

---

# Diagnosing Thoracic Diseases Using Machine Learning and Medical Imaging

---

Wendy Carvalho (02026116), Meriem Elkoudi (02015993), Chris Peters (01989716), Saim Siddiqui (02018510), and Amitha Thalanki (02077527)

## Abstract

**TODO: Write a concise summary of the project and the conclusions of the work. It should be no longer than one short paragraph (e.g. 200 words).**

## 1. Introduction

Thoracic diseases, including pneumonia, emphysema, and fibrosis, are common causes of morbidity and mortality worldwide. Chest X-rays are one of the most widely accessible and commonly used diagnostic tools for detecting these conditions. However, interpreting these images is complex, and clinical diagnoses are often challenging, even for experienced radiologists.

Machine learning techniques are used to address this challenge by enabling automated analysis of medical images. Bringing this leading-edge technology's immense power into the diagnostics field can potentially increase the capacity for trained professionals to diagnose patients properly, sooner, and in more difficult-to-identify cases.

This project aims to build a computer-aided diagnostic (CAD) tool capable of diagnosing thoracic diseases from chest X-ray images, assisting clinicians by automating the detection of common thoracic diseases. This has been implemented using Support Vector Machines (SVM) and Convolved Neural Networks (CNN) in the form of ResNet50 and MobileNet. The efficacy of all are tested.

## 2. Data

**TODO: Describe the data used for experiments and report data statistics as well as interesting observations or patterns in the data.**

The data for this project has been obtained from the NIH ChestX-ray8 dataset, an open-source, hospital-scale collection of medical images, which will be essential for developing a CAD system. The original NIH ChestX-ray8 dataset contains 108,948 images from 32,717 patients, however, the labels are frequently incorrect and misdiagnoses. A subset of 810 images was chosen based on the evaluation of radiologists themselves. The 810 image subset is the only

commonly known verified chest x-ray dataset.

The data can be classified into 14 different labels, with the capability of multi-label classification. The "None" label is also a possibility, but only occurs in the absence of any diseases and cannot be combined with any other label. The 14 possible abnormal labels are:

- Hernia,
- Pleural Thickening,
- Fibrosis,
- Emphysema,
- Edema,
- Consolidation,
- Pneumothorax,
- Pneumonia,
- Nodule,
- Mass,
- Infiltration,
- Effusion,
- Cardiomegaly,
- and Other.

In Figure 1, the dataset's distribution of normal and abnormal x-rays is shown, where abnormal is any image with at least one of the 14 aforementioned labels, and where normal is any image with the "None" label. The abnormal category outnumbers the normal category quite a bit, which may contribute to the class imbalance.

As shown in Figure 2, there is a very clear class imbalance, with "None" and "Effusion" occurring the most and "Pneumonia" and "Hernia" occurring the least.

## 3. Methods

**TODO: Provide a detailed description of your method and explain why the method is a good fit for the problem.**

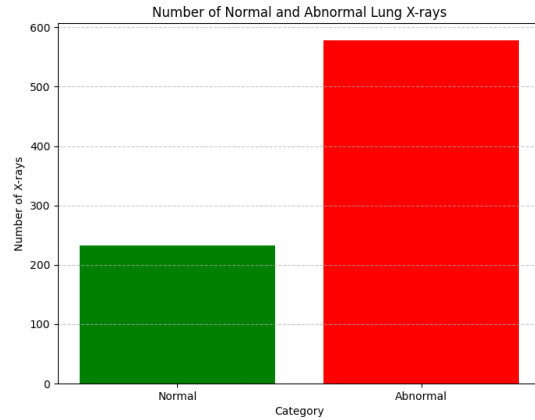


Figure 1. The distribution of normal vs. abnormal lung x-rays

## 4. Results

**TODO:** Briefly describe the evaluation approach and metrics. Report performance metrics for the method(s) through Figures or Tables. Report insights obtained from the results. Good ways to obtain insight are ablation analysis, error analysis, and use of synthetic data.

## 5. Conclusion

**TODO:** In one short paragraph concisely summarize the main points and insights of the project, describe potential directions to extend your project, and describe limitations of your project.

## 6. Contribution Chart

**TODO:** Complete the following Table to clearly report the contributions that each team member made to the final project. (Task/Subtask, Student ID, Commentary on Contribution)

	Student ID	Task	Commentary on Contribution
Name			
Name			
Name			
Name			
Name			

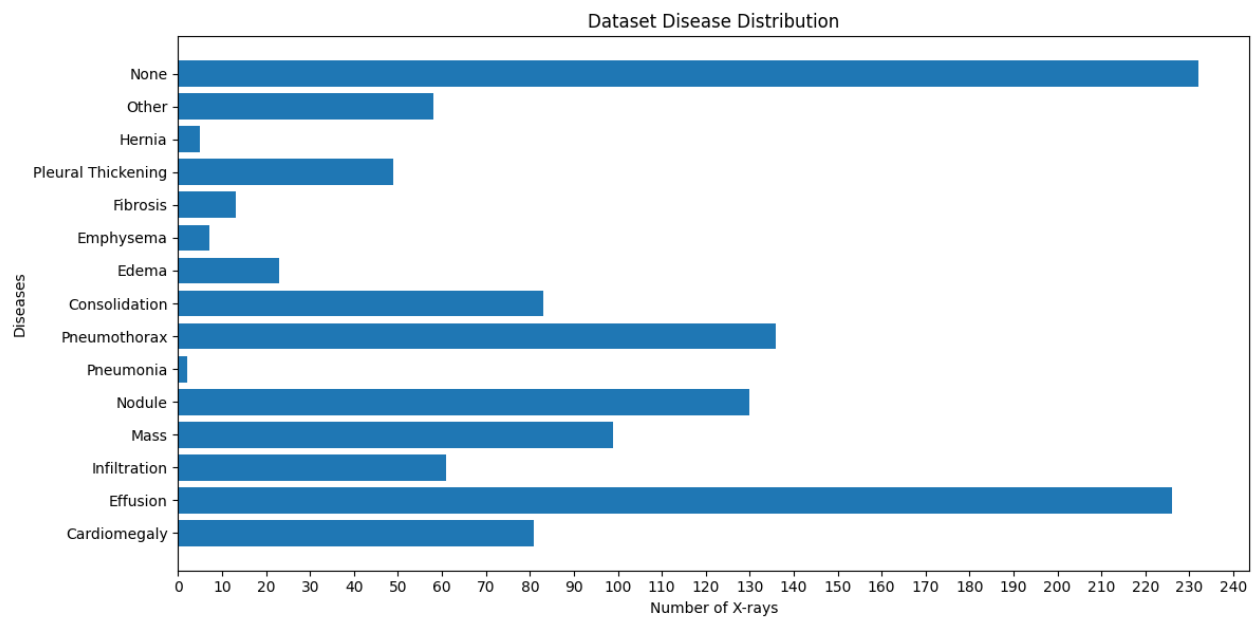


Figure 2. The disease distribution of the dataset with Diseases vs. Number of x-rays. As shown, there are 15 labels, including the "None" label. The class imbalance can also be seen.