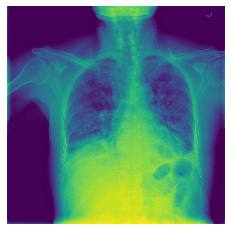
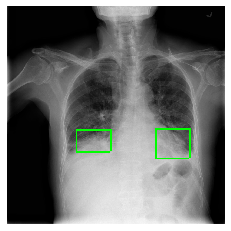
INTRODUCTION

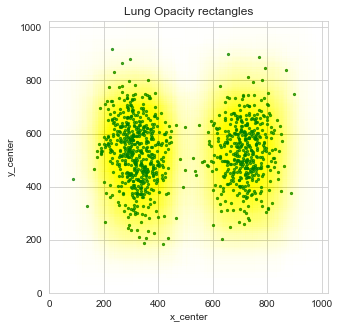
Pneumonia is a waterborne disease that affects the lungs. The air sacs in the lungs, once infected by virus or bacteria, start getting filled with fluids. This causes respiratory problems. Pneumonia is responsible for nearly 15% total deaths worldwide. Pneumonia is also a leading side effect of being affected by COVID-19 or popularly known as the corona virus.

One way of detecting pneumonia is through chest X-rays. These fluid-filled sacs show up as opaque areas in the chest X-rays. These X-rays are generally examined by a trained medical professional. This, however, can be time consuming and also introduces the chance of human error or uncertainty. Also, due to certain unforeseen conditions (like pandemics, widespread diseases, etc.), clinicians might be under a lot of stress to examine X-rays in large numbers. This is where our algorithm comes into play. Our project takes chest X-ray images in the form of DICOM images as input and detects these opaque areas in the chest X-rays to decide whether the patient has pneumonia or not. Radiology Society of North America (RSNA) has hosted this dataset on Kaggle as a part of a competition.

The patients in the dataset have been divided into three categories. Patients with Normal lungs, Opaque lungs and whose results are uncertain. For the purpose of this dataset, the third category is considered to be Normal. Normal lungs have no opaque areas in them. the main result is to find patient X-rays with opacity in them.

 X-ray example X-ray with bounding boxes

It is also important to find out where these opaque regions exist and how widespread they are. The dataset given contains X-ray images in DICOM format along with co-ordinates of bounding boxes. These bounding boxes, when overlaid on the X-ray images, show the exact positions of the presence of opacity. It was observed that the patients can have up to four such bounded boxes int heir X-rays. However, majority patients with pneumonia have between one and two areas. Also, these boxes are pretty evenly distributed throughout the lungs (as shown in the illustration below).

<add data cleaning part and finish>

LITERATURE SURVEY/ RELATED WORK

<https://medium.com/@sebastiannorena/kaggle-rsna-pneumonia-detection-challenge-explained-c140b19bf903>

<https://arxiv.org/pdf/1811.08939.pdf>

Radiology Society of North America hosted this dataset on Kaggle as a means to get more efficient methods to detect pneumonia using patient X-rays. The data is already classified. Therefore, supervised machine learning techniques have been used to classify the data and come up with different methods to do so. Since the dataset was posted as a competition on Kaggle, over 1400 teams participated. Many classification methods were used, like boosting, k-nearest neighbors, decision trees, kernel methods, etc.

Our project made use of CNN and SVM algorithms. CNN works good with computer vision-based algorithms and, hence, is very useful for image data. SVM works best with numerical data. We extract metadata from the images in the form of numerical data and use that as input to our SVM. These help in localizing anomalies.

In other projects, more advanced methods like Mask-RCNN and FCN have been used. However, these work on a pixel level. Each pixel is evaluated and classified in a binary format. Based on that, a binary matrix ix given as output. Therefore, when objects to be identified (like in our case, opaque areas in the lungs), are very clos together, or have overlap between them, the algorithms do not work as efficiently in identifying separate opaque areas.

Another algorithm used is CoupleNet(<https://arxiv.org/pdf/1811.08939.pdf>). A CoupleNet is a type of fully convolutional network. It is used to couple the global structure of the data with the local parts for better object detection. This is what helps in giving CoupleNet algorithm a high prediction accuracy. However, due to such complexity, it is difficult to train and takes a lot of processing power.