Multi-Classification of Chest X-Ray Images

(NIH data)

Neeharika Sinha

April 30th 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 any 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.

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 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.

This project analysis the Chest X Ray data for 14 Common Thorax Disease and possible dependency of the features provided in the dataset. The Machine learning and the deep learning algorithms are build specifically to identify the "Cardiomegaly" infection.

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, 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 Patient Gender

The Figure below [1] shows the Frequency distribution of Patient Gender in CXR data set. Point to be noted here is the this not only includes the CXR images of infectious patient but also after the treatment labelled as ‘No Findings’ in the csv data set , as can be seen from this IPython notebook. At least we can make a hypothesis that the marked infectious diseases are related to female as compared to men.

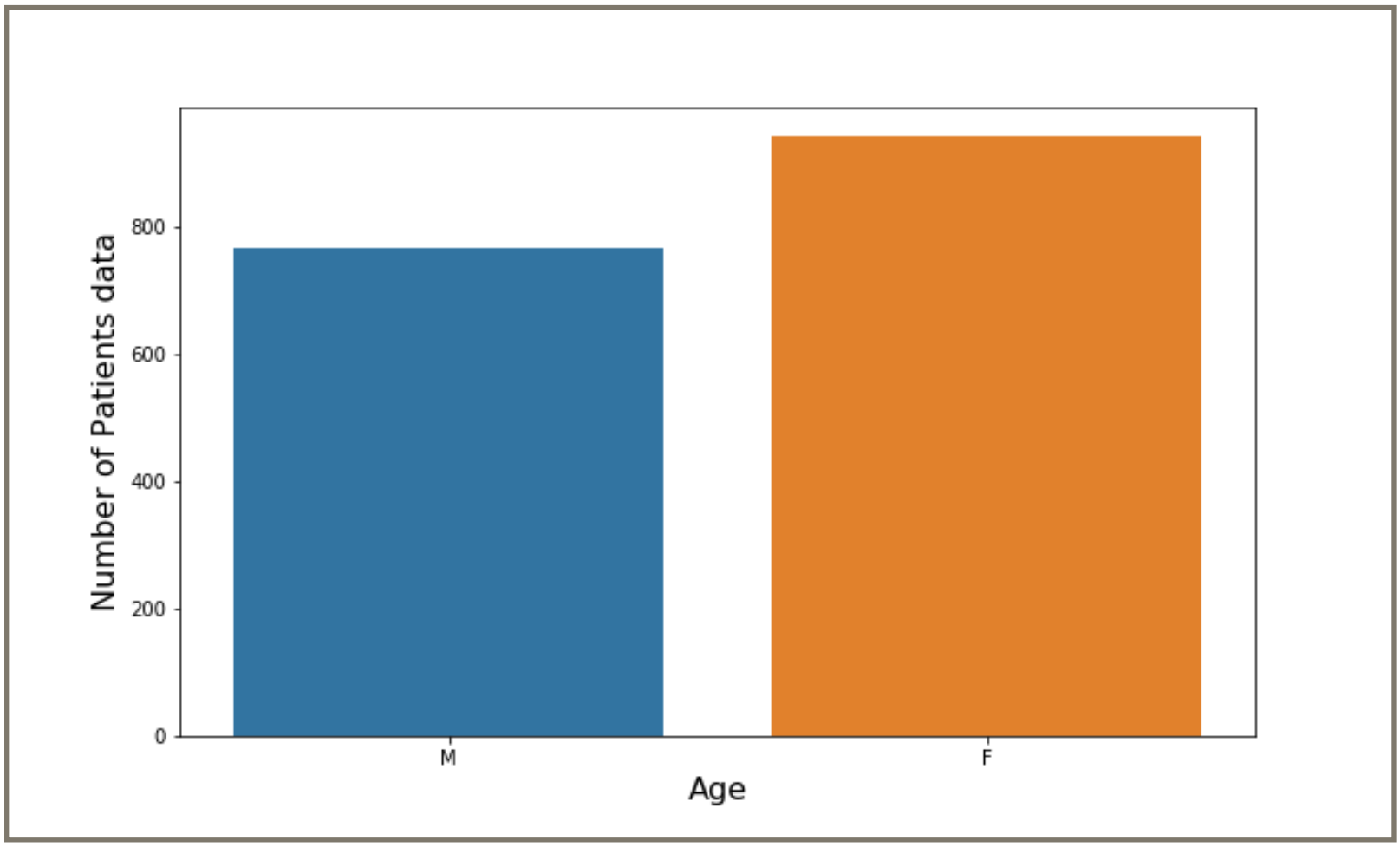


Figure 1. Frequency distribution of Patient Gender in CXR data set

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 2. 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.

```
In [7]: print(tidy_df['labels'].unique())

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

Figure 2. The screen short of the ‘labels’ marked with multiple diseases.

2.3 The unique labels

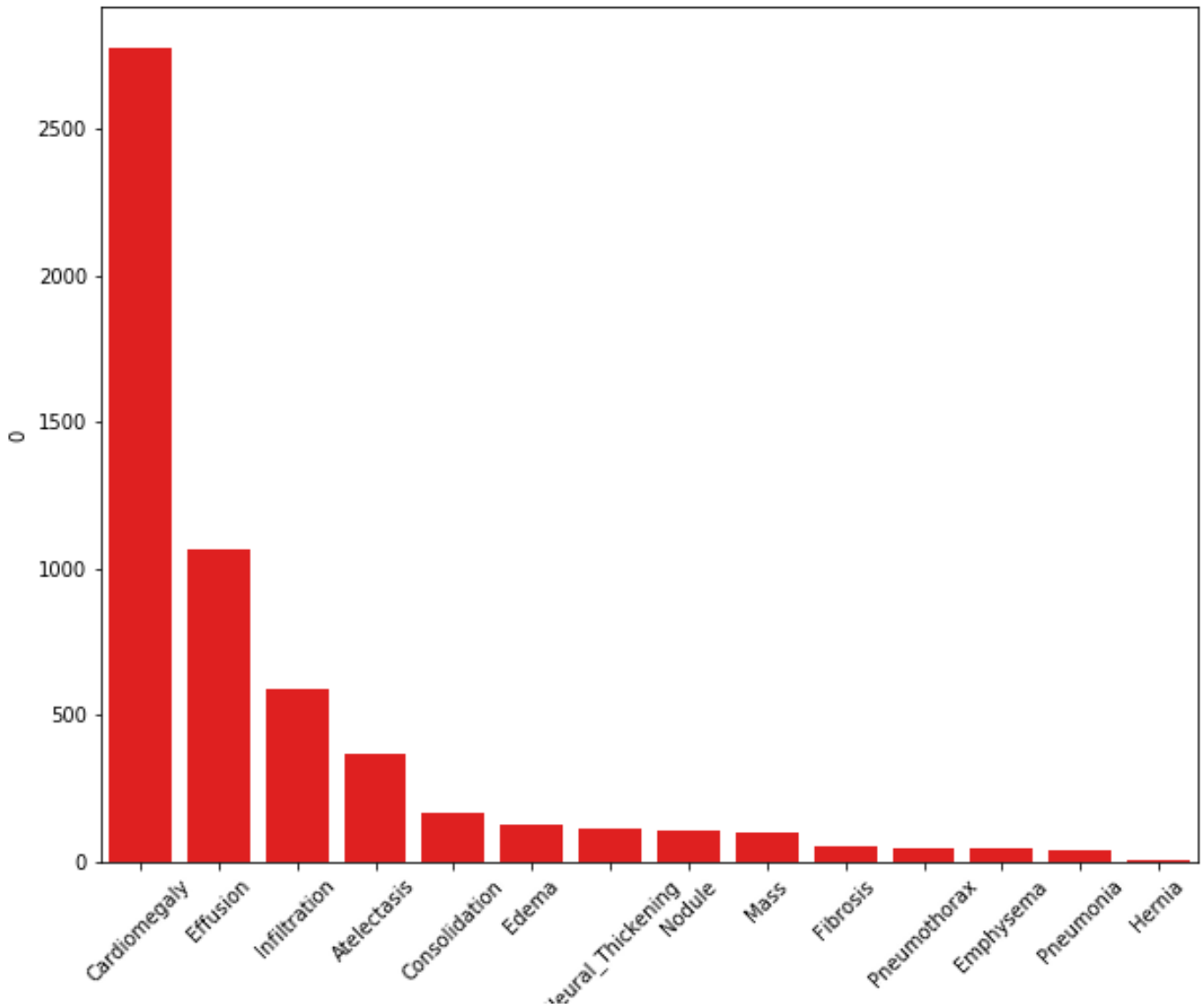


Figure 2. Number of Patients with unique infection of the 14 diseases identified.

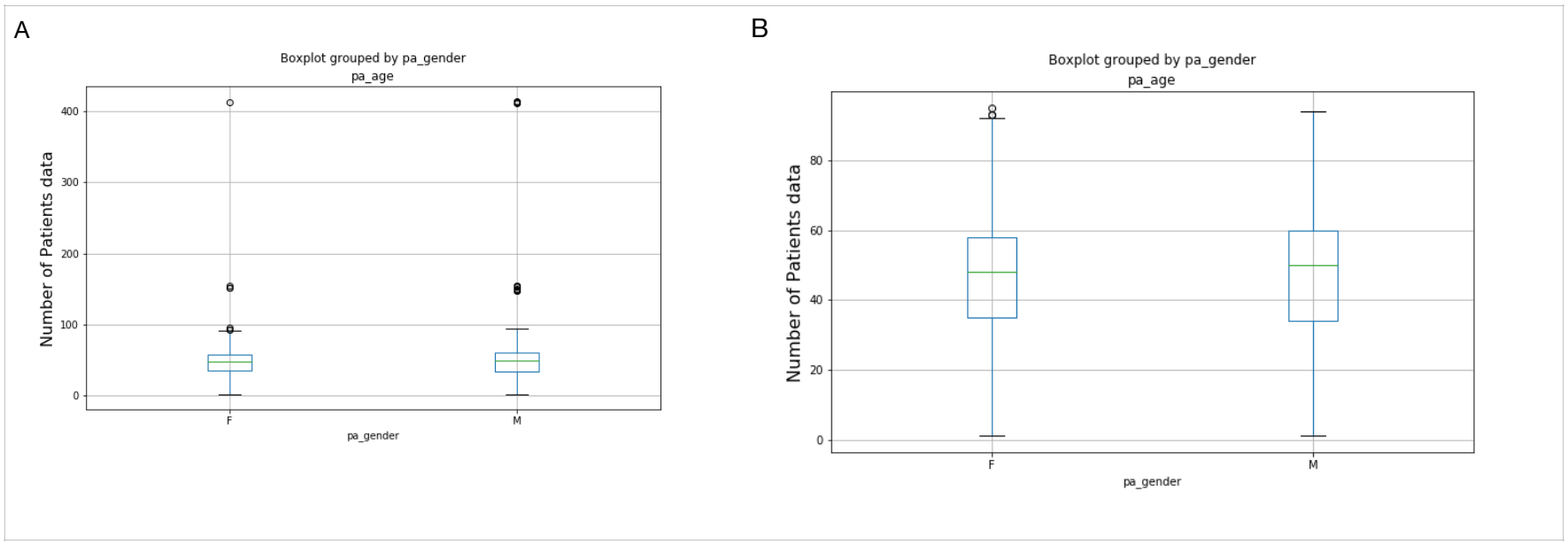


Figure 3. The box plot showing the outliers in the feature “Age” . Considering the reliable age to be 117 years of a human , the plot B gives the plot after removing the outliers.

3. Data Exploration

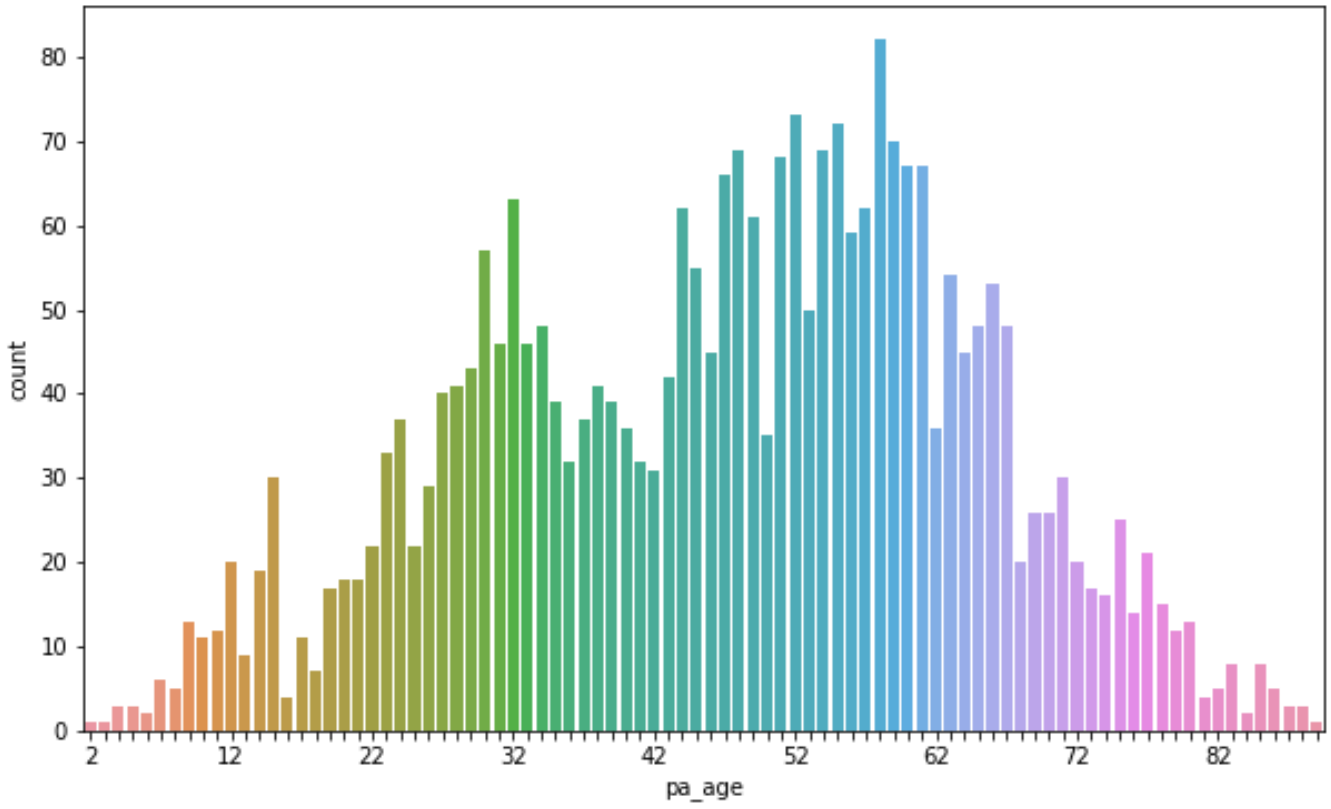


Figure 4. Population of data after removing the outliers for age.

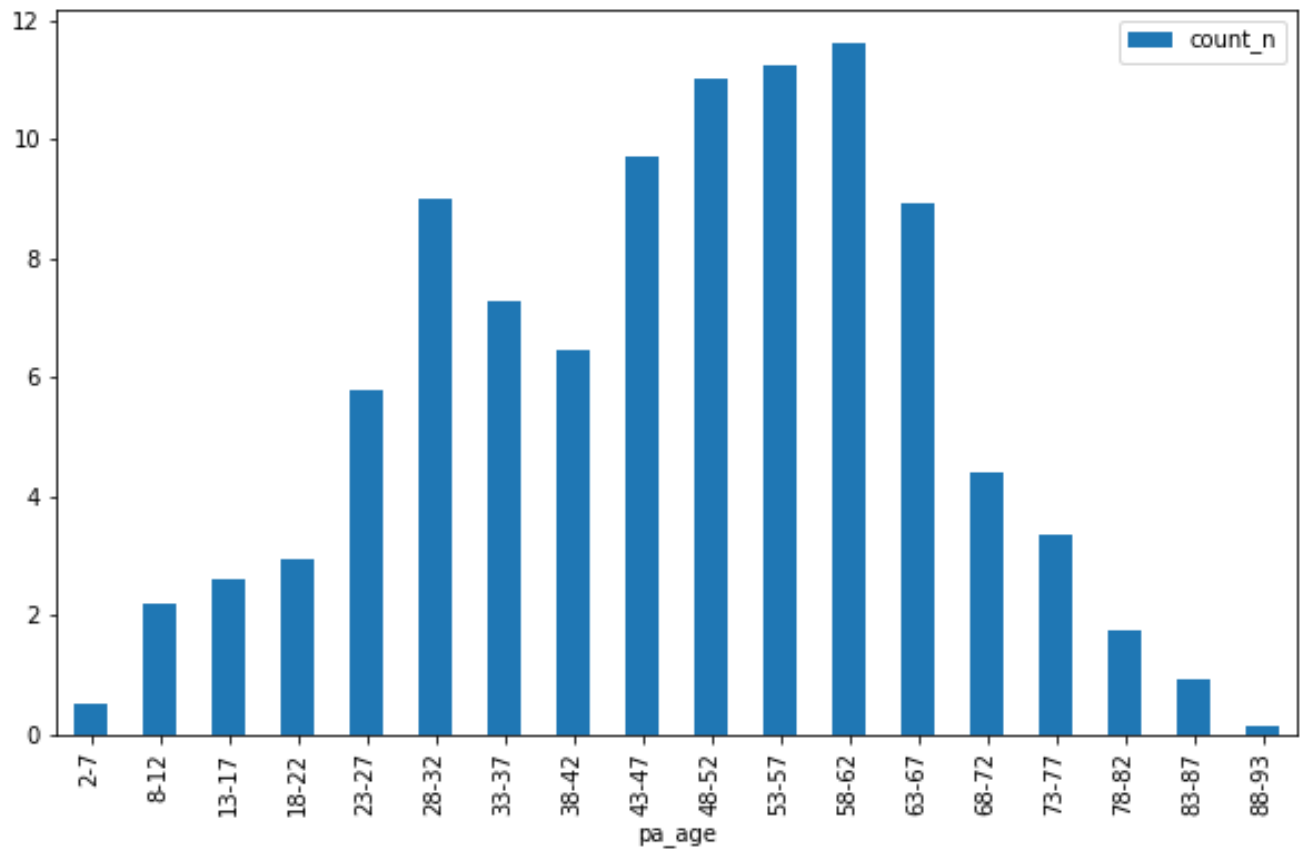
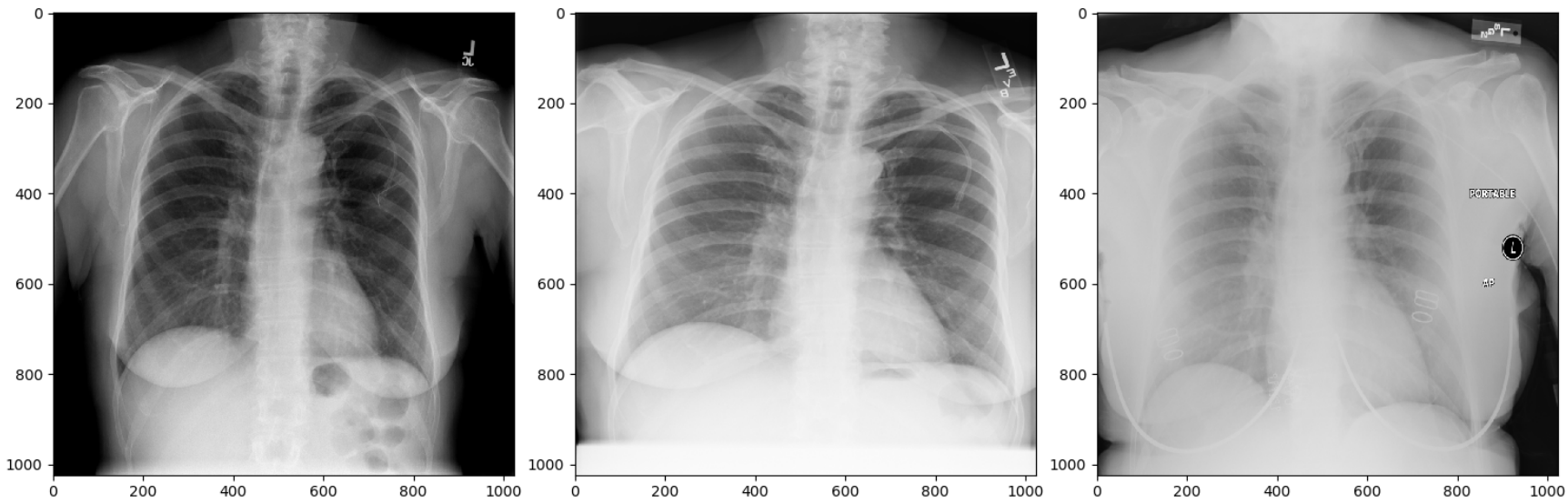


Figure 5. The number of cases with Cardiomegaly is now normalised to total number of data sets.

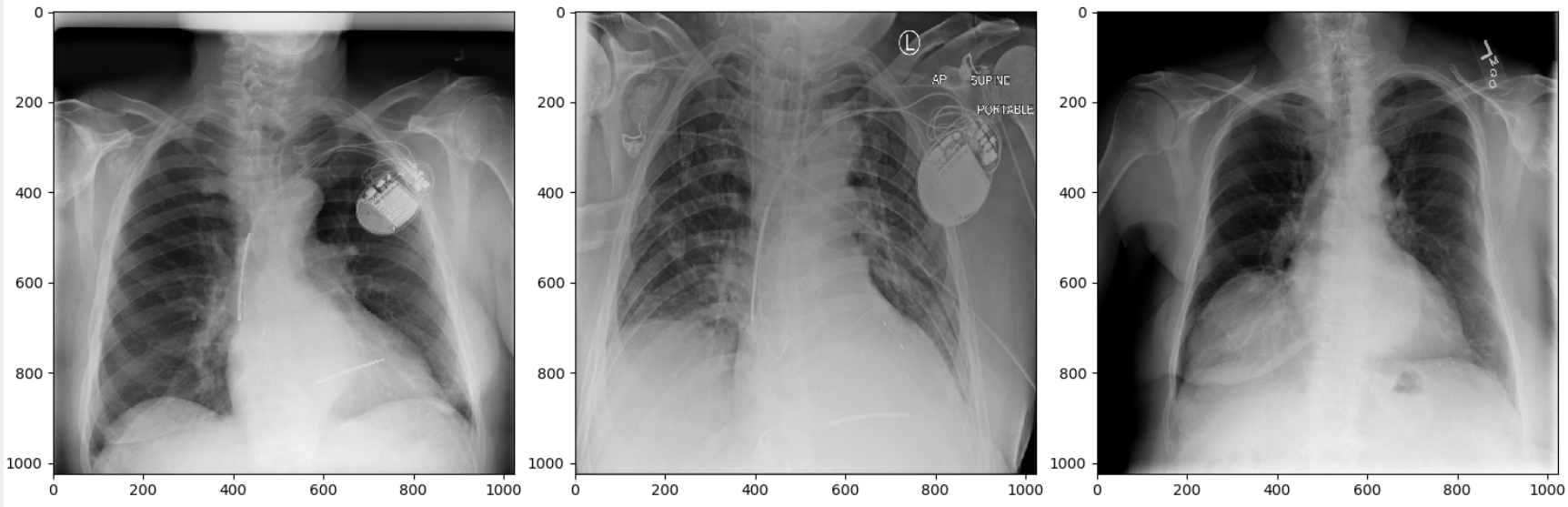
4. Automated reading a Chest X-ray

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.

Normal CXR



Infectious(Cardiomegaly)



4. Modeling

5. Using Model and Recommendations

6. Conclusions