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 Convolutional Neural Networks (CNN) in the form of ResNet50 and MobileNet. The efficacy of all are tested.

2. 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. Each image is high-resolution (1024x1024 px) and in PNG format.

Figure 1 shows an example X-ray from the dataset, with the diagnoses of Atelectasis, Effusion, and Pneumothorax. As shown, each individual x-ray can have multiple labels associated with it, which further complicates the model.

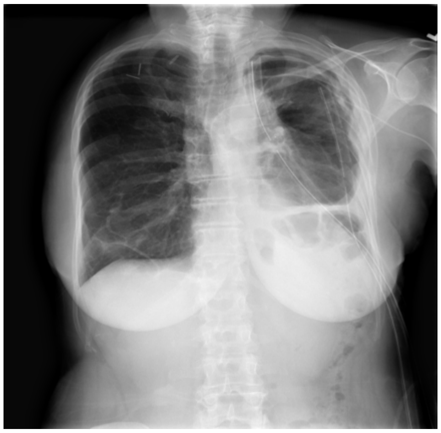

Figure 1. 00018366_048.png, an example X-ray from the dataset with the diseases Atelectasis, Effusion, and Pneumothorax.

The data can be classified into 15 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 15 possible abnormal labels are:

- Hernia,
- Pleural Thickening,
- Fibrosis,
- Emphysema,
- Edema,
- Consolidation,
- Pneumothorax,
- Pneumonia,
- Nodule,
- Mass,
- Infiltration,

- Effusion,
- Cardiomegaly,
- Atelectasis,
- and Other.

In Figure 2, the dataset's distribution of normal and abnormal x-rays is shown, where abnormal is any image with at least one of the 15 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.

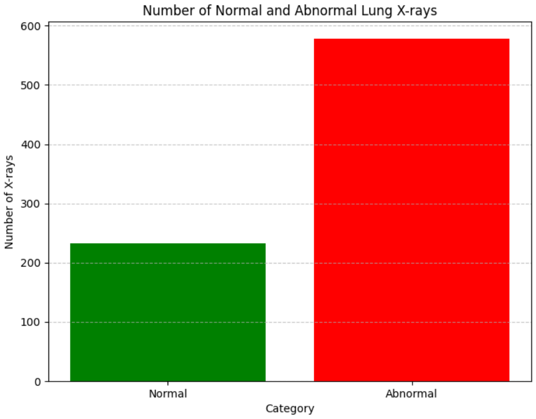

Figure 2. The distribution of normal vs. abnormal lung x-rays

As shown in Figure 3, there is a very clear class imbalance, with ”None” and ”Atelectasis” occuring the most and ”Pneumonia” and ”Hernia” occuring the least.

3. Methods

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

Three separate models were used in order to accomplish this project. The first was a Support Vector Machine (SVM) model whereas the other two relied on Convolutional Neural Networks (CNN) in the form of ResNet50 and MobileNet. The data was split into a 60% training, 20% testing, and 20% validation separation.

3.1. Support Vector Machine

Support Vector Machine or SVM is a supervised machine learning model used to separate classes based on their ability to be separated linearly. The class instances are graphed and ideally separated into clusters. SVM models take this data and create a linear separator that creates the maximum margin between clusters. In the case where a linear separator is not successful, the model can employ the kernel trick, which adds higher dimensions in order to create a substantial margin between classes.

3.2. Convolutional Neural Networks

Convolutional Neural Networks or CNN are a type of neural network that are designed to process images. Neural networks are machine learning models that use artificial neurons, or nodes, to process data. Each node takes in the weights of each feature, computes the sum, and runs it through an activation function. The activation function can be any function, such as ReLu or the sigmoid function.

CNNs differ from neural networks by containing a convolutional layer, which takes in input data, a filter, and a feature map.

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			

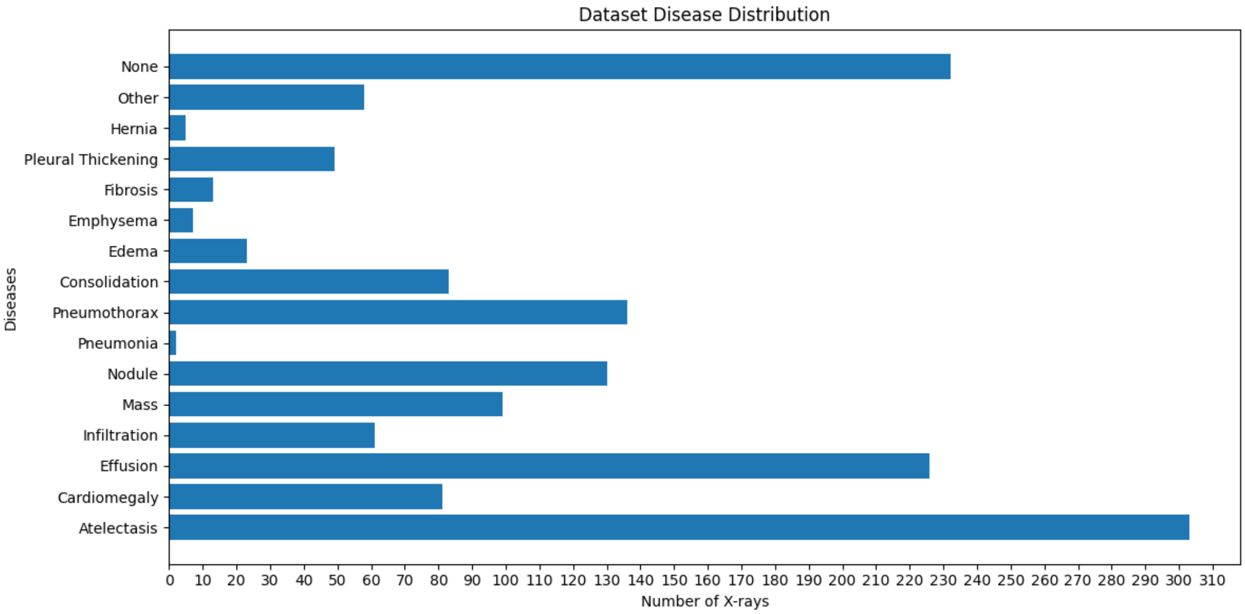

Figure 3. The disease distribution of the dataset with Diseases vs. Number of x-rays. As shown, there are 15 labels, including the ”None” label. The class imbalance can also be seen.