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) vs healthy patients, 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/)