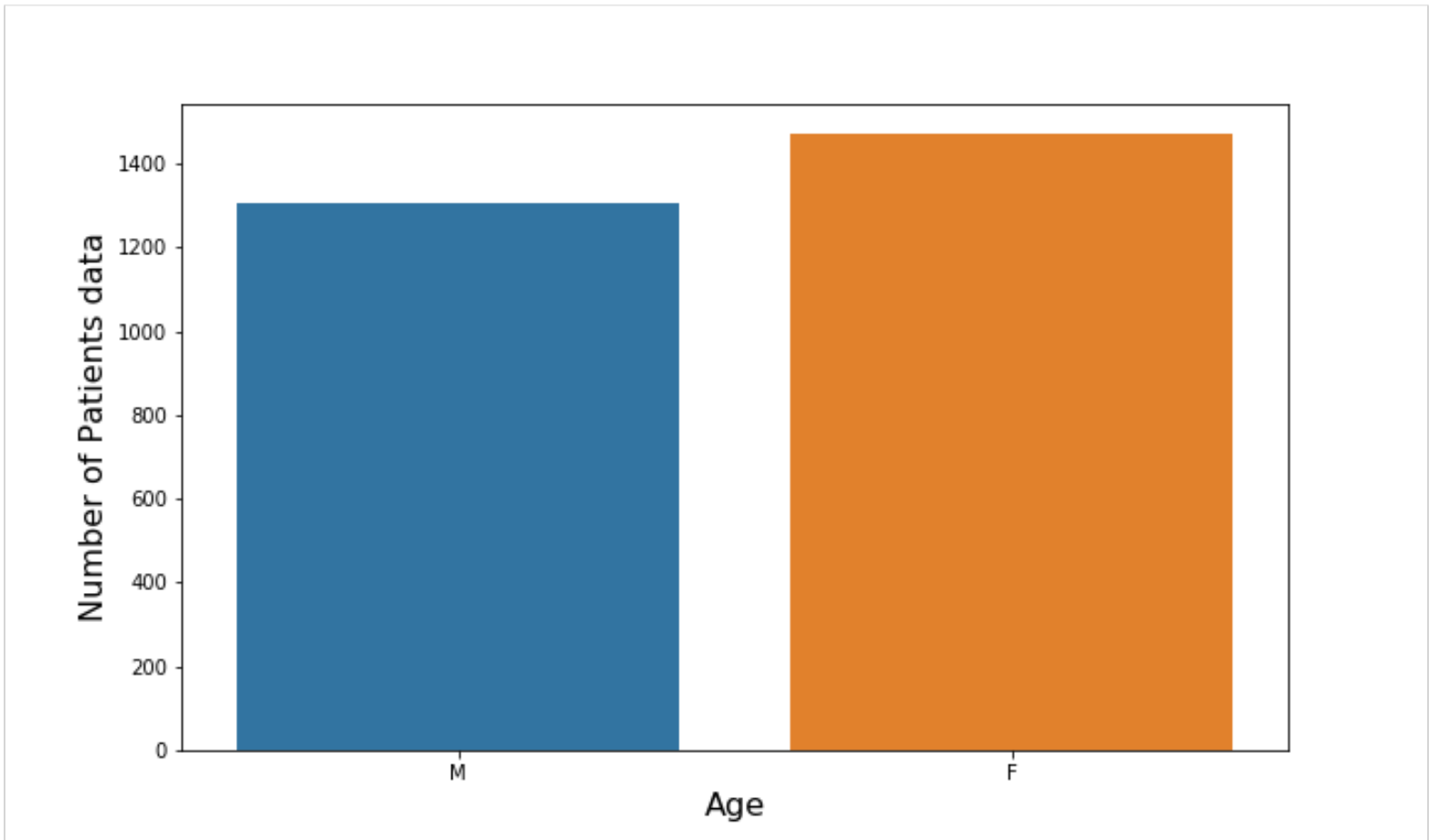

Figure 2. Frequency distribution of Patient Gender in CXR data set

2.2 The image label accuracy

The image labels are mined from the radiology reports using Natural Language Processing(NLP) techniques. The 14 disease keywords are purely extracted from the reports. The radiologists have multiple sources available together with the CXR images to assign labels for diseases. This gives more than one label to major CXR images. Figure 3. is the screen short showing the labels with multiple diseases. Images labeled with ‘No finding’ could contain disease patterns other than the listed 14 or uncertain findings within the 14 categories referring to IPython notebook. In this project it has been considered as images with no diseases or “normal”.

```
In [21]: print(df['labels'].unique())
```

```
['Cardiomegaly' 'Cardiomegaly|Emphysema' 'Cardiomegaly|Effusion'
 'No Finding' 'Hernia' 'Hernia|Infiltration' 'Mass|Nodule' 'Infiltration'
 'Effusion|Infiltration' 'Nodule' 'Emphysema' 'Effusion' 'Atelectasis'
 'Effusion|Mass' 'Infiltration|Mass' 'Infiltration|Mass|Pneumothorax'
 'Mass' 'Cardiomegaly|Infiltration|Mass|Nodule'
 'Cardiomegaly|Effusion|Emphysema|Mass'
 'Atelectasis|Cardiomegaly|Emphysema|Mass|Pneumothorax' 'Emphysema|Mass'
 'Emphysema|Mass|Pneumothorax' 'Pneumothorax' 'Emphysema|Pneumothorax'
 'Atelectasis|Pneumothorax' 'Cardiomegaly|Emphysema|Pneumothorax'
 'Mass|Pleural_Thickening' 'Cardiomegaly|Mass|Pleural_Thickening'
 'Pleural_Thickening' 'Effusion|Emphysema|Infiltration|Pneumothorax'
 'Emphysema|Infiltration|Pleural_Thickening|Pneumothorax'
 'Effusion|Pneumonia|Pneumothorax' 'Effusion|Infiltration|Pneumothorax'
 'Effusion|Infiltration|Nodule' 'Atelectasis|Effusion|Pleural_Thickening'
 'Fibrosis|Infiltration' 'Fibrosis|Infiltration|Pleural_Thickening'
 'Fibrosis' 'Infiltration|Mass|Nodule' 'Cardiomegaly|Edema|Effusion'
 'Atelectasis|Effusion|Infiltration'
 'Atelectasis|Consolidation|Edema|Pneumonia' 'Consolidation'
 'Edema|Infiltration' 'Edema' 'Cardiomegaly|Consolidation'
```

Here the data is analysed with the frequency of occurance of each named deseases.

Figure 3. The screen short of the ‘labels’ marked with multiple diseases

2.3 The unique labels

Figure 4 identifies the uniques ‘labels’ or the 14 diseases. It shows that ‘Infiltration ‘is the dominant cases and is mostly accompanied with other pathologies. Infiltration is the diffusion or accumulation of foreign substances or in amounts in excess of the normal in the lungs.

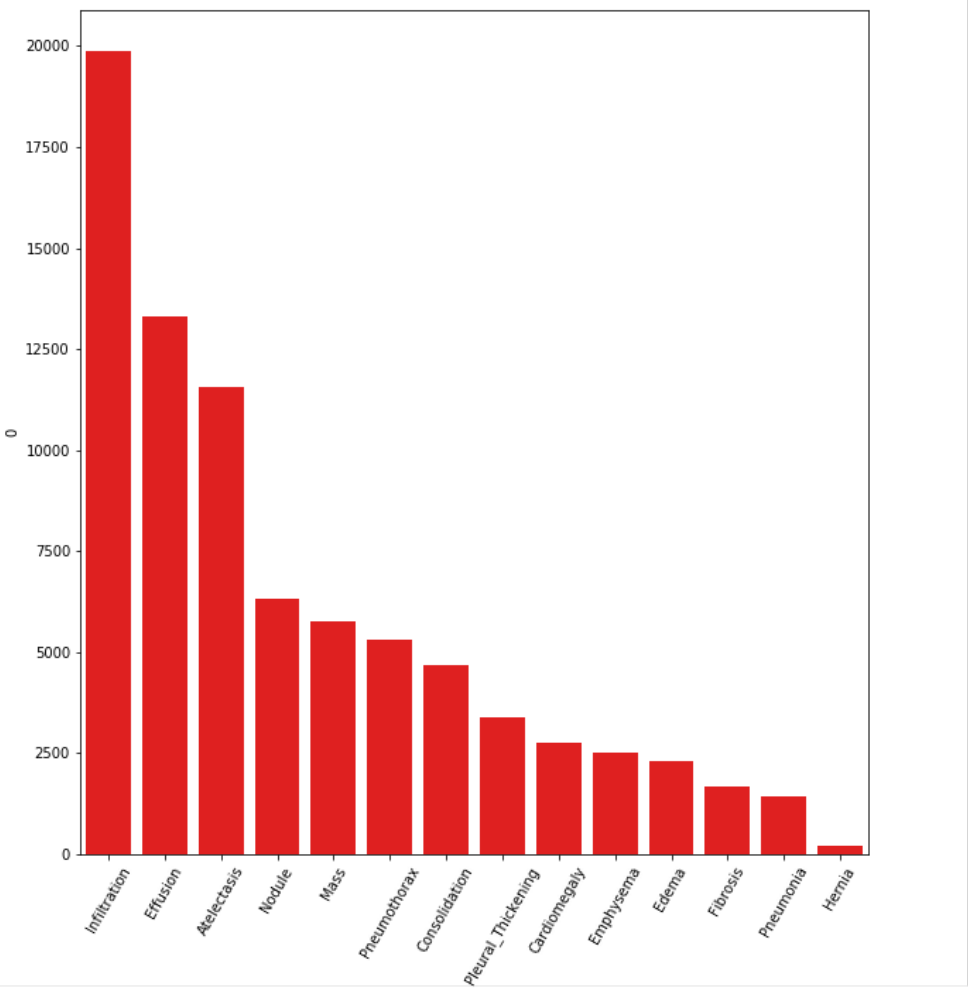

Figure 4.Number of Patients with unique infection of the 14 diseases identified

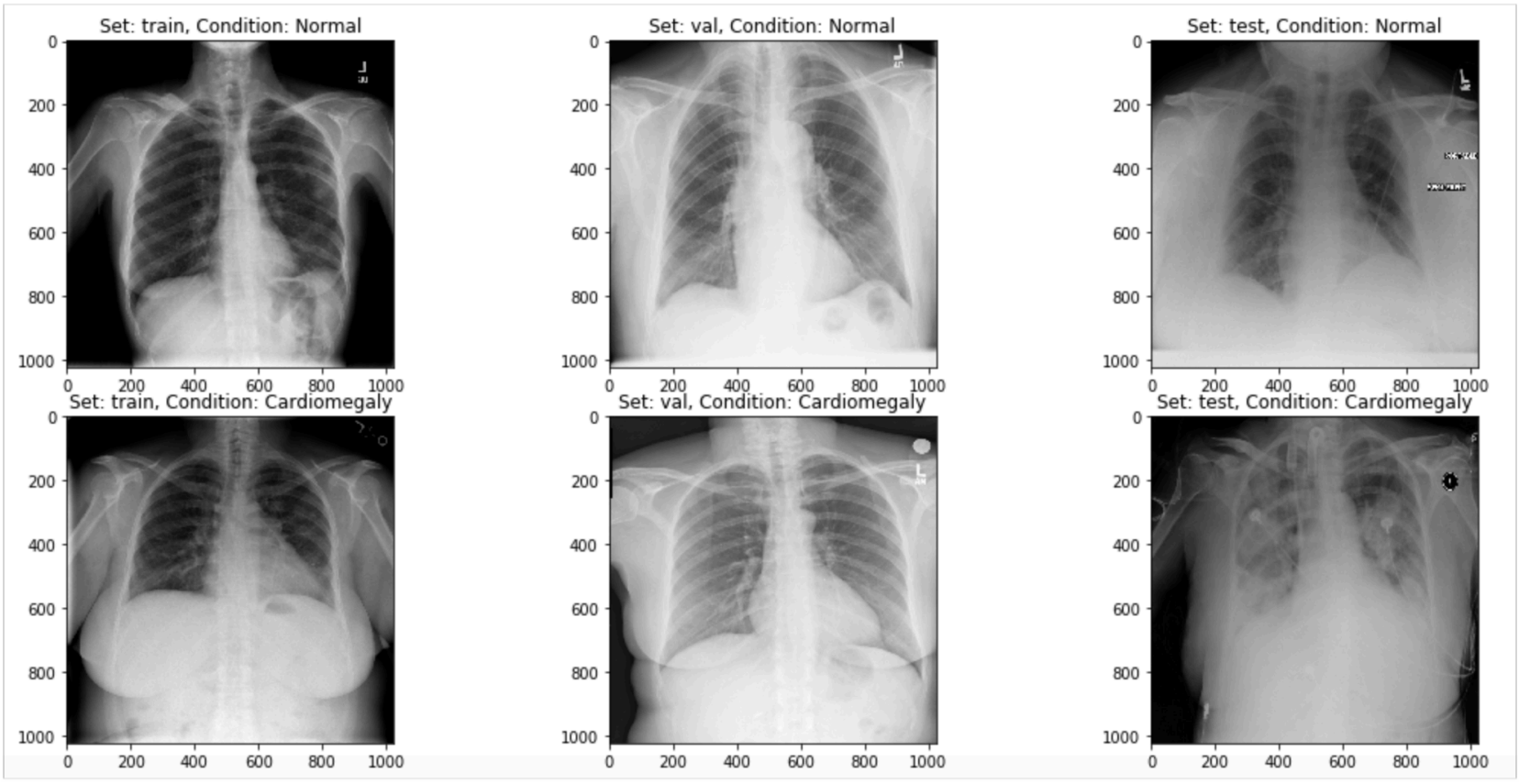

Figure 5. Chest Xray images selected randomly from the labels as Normal and one marked as Cardiomegaly

3. Data Exploration

All the images from the NIH website were downloaded and unzipped by this IPython notebook. A random image selection with marked 'labels' shows that it's not easy to identify the infection or to say they are not definite pattern to say the image has Cardiomegaly as shown in Figure 5. This requires a definite and precise models which can train the machine to have a quick and precise results to identify the disease.

It is hard if not impossible to distinguish certain pathologies solely based on the findings in the images. However, other information from multiple sources is supposed to be available to the radiologists (e.g. reason for exam, patients' previous studies and other clinical information) when he/she reads the study. The diagnostic terms used in the report (like 'cardiomegaly') come from a decision based on all of the available information, not just the imaging findings.

4. Modeling

Knowing the labels for Cardiomegaly, i.e. 0 for Normal i.e. "No Findings" and 1 for "Cardiomegaly", we use supervised Machine learning together with Deep learning algorithms to build a predictive model. Furthermore, since there are only two outcomes (or classes) in the data (0 and 1), we use binary classification algorithms.

4.1 Data Pre-processing

The data provides the list of images to be considered as training and test sets in form of txt file as "train_val_list" and "test_list". The model made here requires a validation list. We randomly selected the image data from the training list by using this IPython notebook. The models are trained using the 75% of the data, 25% is for validation and the marked test data for testing the the performance of the models as shown in this IPython notebook.

4.2 Modeling Pipeline and Evaluation Metric

Deep learning can be effectively used for spotting anomalies in X-rays to detect various diseases. The application of deep learning to medical images has become easier due to the availability of open-source images.There is a theoretical perspective to understand why (single layer) Convolution Neural Network (CNNs) work better than fully-connected networks for image processing.

We pipelined our model in the following steps. The model has nine convolutional blocks comprised of convolutional layer, followed by the flatten layer and two dense layers. The dropouts functions are applied to reduce the over-fitting of the model. The activation function applied is Relu throughout the model pipeline except for the last layer we applied softwmax as the model is a binary classification problem. The model is able to achieve an accuracy of 89% which is quite good considering the size of data that is used as shown in this IPython notebook. It means that there is 89% chances that the model will be able to predict Cardiomegaly correctly if we say that the Radiologist is 100% correct.

Figure 6 plots the accuracy with 70 epochs. The Figure 6 below shows that the model is converging which can be observed from the decrease in loss and validation loss with epochs. Figure 7 shows that there is an accuracy is reached around 89% in the beginning of the epochs and is very consistent. This seems to be very ideal simple model.

The model was simple but well built as we see its performance with the validation set of CXR images. The performance and effectiveness was performed on the test set of CXR data. On the choice of the evaluation metric we selected, area under the curve (AUC) of receiver operating characteristic (ROC) curve and AUC of precision recall (PR) curve and F1 score through confusion matrix.

The effectiveness and the performance measurement of the classification Machine Learning model is shown in Figure 8. We have good number of true positive cases of 1027, with few false negative cases of 509.

The ROC curve as shown in Figure 9 having an AUC as ~0.7 is considered to be acceptable although not ideal.

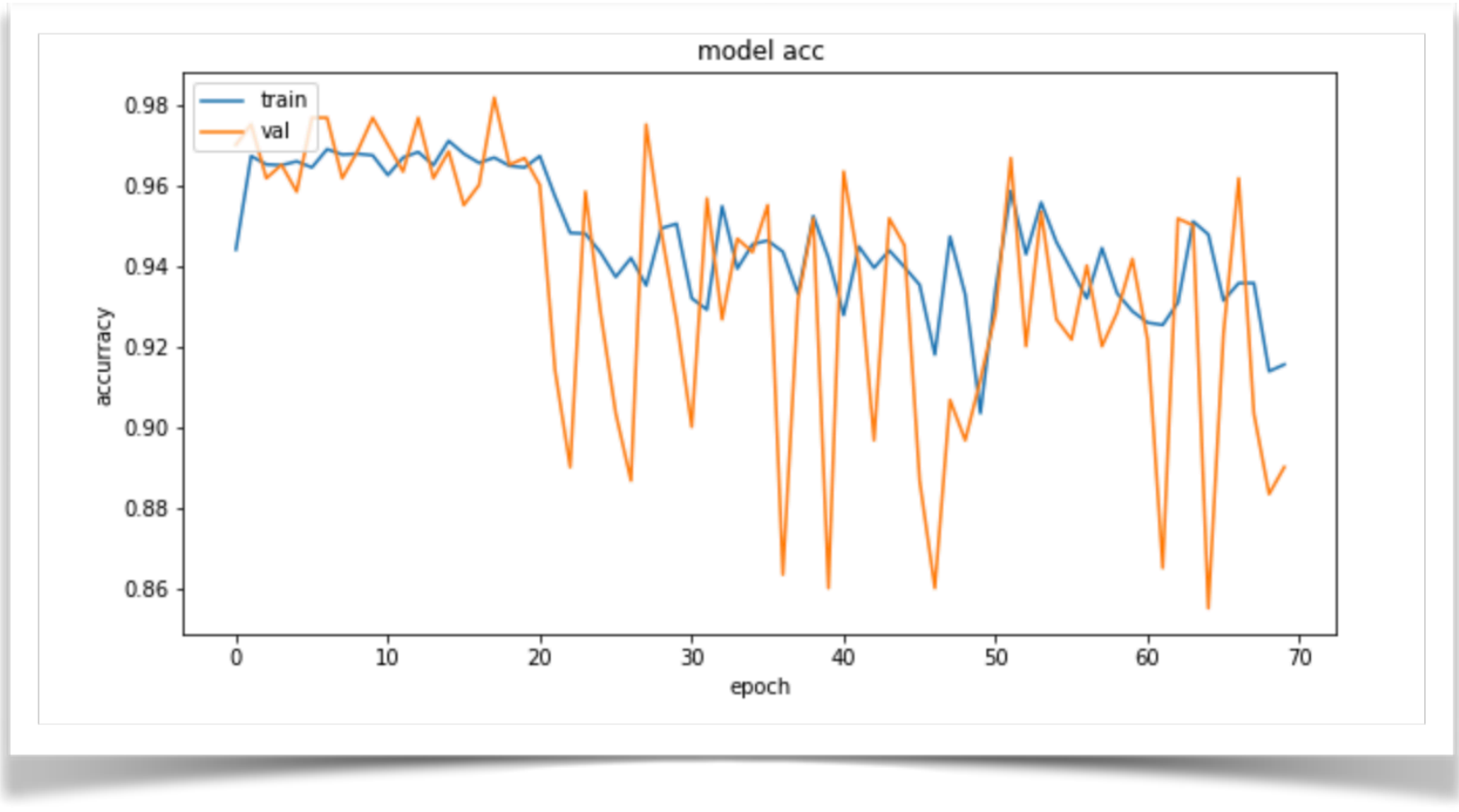

Figure 6. The model accuracy with 70 epochs

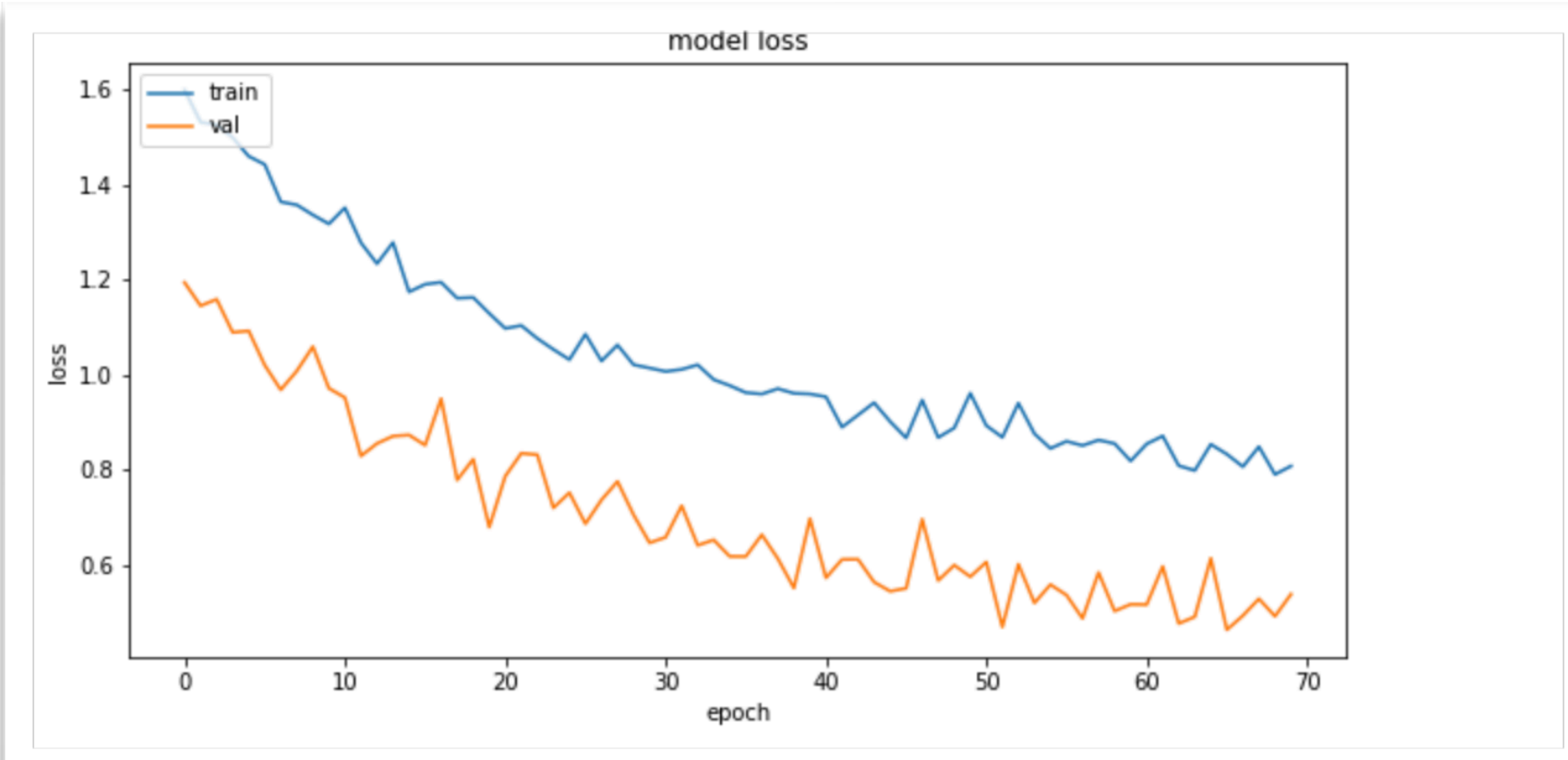

Figure 7. The model validation loss with the epoch

```
CONFUSION MATRIX ------------------
[[1027  201]
 [ 509  560]]

TEST METRICS ----------------------
Accuracy: 69.09011754462342%
Precision: 73.58738501971091%
Recall: 52.38540692235735%
F1-score: 61.20218579234973

TRAIN METRIC ----------------------
Train acc: 91.55
```

Figure 8. The confusion matrix and test metrics

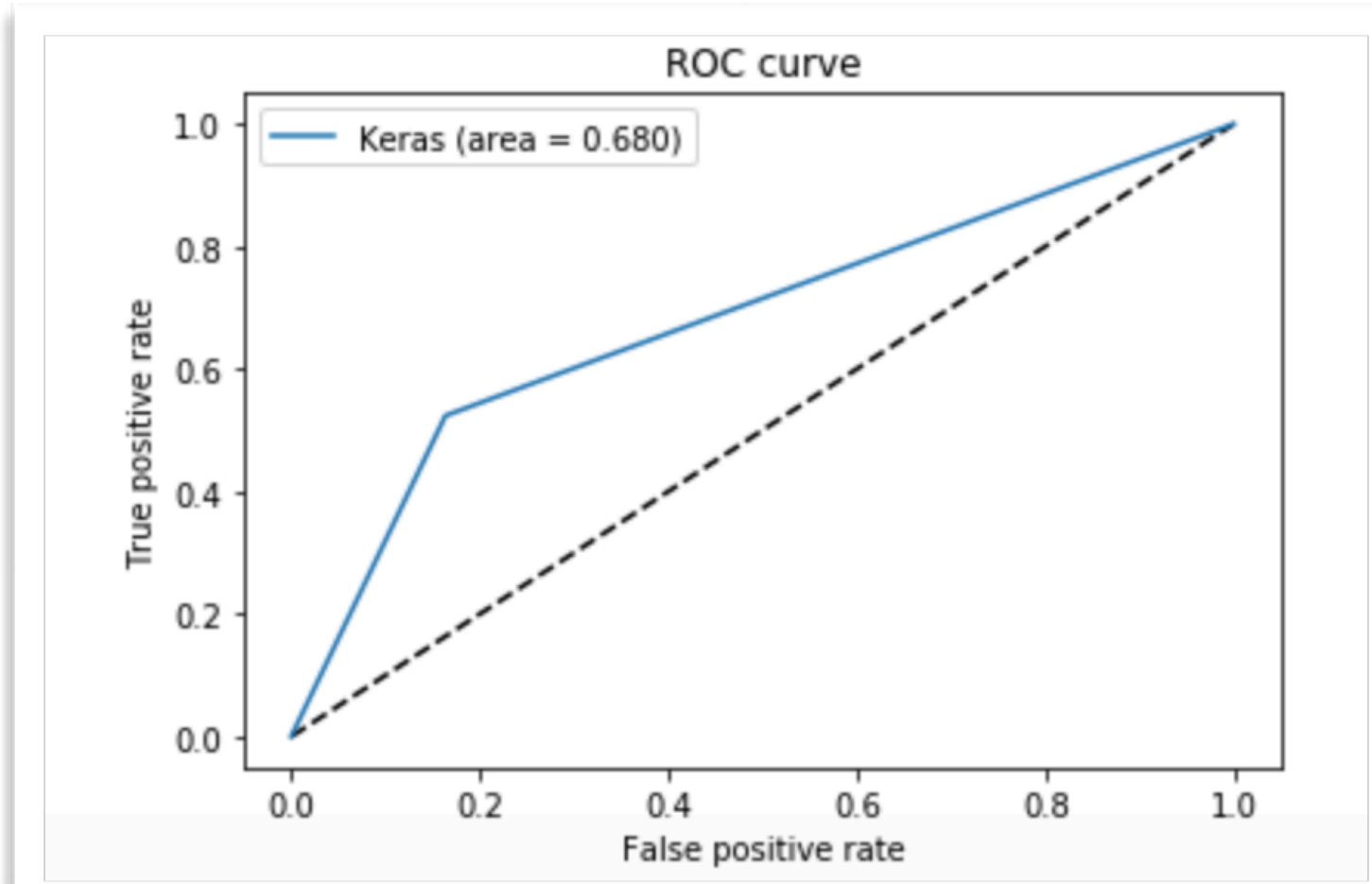

Figure 9. The ROC curve for the model

5. Assumptions and Limitations

Most existing computer vision neural networks are designed for colorful natural images and takes advantage of the rich textures present in them. This makes it hard to directly apply off-the-shelf solutions on CXR. The number of image data are still limited for training.

The major issue which we had to deal with in this Deep Learning model was the class_weight. As the number of images marked as normal, that is the class label “0” was 30 times more than that marked as Cardiomegaly, that is class label ”1”. The model had to fine tune with different class weights in order to get an acceptable ROC curve.

Another issue which we encountered during the testing phase of the model was the ratio of number of binary class. It was observed that a balanced ratio was better predicting the model for test CXR images as compared to an unbalanced labels.

6. Conclusion

Although this project is far from complete but it is remarkable to see the success of deep learning in such varied real world problems.

The Chest Xray images encourages the data science community and the group of radiologists to share their own labels, so that observer variability can also be assessed. The published image labels will be a great initiative enabling other researchers to start looking at the problem of ‘automated reading a chest X-ray’ on a very large dataset, and therefore the labels will be improved by the community. The heat map would have been an ideal method to have a complete training and AI to read Chest X-rays like Radiologists.