**Problem Statement:**

Pneumonia is an infection in one or both lungs. Bacteria, viruses, and fungi cause it. The infection causes inflammation in the air sacs in your lungs, which are called alveoli. In this capstone project, the goal is to build a pneumonia detection system, to locate the position of inflammation in an image. Tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear black in the image. Dense tissues such as bones absorb X-rays and appear white in the image. While we are theoretically detecting “lung opacities”, there are lung opacities that are not pneumonia related. In the data, some of these are labeled “Not Normal No Lung Opacity”. This extra third class indicates that while pneumonia was determined not to be present, there was nonetheless some type of abnormality on the image and oftentimes this finding may mimic the appearance of true pneumonia.

**Data Format and attributes in the data:**

Dicom original images:

Medical images are stored in a special format called DICOM files (\*.dcm). They contain a combination of header metadata as well as underlying raw image arrays for pixel data.

The below screen shot obtained during data exploration gives us the meta data present in files:

A picture containing text

Description automatically generated

Out of the meta data present in the files, the ones that are of most value are:

* Patient Id: The DICOM files are named according to the patient-id
* Sex of the patient under examination
* Age of the patient under examination
* Pixel Data: Actual X-Ray of the patient
* Size of X-ray: 1024x1024 for all patients

**Data Statistics:**

1. The number of DICOM files provided in the training set is 26684
2. The number of DICOM files provided in the testing set is 3000
3. Data split according to the gender is equal in both the training and testing set. The following plots illustrate how the data is distributed according to the genders

**Chart, waterfall chart

Description automatically generated**

|  |  |  |
| --- | --- | --- |
|  | **Training Set** | **Testing Set** |
| **Number of Male Patients** | **15166** | **1714** |
| **Number of Female Patients** | **11518** | **1286** |
| **Total** | **26684** | **3000** |

Percentage of Males data in training set:56.8%

Percentage of Females data in training set:43.2%

Percentage of Males data in testing set:57.1%

Percentage of Females data in testing set:42.9%

From the above numbers we can see that there is a fair distribution of data according to the gender of the patients.

1. When we look at the distribution of data in terms of age of the patients the distribution has peak around 55yrs indicating that the age group near to 55yrs could be more vulnerable to having pneumonia and have been tested more. The distribution is plotted below:

Chart, histogram

Description automatically generated

The above-mentioned point is valid for both the genders. It is seen that both male and female patients are at risk of pneumonia if their age is around 55-60years.Below plot has only data from patients with pneumonia.

Chart, line chart

Description automatically generated

**Labels:**

The csv file provided with the labels contained the bounding box information and another label was provided describing the class. They both were merged into one data frame as shown below:

Table

Description automatically generated

Also, one thing noticed is that though there were 26684 patient-ids in training set, but in the labels it is found that there were 30227 rows indicating either duplicate information. After inspecting we found that the there is no duplicate information but for a particular patient, we could have inflammation detected in both the lungs. Some examples for different cases are shown below:

Graphical user interface

Description automatically generated with medium confidenceGraphical user interface

Description automatically generated

A picture containing graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated

One thing to notice here is that the bounding boxes come in different sizes.

**Data Preparation for Model:**

1. The images have been resized to have a size of 224x224. Since these are gray scale images, we have only one channel.
2. Using the bounding box information we have prepared the segmentation mask(28x28) which look like below (samples provided):

Chart

Description automatically generated

A sample segmentation mask.

**MobileNet Results:**

**The dice coefficient we are getting on the model is 0.6**

On the left we have the segmentation mask generated by the model and on the right we have the actual mask.

* Case 1: Target=1 that is, pneumonia detected

Chart

Description automatically generated

Case 2: Target=2 that is, pneumonia not detected

Chart, histogram

Description automatically generated