Project Milestone 2

Kiran Chowdary Komati

College of Science and Technology, Bellevue University

DSC680-T301: Applied Data Science (2233-1)

Dr. Catherine Williams

January 26, 2022

Proposal:

Topic: "Covid-19 Image Classification" aims to analyze the x-rays of patients to determine if they have covid or not.

Business Problem:

We all are aware that covid-19 has taken over the world from the year 2020. Most people who fall sick recover under 2 weeks with minimal symptoms and without any special treatment. But there were cases where they needed medical treatment due to severe symptoms. The diagnosis is done primarily by nasal / throat swab tests. This project explores the alternative way by analyzing the x-rays of the patients diagnosed with Covid-19.

Background/History:

The Covid-19 tests are primarily taken using swab test for nose/throat. This can be painful if this must be done repeatedly, especially, in these times, where the covid can attack the same person multiple times. The aim of this project is to explore the possible usage of x-rays to diagnose covid and reduce the need for swabs.

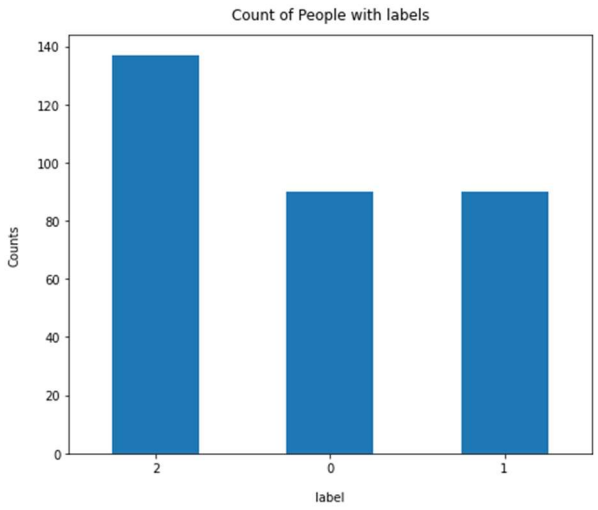
Data Explanation:

Data sets:

I'm using the dataset https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset I found in Kaggle. This has a total of 454 images organized in to 3 folders representing 3 classes out of which 137 images are of Covid-19 and 317 in total with the viral pneumonia and the normal chest X-rays grouped into training and test folders.

Methods:

Exploratory Data Analysis was performed on the data set using Python. Initial checks were done see how many images were present in each of the categories. We checked for the duplicates as well to make sure that we are not dealing with duplicate data. Later we plotted different graphs to better understand the distribution and the data.



Analysis:

The data is split into train and test data sets in the ration 80:20. Sample Train images were plotted to check the classification. Later a VGG16 model was built and fitted against the training data and the testing data was used for validation. VGG16 is object detection and classification algorithm which can classify 1000 images of 1000 different categories with 92.7% accuracy. It is one of the popular algorithms for image classification and is easy to use with transfer learning. The initial accuracy came at 90%. We later performed image augmentation on the training images and then ran our model again. This time the accuracy came around 92% which is better than the initial result.

Conclusion:

After performing image augmentation on the training data set, the accuracy came at 92% which means that the model has a better performance.

Limitations:

This model needs to be tested on a larger data set to make sure that the accuracy of the model is not affected. One challenge to this is that it'll take a good amount of time to train and the test the model. One of the crucial downsides of the VGG16 network is that it is a huge network, which means that it takes more time to train its parameters. Because of its depth and number of fully connected layers, the VGG16 model is more than 533MB. This makes implementing a VGG network a time-consuming task.

Challenges:

In the real world, it could be a little difficult to gather the data to related to the patients due to the PII restrictions although the data I found from Kaggle was publicly available. Also, it could be a little challenge to gather data from multiple sources if it needs different types of tests to be performed on the patient.

Future uses/Additional Applications:

A thorough Analysis of the models using vast amount of real data will help in better prediction of Covid-19 and also faster than the conventional way of testing through nasal/throat swabs. Also, this model can be used to predict other similar diseases such as viral pneumonia that affects the lungs.

Recommendations:

Based on the available information and the analysis performed on the limited data available, this model has a better accuracy of predicting if a person has Covid-19. But as this involves health of person, much more analysis needs to be performed when the larger data sets are available. Also, other ways need to be explored to reduce the time taken to train the model.

Implementation Plan:

Once we can find the large data sets, this model needs to be trained and tested to make sure that the accuracy is not affected. Backup options or models needs to be developed as well to make sure that we have a working model that is better than a random guess incase the original model doesn't workout with the larger data sets.

Ethical Assessment:

As this project is related to the health of individuals and as it contains sensitive information, I made sure to not use any PII information such as individual names. Also, the data needs to be presented in accurate form without any modifications or misrepresentations. Also, steps must be taken to avoid any bias towards any gender. The users of this model also need to be aware of the limitations of the model and not use it for self-medications and should seek doctor incase of any discomfort even though their result may suggest otherwise using the model.

References:

Raikote, pranav (2020). *Covid-19 Image Dataset*

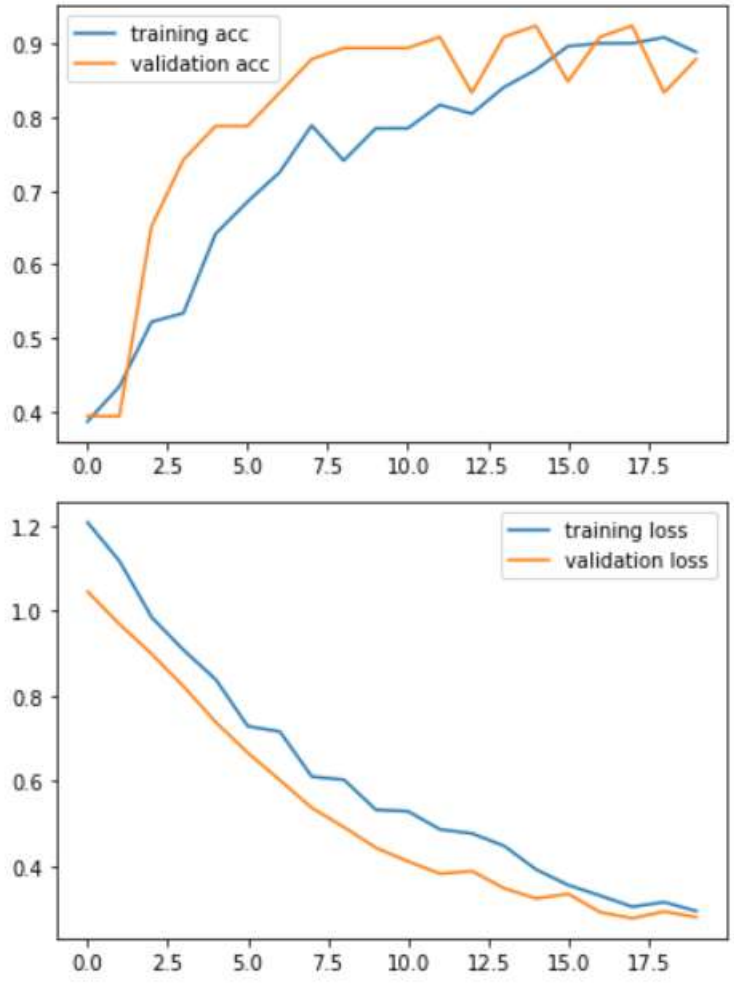
Retrieved from https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset

Rohini G (Sept. 23, 2021). *Everything you need to know about **VGG16***

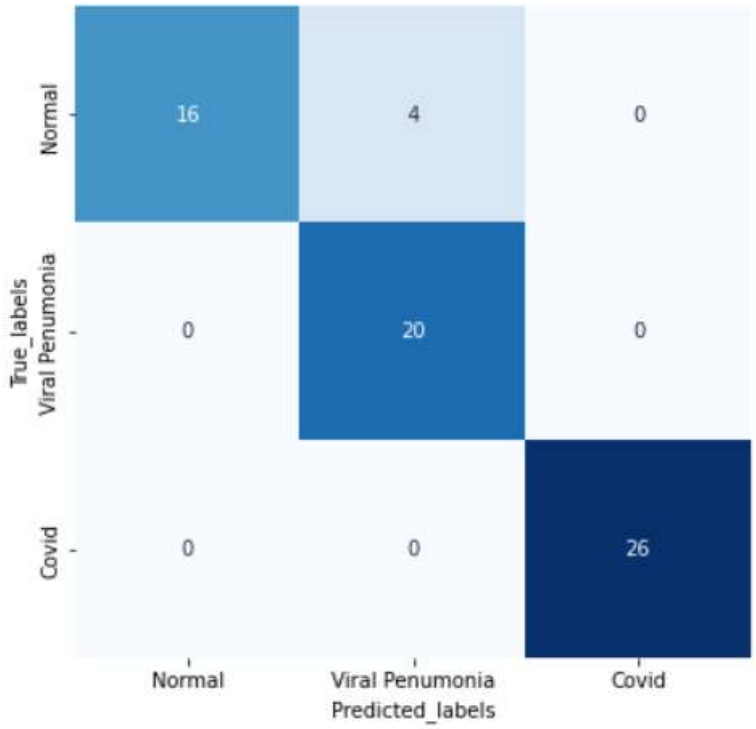
Retrieved from https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918

Appendix:

Learning Curves:



Confusion matrix:



Questions:

1.Why do we need to predict the Covid-19 disease?

2. How does this project helps in predicting the Covid-19 disease efficiently?

3. What is the accuracy rate compared to the Nasal swab tests?

3. Why is the image augmentation done? Will it have any affect on the original outcome?

4. What parameters are used in the model?

5. What are the different models that are built as part of this project?

6.Are there any new features built from the existing features?

7. Can adding any other features affect the model accuracy?

8. is there any additional data to supplement the dataset used?

9. How's the performance of the model?

10. Can this methodology be used in other areas of medical sciences?