Chest X-Ray Image Classification: Project Report

Zach Wimpee

February 8, 2021

Introduction and Overview

1 Motivation, Problem, and Client

The COVID-19 pandemic has presented seemingly all areas of society with a host of new and unprecedented challenges, many of which are faced in the healthcare field.

For example, while many people living in developed countries are fortunate enough to have relatively easy access to COVID-19 testing, those living in developing or impoverished areas may not be afforded this same luxury. The virus, however, does not discriminate, and therefore additional tools are needed to aid in diagnosing infected patients when available tests are limited or perhaps even nonexistent.

2 Proposed Solution and Approach

Consider the following hypothetical scenario: Dr. Smith runs a clinic for the impoverished in a developing country, and has a new patient exhibiting a variety of symptoms including chest pain, coughing, and fever. Dr. Smith suspects that the patient has pneumonia, but is unsure if it is a viral or a bacterial infection. He would like to rule out the possibility of a COVID-19 infection, but the clinic is still waiting to be resupplied with a shipment of tests. The patient's condition is declining, and the fastest diagnostic tool available is a chest X-ray.

In this hypothetical scenario, Dr. Smith and his patient would benefit from a tool that could determine from an X-ray whether or not the pneumonia is being caused by a COVID-19 infection. The project being proposed here is an exploration into the potential use of neural network machine learning models in developing such a tool.

2.1 Finding a Dataset

A brief search identified a promising dataset to use for this project's development. It is a curated set of chest X-rays for which each image is labeled one of either COVID-19, normal, viral-pneumonia, or bacterial-pneumonia.[1] The dataset has already been checked for duplicates samples and defective images, and therefore minimal cleaning and preprocessing will be required.

[1] https://data.mendeley.com/datasets/9xkhgts2s6/2

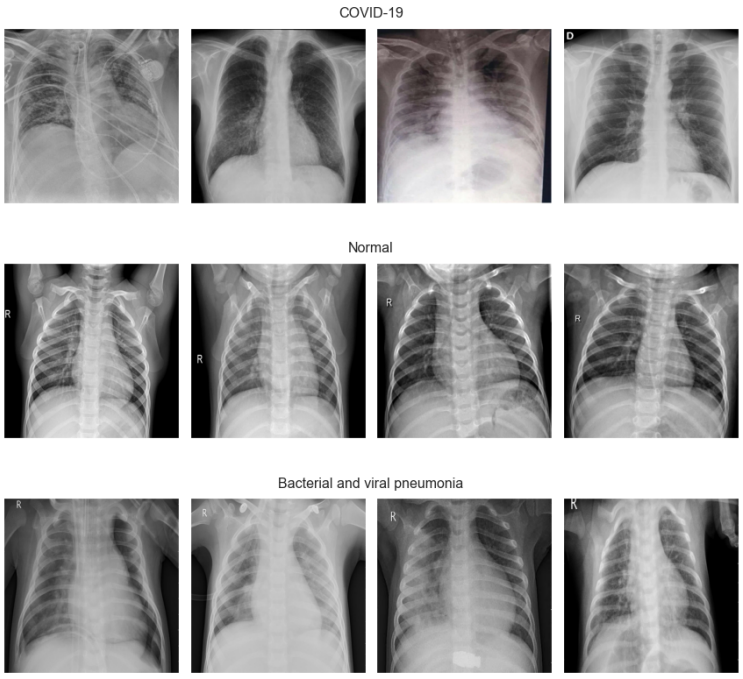


Figure 1: Sample chest X-rays from curated dataset

2.2 Outlining the Deliverables

The final goals for this project are twofold: The first is to design and train a neural network model to classify chest X-rays, and achieve an acceptable level of performance upon test set evaluation. After acquiring a decent model, the final goal will be to implement visualizations that give insight into what the model deems important in its decisions so as to provide model interpretability.

Data Exploration and Preprocessing

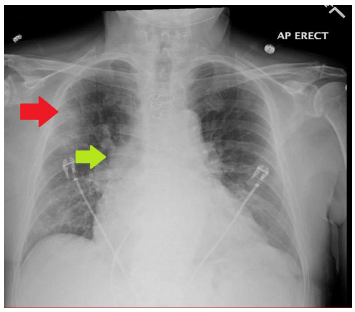
3 Examining the Dataset

The original dataset contains a total of **9208 chest X-ray images...**

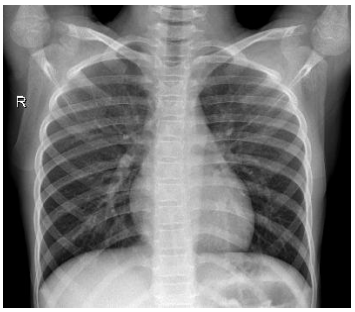
- 1281 COVID-19
- 3270 Normal
- 3001 Bacterial Pneumonia
- 1656 Viral Pneumonia

For the purpose of minimizing complexity the viral and bacterial pneumonia samples will be consolidated under the single class label of pneumonia. This reduces the goal of the project to obtaining a model to make predictions between 3 classes instead of 4. While this simplification is less realistic than keeping the distinction between pneumonia cases, it will facilitate the development of a baseline model with maintained focus on distinguishing between the COVID-19 and pneumonia samples.

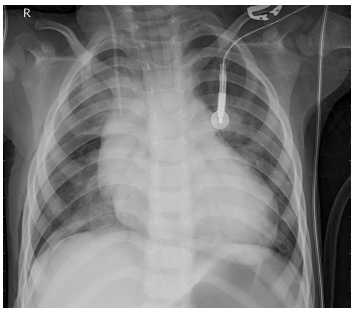
3.1 Issues and Concerns



(a) COVID-19



(b) Normal



(c) Pneumonia

Figure 2: Sample X-ray for each class label

A brief examination of the data reveals several causes for concern that should be noted. **Many of the COVID-19 and pneumonia samples have annotations and text highlighting points of interest, as well as what appears to be pieces of equipment that are not seen in the normal X-rays.** These artifacts pose the threat of artificially inflating the performance metrics for a model that has been trained on these images. This concern will be addressed in two places:

- First during data loading procedures...
  - Data augmentation is applied to training set
- Then during model testing and evaluation...
  - Assess attribution of input to model predictions using Captum

4 Splitting Data and Addressing Class Imbalance

The data is now split into disjoint training, validation, and testing sets. 60% will be used to construct the training set, while the validation and testing sets will evenly split the remaining 40% of the data. Although the full dataset has a high level of class imbalance, each class is represented by a relatively large number of samples. Therefore the issue of class imbalance is handled as follows:

1. Get train/validation/test sets such that the distribution of class labels is approximately the same for each split, as shown below.

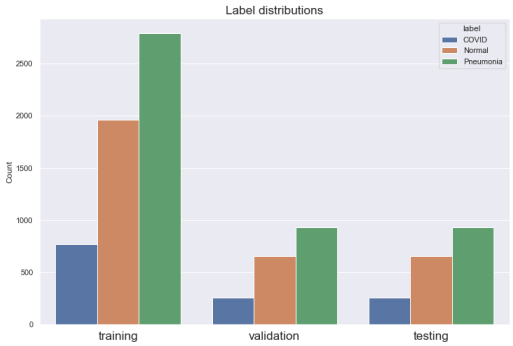


Figure 3: Class label distributions after splitting

2. Compute the accuracy during training steps by taking the class frequencies into account

   - Consider a single training epoch in which label predictions are made for N samples. Then the raw epoch accuracy is given by

   $$\textit{Raw Epoch Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i y_i}$$

   Where,

   - $x_i \equiv$ Predicted label for sample i
   - $y_i \equiv$ True label for sample i
   - $\delta_{x_i y_i} = \begin{cases} 0 & x_i \neq y_i \\ 1 & x_i = y_i \end{cases}$

   - Here we will modify the calculation of epoch accuracy by weighting each prediction by the inverse frequency of the true class labels.

   $$\textit{Balanced Epoch Accuracy} = \frac{1}{C}\sum_{j=1}^{C}\sum_{i=1}^{N_j}\frac{\delta_{x_{ji} y_{ji}}}{N_j}$$

   Where,

   - $C \equiv$ Number of classes
   - $N_j \equiv$ Sample counts for class label j, where $j = 1, ..., C$
   - $x_{ji}, y_{ji} \equiv$ Predicted and true labels, respectively, for sample index ji
   - $\delta_{x_{ji} y_{ji}} = \begin{cases} 0 & x_{ji} \neq y_{ji} \\ 1 & x_{ji} = y_{ji} \end{cases}$

   Note that this is equivalent to taking the average of the recalls for each class.

3. Use metrics that can encapsulate the imbalance into the scores during model evaluation on test data

   - In addition to both raw and balanced accuracy scores, test data predictions will be evaluated using a variety of performance metrics and techniques:
     - Confusion Matrix
     - Precision-Recall Curves
     - ROC Curves
     - Precision, Recall, and F1-Scores