Intro

|  |  |
| --- | --- |
| Pneumonia Testing  Machine Learning | Abstract  This is a project that uses images in order to classify whether a patient has pneumonia. We want to know whether we can use a popular machine learning, Convolutional Neural Network, to classify pneumonia.  Special Topics – Fall 2018  AJ Abundez, Jordon Jones, AJ Romanoski, Jack Nordell, Rachael Ostlund |

*As our knowledge of machine learning deepens and the questions in bioinformatics become more complex it is crucial to allow for connection to be utilized*. By using deep learning techniques, we are able to use automatic feature learning to allow the learner to make complex predictions. Machine Learning has been used in many different areas of bioinformatics including: genomics, proteomics, and system biology. It is now being used help improve accuracy and efficiencies when making medical diagnosis.

Pneumonia is the second most common reason for adult hospitalizations and is the leading cause of infant hospitalizations in the United States. With pneumonia being such a prevalent infection within the medical system it is critical to allow for timely and reliable diagnosis. Our goal was to utilize machine learning techniques to efficiently and accurately classify X-Ray images of lungs to solve this problem, this is a possibility because of the advanced features machine learning has available. This would allow for the medical system to make decisions in a timely manner and could be used as a “second” opinion.

The following report will cover the biological background of pneumonia including the causes, prevention and an explanation of the diagnosis process. Furthermore, the process and techniques our team used implement a CNN to classify pneumonia through X-Ray images.

# Biological Background

Pneumonia is an infection that inflames the alveoli, air sacs, in one or both lungs; the air sacs may fill with fluid (purulent material). This causes for patients to have difficulty breathing and allowing enough oxygen to reach the bloodstream. In infected patients this can cause a cough with phlegm, fever, chills, and difficulty breathing. The severity of pneumonia can range from mild to life-threatening; the most serious cases are common in infants and young children, people older than age 65, and people with health problems or weakened immune systems.

## Causes and Prevention

A variety of organisms, including bacteria, viruses and fungi, can cause pneumonia and is commonly transmitted through the air we breathe, meaning that it can be spread to someone else. *Currently there is not reliable research for the SIR model of pneumonia in the US*. Pneumonia is classified according to the types of germs that cause it and where you got the infection.

Pneumonia prevention is best achieved through maintaining positive selfcare. Practicing good hygiene to protect yourself against respiratory infections, washing your hands regularly and upholding a clean environment. Additionally, do not smoke, it damages your lungs as a natural defense against respiratory infections. Lastly, get vaccinated, for the flu to keep your immune system strong, pneumonia vaccination are also available as a course of prevention.

## Pneumonia Testing

There are a variety of different ways that doctors can diagnosis pneumonia in a patient. When showing some of the common symptoms (chest pain, coughing, fever, nausea, shortness of breath, etc.) a doctor can recommend getting a blood test, pulse oximetry, sputum test, or a chest X-Ray. The most common of these is a chest X-ray, allowing the doctor to determine the extent and location of the infection, although, it will not tell your doctor what germ is causing the infection.

Chest x-rays show lung abnormalities through increased or decreased density. When diagnosing pneumonia doctors look for consolidation to be present on the scans. This is a cloudy looking representation that shows a replacement of air in the alveoli. A healthy lung is shown in Figure 1 and a pneumonia infected lung is shown in Figure 2 below.

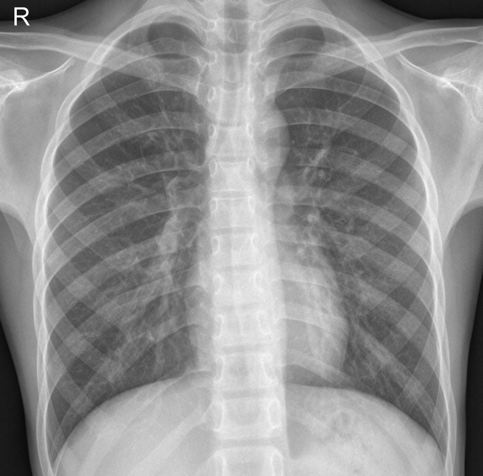
 

Figure 1 Figure 2

## Project Data Set

The data set that our team utilized for this project was images in order to classify whether a patient has pneumonia. The dataset that we are using is open-source hosted by Kaggle. Kaggle is an online community of data scientists and machine learners, owned by Google Inc. allowing users to find and publish data sets, explore and build models in a web-based data-science environment.

The Kaggle data set we used during this project was Chest X-Ray Images (Pneumonia), provided by Paul Mooney. It provided 3 sets of data (train, test, validating) with photos that were categorized as normal or with pneumonia. We found that this dataset was beneficial for our work because we did not have to create or divide a larger dataset up.

# Machine Learning Implementation

For this project we will be testing on a dataset of 234 normal images and 390 pneumonia images, training on a dataset of 1341 normal images and 3875 pneumonia images and validating on a dataset of 8 normal images and 8 pneumonia images. The large dataset for training helped improve our overall accuracy. We implemented supervised learning on a convolutional neural network to answer the question at hand.

## Pre-Processing

The data that we are using from Kaggle provides images that have a variety of different sizes. The first step of our preprocessing was creating images that were all the same dimensions while losing the least amount of usable data. Furthermore, in our processing we used TensorFlow, which can only process images of the same size. In addition, this helped us create a more efficient program by creating less data to work with by removing the dynamic sized layer.

The resizing of images was completed in java. The initial step was to remove excess space around the lungs that contained no useful data. This was done by creating a new image that was decreased by a factor of 4 from the original image. Then redrawing only the center section of the original image to this new image. Effectively removing the unneeded parts and focusing the image on the lungs. We then needed to resize all the images to the same size which is required for a CNN. We read in and averaged all of the image’s heights and widths, this information was then used to create a common image dimension for all images. Once all images were scaled correctly, the next step was to transform the images into a 2D array for TensorFlow to read in.

## CNN Implementation

Our NumPY array will then be put into our CNN where it will run through the convolutional layer, ReLU layer, pooling layer and fully connected layer to create a binary classification. While the program is running initially it starts to recognize lines and shapes within the image to learn the placement of the ribcage. It then pays attention to the different densities throughout the image to recognize the cloudiness within the lungs to classify it as pneumonia.

The ReLU and pooling layer will run fixed functions. In the pooling layer we recognized that are inputs were large we had to adjust with padding as we continued. With a kernel size of **XXXxXXX** we needed to create an initial padding on **XXX** in each direction and increasing by **XXX** after each convulsion.

After running for **XXX** epochs it enters the fully connected layer.

Weights and biases

Cross entropy

Over fitting

Learning rate

SoftMax

# Experiment Findings and Comparisons