horizontal line**IPHIGENIA**

CHINMAYEE DS

CHANDANA V KUMAR

Covid-19 Diagnosis using X-RAYS

**18th November 2020**

# OVERVIEW

The outbreak of the novel coronavirus has shaken up the world and as a developing nation, with one of the largest populations in the world, we clearly have proven to be under equipped. Although this pandemic has been brutal to all sectors of economy, the health care sector has borne the brunt of it; overworked doctors, shortage of nurses, paramedics, radiologists, shortage of medical equipment, medicines and testing kits.

The current testing facility takes anywhere from a few hours to several days depending on the type of testing. This will only contribute to unmonitored spread of the infection and increase in fatality rate. As this virus affects lungs, chest X-rays are often referred during the course of treatment. Attempts are on by various researchers to apply image analytics in identifying the COVID-19 infection with the X-rays available. This methodology can be tried out in a vast country of our size and can be used as a method of primary diagnosis to begin with.

With this motivation, we aim to develop an algorithm which will analyse the x-ray scans and make the diagnosis. This model will specifically help in identifying if a patient has been infected by COVID-19 or Pneumonia, both of which exhibit very similar symptoms.

Development of this methodology will significantly change the detection methodology from the current scenario. As infra for the above methodology is well established, testing time and cost can be reduced. Due to the low RT-PCR ( Reverse transcription polymerase chain reaction ) sensitivity of 60%–70%, even if negative results are obtained, symptoms can be detected by examining radiological images of patients.

# GOALS

1. To Develop the most accurate diagnostic deep learning algorithm.
2. To save time and resources spent in diagnosis using RT-PCR.

# SPECIFICATIONS

Since we wanted to build the most accurate diagnostic algorithm, we needed reliable, true data. We used the dataset from Kaggle’s “COVID-19 RADIOGRAPHY DATABASE (Winner of the COVID-19 Dataset Award by Kaggle Community)”.

**DATA DESCRIPTION**

This radiology database contained 215 Covid-19 chest X-RAYs, 1342 chest X-RAYs of Viral Pneumonia and 1347 Normal, healthy chest X-RAYs. From these numbers, it's evident that the database is extremely imbalanced and fitting the models now is not an option. We then studied our data by performing some Exploratory Data Analysis (EDA) to get a deeper understanding of the parameters and/or hyper parameters we have to consider.

**EDA**

We started out by printing a few random images from the dataset to see if there is any visible difference. Upon learning about the virus itself, there are a few conclusions we arrived at:

1. The visual changes in the lungs of patients infected by Covid-19 can be caught in later stages of the disease.
2. These changes are in the form of Ground-glass opacification/opacity (GGO) which is a radiological term used to refer to haziness in the X-RAY of lungs.
3. This type of haziness is also very common in lungs of patients suffering from Viral Pneumonia.

To dig deeper into this, we did the following:

1. Printed the average image of all Covid-19 images in the dataset, and the same for all Viral Pneumonia as well as normal lung X-RAYs.
2. Printed the contrast image of all three sections of the dataset.
3. Printed the Standard Deviation of images in the same fashion.

**BALANCING THE DATASET**

As we mentioned earlier, the dataset was imbalanced and we could not proceed further with this range of imbalance as the result will be biased towards the class with more data. The difference of the length of the classes was too large to ignore. So, we implemented a few data augmentation techniques to generate artificial images of Covid-19 X-RAY. These techniques were applied after the dataset was split into test and train and the training data was augmented and not the test data.

**DATA AUGMENTATION**

We implemented 4 techniques for data augmentation:

1. Rotation - This is done by finding the center of an image and rotating it at a random angle between -25° to +25°.
2. Flipping - This is done by flipping the image in the left to right direction.
3. Translation - Done by displacing the image by 50 pixels in random directions (left, right, up, down) and filling the empty space with black pixels.
4. Gaussian Blur - Done by blurring the image by using a gaussian function.

These techniques were randomly applied to the training dataset. This produced artificial images which we then added to our original training dataset to balance it.

**MODEL BUILDING**

Using our now balanced dataset, we fit a few linear and non-linear models for comparison. We chose models which are known to perform well for image datasets and/or especially X-RAYs. These models with their accuracy are:

[0- Covid-19; 1- Normal ; 2- Viral Pneumonia]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | F1 score | Recall |
| Random Forest Classifier | 90% | 0- 94%  1- 88%  2- 91% | 0- 80%  1- 91%  2- 90% | 0- 70%  1- 94%  2- 89% |
| Support Vector Classifier | 93% | 0- 87%  1-95%  2- 92% | 0- 89%  1- 94%  2- 93% | 0- 91%  1- 93%  2- 94% |
| K Neighbours Classifier | 87% | 0-94%  1-89%  2-85% | 0-86%  1-88%  2-87% | 0-79%  1-87%  2-89% |

From the above comparison, it's evident that the F1 scores and recall is low. So, we built a Convolutional Neural Network (CNN) to see how different the results would be.

**CNN MODEL**

The model was built with alternating layers of Convolutional 2D and MaxPooling 2D layers. Eight such alternating layers(Conv2D-MaxPool2D) were added with a dropout layer in the middle to avoid overfitting. The activation function used was “relu”. Three dense layers were added.

In order to use haziness in the lungs as a parameter, we had to choose one of these pixel scaling methods to train our data.

1. Normalization.
2. Centering.
3. Standardization.

After training our model with each of these methods one by one and comparing the results, we concluded that Centering did not yield as accurate results as the other two methods, and Normalization yielded the best results.

[0- Covid-19; 1- Normal ; 2- Viral Pneumonia]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | F1 score | Recall |
| CNN | 92% | 0- 97%  1- 88%  2- 97% | 0- 94%  1- 93%  2- 92% | 0- 91%  1- 98%  2- 87% |

**Web Application**

We built a basic web application using Flask to implement our model. This app takes an X RAY image as input and predicts if the X-RAY belongs to a patient who is normal or infected by Covid-19 or Viral Pneumonia.

**Platforms/Technology Stack used for this project:**

* Google Colab
* Flask
* HTML
* CSS
* Python
* TensorFlow

**CONCLUSION**

Even though the accuracy of our model is 92% which is less than that of SVC, the F1 score and Precision is better. High precision relates to low false positives and F1 score is high only when both precision as well as recall is high, which indicates less false positives and false negatives.

# 