**Data Science Project Protocol**

*Author(s):*

*Michal Gimmon*

*Liron Walter*

# Introduction

X-ray imaging is one of the most widely used diagnostic examinations for the diagnosis and research of numerous types of illnesses, making it a high-potential source of data for the development of computer-aided detection and diagnosis tools.

Previous work has been done in the past in the field of X-ray images classification, (for example, the identification of Tuberculosis) using various deep learning techniques, yet it is still a challenge to be explored, as the available data is limited to only thousands of images for training, and poor image-labeling techniques which impact classification accuracy.

The dataset at hand is comprised of 112,120 chest X-ray images of 30,805 unique patients, which were classified into 14 different types of lung diseases (including images which demonstrate multiple diseases), and images of healthy individuals.

Several groups have addressed this multi-label dataset in attempt to classify the images according to different research goals - some classified all 14 classes (diseases) [1], [2] or part of the diseases [3], some identified patients with one specific disease vs healthy patients, and some looked for the correlation between two (or more) diseases.

Due to the complexity of multi-label classification in this dataset, we decided to simplify the problem and focus on classifying the images to patients who demonstrate Effusion (a fluid in the space around the lung), other diseases or healthy (3 classes), using mainly CNN (deep learning method) and/or other ML models (e.g. SVM). The main reason for this decision is the high frequency of the disease in the available dataset. Moreover, according to radiologist [4], this phenomenon labeling seem to be more accurate than other diseases.

As mentioned above, the outcome of this classification challenge will also be affected by the fact that the disease labels in this dataset were retrieved using text mining (NLP) tools from radiologists’ reports, which may lead to seemingly false identifications and impact the model’s accuracy. The task is also challenging in matters of computational resources and memory and is of course much more complicated than our simplified preliminary problem, though is extremely interesting. Our goal is to find a model that predicts Effusion cases better than a random probability.

*[1] Pranav Rajpurkar et al., CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, (2017), URL https://arxiv.org/abs/1705.02315*

*[2] Pranav Rajpurkar et al., Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, (2018), URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002686*

*[3] Xiaosong Wang et al., ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, (2017), URL* [*http://openaccess.thecvf.com/content\_cvpr\_2017/papers/Wang\_ChestX-ray8\_Hospital-Scale\_Chest\_CVPR\_2017\_paper.pdf*](http://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf)

*[4]* [*Luke Oakden-Rayner*](https://lukeoakdenrayner.wordpress.com/)*, Exploring the ChestXray14 dataset: problems, (2017), URL* [*https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/*](https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/)

# Methodology (Project design)

## Data

**Data Sources**

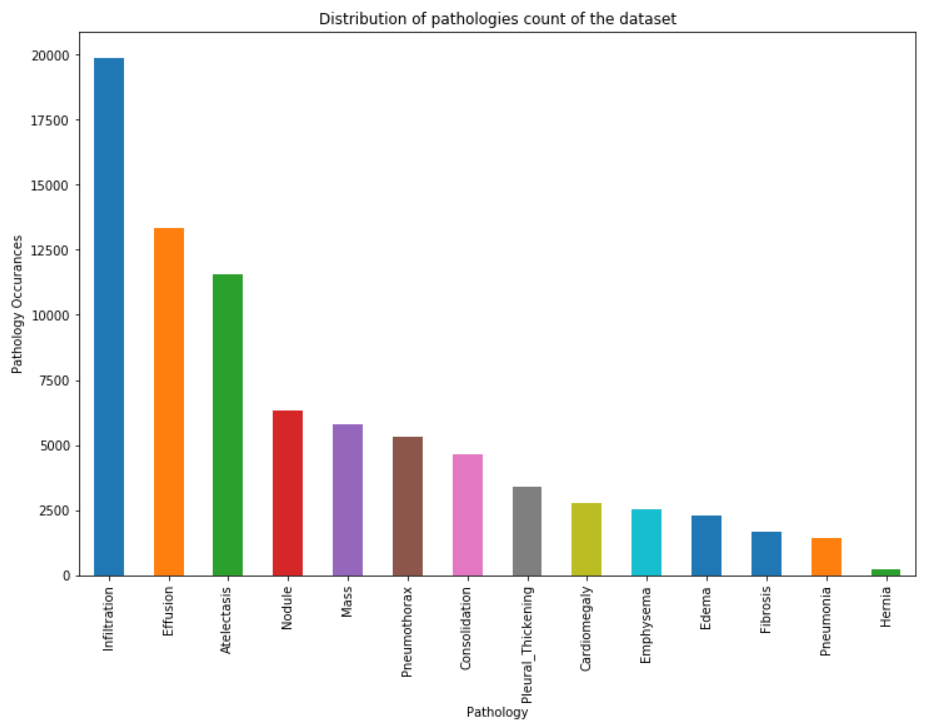
The dataset at hand, as described above, was collected, labeled and published in 2017 by the *NIH Clinical Center* – a clinical research hospital operated by the United States’ National Institute of Health.

The images and its corresponding data were retrieved from the NIH publicly available website - <https://nihcc.app.box.com/v/ChestXray-NIHCC>.

**Defining the Outcome Variable**

As mentioned before,due to the complexity of multi-label classification in this dataset, we have currently decided to simplify the problem and focus on classifying the images between one type of disease, normal or other diseases.

In order to define the disease on which we want to focus, we had to understand the data in terms of the frequency of each pathology, as described below:



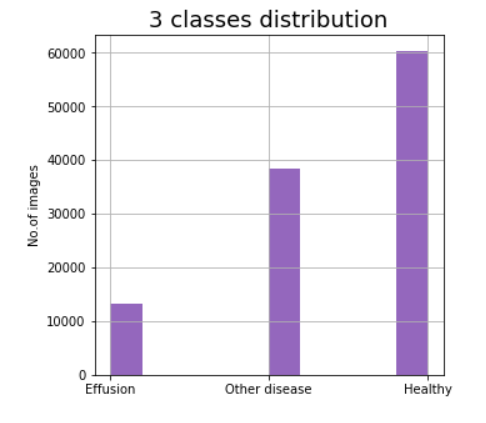
Based on the above, we can see that 'Effusion' is one of the most common observations in this dataset, meaning there is a higher chance for the model to classify it correctly, as it will have more images to be trained on.

In addition, when defining our final outcome, we had to consider the relative accuracy of the disease labeling by NLP, which may significantly affect our final result.

Based on previous work done on this dataset [4], we learned that ‘Effusion’ was labeled with relatively high accuracy, therefore we have decided to define our final classification to the 3 following categories: **Effusion (1), Other disease (2), Healthy (3).**

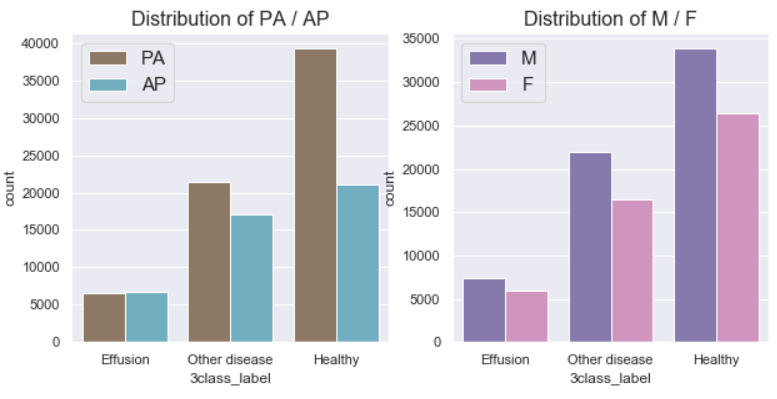
**Exploring the data for correlative / confounding variables**

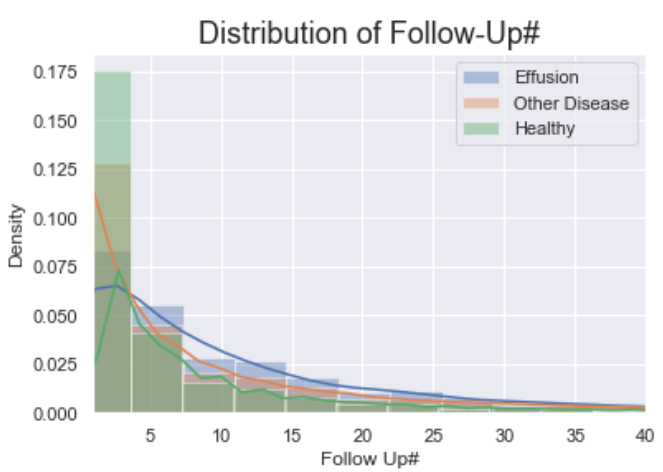
After classifying the images between the 3 new categories we defined, we first checked for the new distribution of the images:

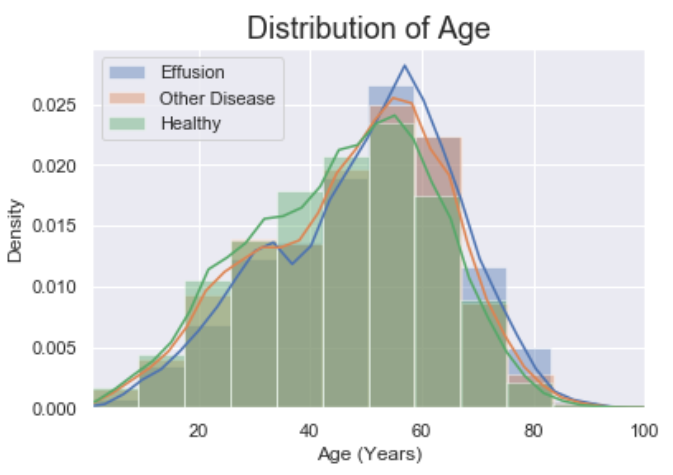


Our 3 groups are comprised of 13,317 images of 'Effusion', 38,442 Images of 'Other diseases' and 60,361 'Healthy' images.

In order to decide if / which variables should be included in our model, and in order to avoid creating biased sub-groups for training and validation, we checked for a relation between the parameters ‘Patient Gender’, ‘View Position’, ‘Patient Age’, ‘Follow-Up #’, and the detection of ‘Effusion’:



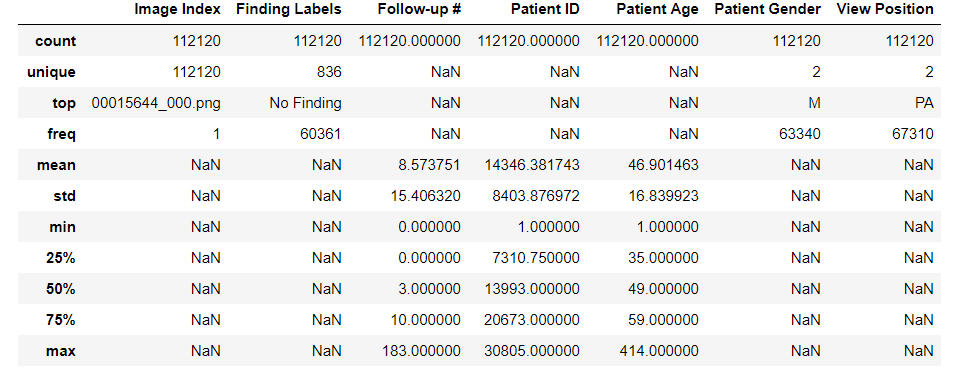




As we can see, the variables ‘View Position’ and ‘Patient Gender’ are distributed almost equally in the outcome population (‘Effusion’).

Moreover, we can see that ‘Patient Age’ is distributed in a relatively normal matter between ‘Effusion’ patients, which means a random selection of images for the training / validation sets will not affect the model’s accuracy.

**Treatment of Outliers and Missing Values**

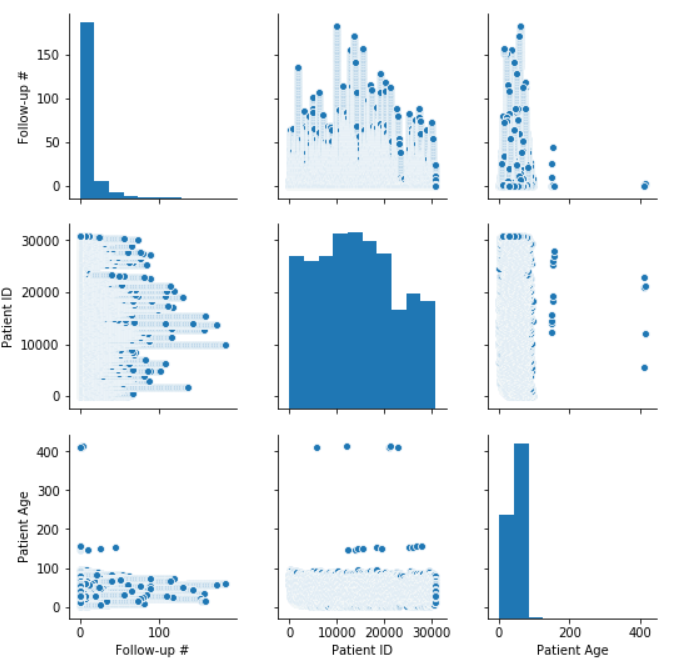


As represented in the table above, there are no missing values for all variables (both categorical and numerical).

According to 'Patient ID', the database includes 112,120 x-ray images of 30,805 unique patients.

It can be observed that 60,361 out of 112,120 images represent healthy subjects ('No Finding'). Moreover, 56% of the subjects are males and 60% of the images' view position is posterioranterior (PA, when patient's back faces the machine).

It can also be observed that there are outliers of patient age, since 414 years is the maximal value, and that the follow-up median # is 3, while the mean is 8.5, meaning the distribution is not normal.



According to the pairplot, there are 5 images with patient age over 400, and another 10 images having age above 100. Moreover, there is no significant correlation between the follow up # and the patient age.

Yet, since we currently are interested in classifying only the images data, we decided not to remove the outliers of the 'age' variables, since we do not mind that they were mistakenly documented. If we would have decided to include the ‘age’ variable in the model, we would have removed these outliers.

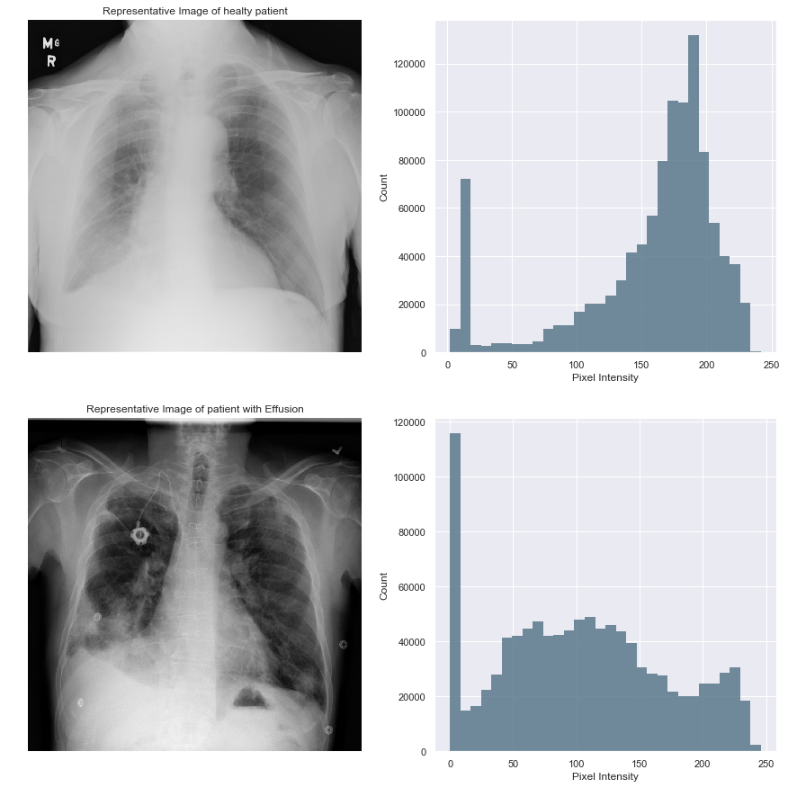
**Images Analysis & Treatment**

* **Images Resizing**

Images original size was 1024x1024, we are currently resizing the images to 512x512 in order to 'save' computing resources which will be required to run the models in later stages. We chose this specific size (and not smaller) to make sure no significant data is lost.

* **Pixels Histograms of representative images**

In order to understand the patterns of pixels distribution between an image which demonstrates ‘Effusion’ and an image of a healthy patient we analyzed two representative images:



## 

From exploring representative images (see additional sample images in the Jupyter notebook), we find that the variability in images’ pixels intensity is extremely high - many of the images are highly saturated / dark, which means they may act as outliers.

We plan to overcome those outliers and enrich our data by doing the following:

1. Enlarge the minority group (‘Effusion’) to 33% (by different data augmentation techniques such as modifying the contrast), and randomly reduce the majority group (‘Healthy’) to 34%. Third group (‘Other diseases’) will remain the same.
2. Localization – define a specific image area to be used for training the model (based on the common appearance area of ‘Effusion’ according to literature).
3. Creation of new variable for pixels’ intensity cutoff – an indication for images with extreme intensity values which will be assigned lower weights when designing the model

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how did you get the final prediction. You have to discuss also the limitations of the model, when it can be used and when not.