Thomas Stern

31 July 2022

Capstone 3 Proposal: CXR Pathology Prediction

Chest X-rays are a quick and inexpensive diagnostic tool used in emergency rooms and primary care all over the world. They have proven themselves to be relatively reliable but misdiagnosis is still a large problem compared to other radiographic techniques of this region (ie chest CT or Ultrasound). One meta-analysis from the Society of Critical Care Medicine found that chest X-rays have an overall sensitivity of 48%, whereas Ultrasound sensitivity was 95% and both had similarly high specificities of 92-94%. It seems that either the images are not showing pathology or radiologists are not able to effectively diagnose based on the images. The American Journal of Emergency Medicine puts the sensitivity of chest X-rays for pneumonia specifically to be between 38% and 76%. It is my hope that machine learning tools, such as convolutional neural networks, may be able to diagnose based on image data better than human experts.

In this project I aim to make a model that performs at least as well as human experts using data from Guangzhou Women and Children's Medical Center. There are 5863 chest X-ray images of pediatric patients between 1 and 5 years old. Each image is labeled as Normal or Pneumonia. Images labeled Pneumonia were further classified into Bacterial/Viral categories. These X-rays were taken during routine visits and analyzed by 2 professional radiologists.

The scope of this project is far reaching as chest X-rays are a first-line diagnostic tool all over the world. They are quick, cheap and very useful to doctors everywhere. Unfortunately, they

may be relied on too heavily. With enough data we may be able to change hospital convention so that it is not the first and only tool used for respiratory symptoms. Although CTs are the gold standard, they are more expensive and invasive, there is also the chest Ultrasound which could be a replacement for chest X-rays as the first line. This would benefit all those involved in the healthcare experience by making cheaper, more accurate diagnoses.

There are a number of constraints that present a challenge to this mission. Obtaining more data is always a difficulty when dealing with medical data. This dataset is relatively large for healthcare data but it may be difficult to make highly accurate models on so few instances of data. Another limitation is the labels for each image because they are based on human judgment. If we are basing all the modeling around imperfect labeling by human experts, then we are unlikely to ever get higher accuracy levels. It is also limited in that it is only pediatric patients, age 1-5.

I plan to use a CNN model to most accurately predict whether pneumonia is present in any given chest X-ray. This will all be delivered in the form of a github repository folder containing all code and modeling within jupyter notebooks as well as a report document and slide deck for presentation.