Pattern Recognition and Machine Learning Project: COVID Detection using X-ray Images

Hitesh Shanmukha(B22AI013)

Abstract

COVID- 19 global pandemic affects health care and lifestyle worldwide, and its early detection is critical to control cases' spreading and mortality. The actual leader diagnosis test is the Reverse transcription Polymerase chain reaction (RT-PCR), result times and cost of these tests are high, so other fast and accessible diagnostic tools are needed. This projects' approach uses existing deep learning models (VGG19 and U-Net) and various machine learning models to process these images and classify them as positive or negative for COVID-19. The proposed system involves a preprocessing stage with lung segmentation, removing the surroundings which does not offer relevant information for the task and may produce biased results; after this initial stage comes the classification model trained under the transfer learning scheme; and finally, results analysis. The best models achieved a detection accuracy of COVID-19 around 96%.

1 Introduction

The Coronavirus Disease 2019 (COVID-19) has brought a worldwide threat to the living society. One of the areas where machine learning can help is detecting the COVID-19 cases using chest X-ray images. The task is a simple classification problem where given an input chest X-ray image, the machine learning-based model must detect whether the subject of study has been infected or not. In this project, we've analyzed the given chest x-ray image dataset using essential exploratory data analysis techniques and drawn predictions about whether the subject of study has been infected or not.

2 About the dataset

COVID Detection using X-ray Images dataset

- The COVID-19 dataset consists of Non-COVID and COVID cases of X-ray images.
- The associated dataset is augmented with different augmentation techniques to have about 17099 X-ray images.

- The dataset contains two main folders, one for the X-ray images, which includes two separate sub-folders of 5500 Non-COVID images and 4044 COVID images.

3 Importing the Dataset

The dataset was downloaded and then uploaded to google drive from where it was imported using *tensorflow.keras.utils.image dataset from _directory()* with batch size = 32, image width and height = 224, color mode as grayscale and using 80% for training and rest 20% for validation.

4 Data Preprocessing and Analysis

As the images come from several datasets with different image sizes and acquisition conditions, a preprocessing step is applied to reduce or remove effects on the performance of the models due to data variability.

- Preprocessings like normalisation of the images, conversion to gray scale, resizing to standard 224x224 size was handled during the import by keras library.

Figure 1: Images after preprocessing

- Performed dimensionality reduction using PCA, explained in brief in section 5.

- Performed Lung Segmentation using U Net architecture, explained in brief in section 6.

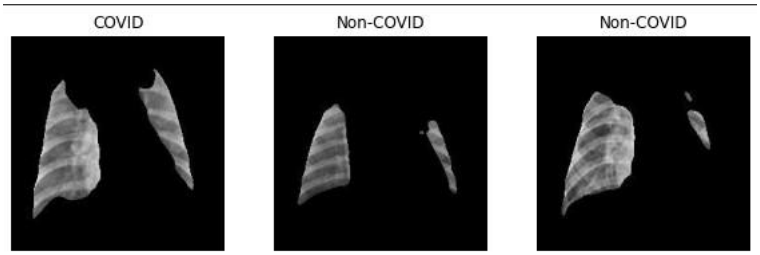


Figure 2: Masked Images

5 Dimensionality reduction using PCA

PCA was done to :

- Reduce the time and storage space required. • Remove multi-collinearity which improves the interpretation of the parameters of the machine learning model.

For deciding the number of Principle components, a scree plot was plotted and a threshold variance of 95% was set and image dimensions were reduced as shown.

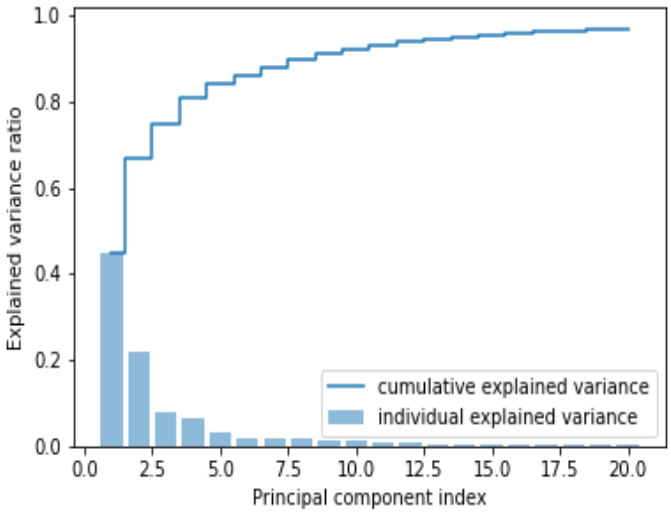


Figure 3: Scree Plot

20 n components were used as they covered 95% of the cumulative variance.

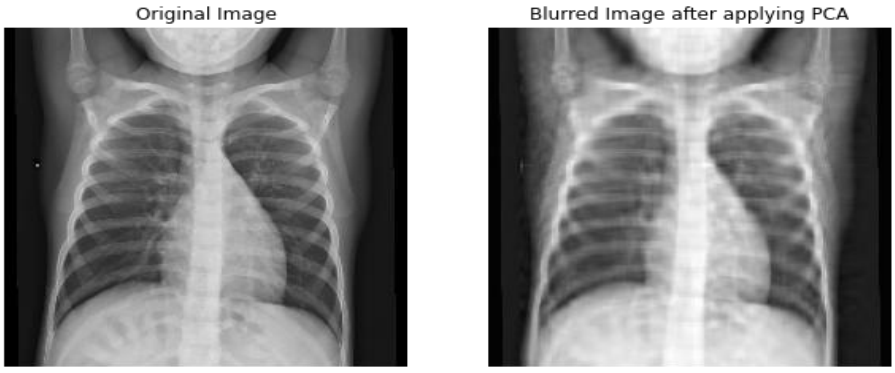


Figure 4: Dimensionality reduction using PCA

6 Lung Segmentation

Segmentation was done using a Deep Learning model based on U-Net architecture. U-Net architecture is accurate for the segmentation of medical images. This kind of model receives the target in the form of an image mask with ones (1) on the reconstruction area and zeros (0) on the rest; consequently, in a production setting, model input is an X-ray chest image, and the output is the predicted mask.

Figure 5: Lung Segmentation

6.1 U Net Architecture

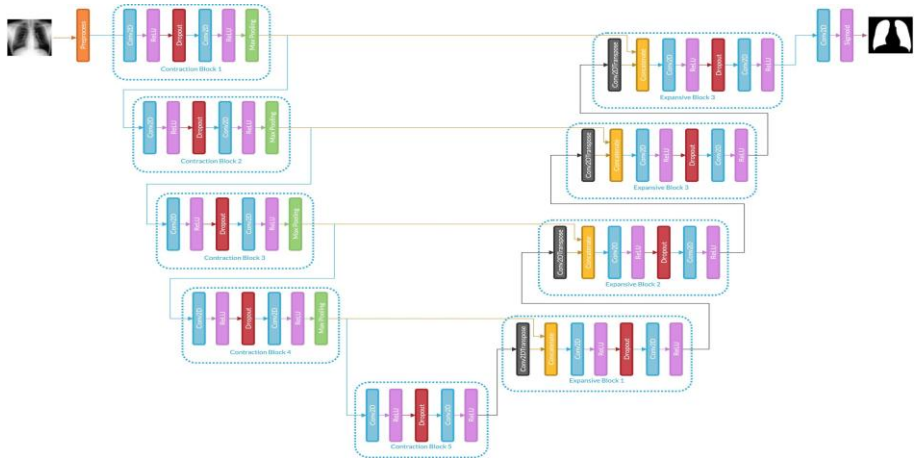


Figure 6: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

7 Application of machine learning models

7.1 Creation of dataset

Dimension of each image were reduced to 224x20 using PCA, which then was flattened and stored in a numpy array.

7.2 Train Test Split

Using train test split function from sklearn the given dataset was split into train and test dataset in 70:30 ratio.

7.3 Results on Unmasked dataset

Models report			
Classification Model	Accuracy	F1 Score	AUC
Random Forest	0.79	0.74	0.79

Decision Tree	0.73	0.69	0.72
Logistic Regression	0.68	0.61	0.67
XGBoost	0.81	0.77	0.81
Light GBM	0.82	0.78	0.82
SVM	0.76	0.69	0.75
K Means	0.45	0.43	0.46

The plot for comparing the accuracies of different models is shown below

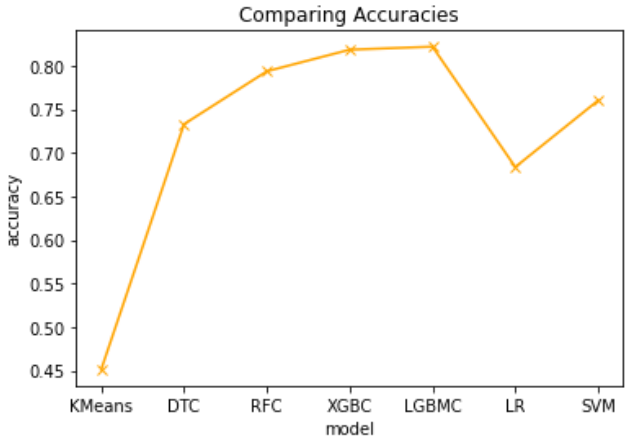


Figure 7: Comparing accuracy

7.4 Results on Masked dataset

Models report			
Classification Model	Accuracy	F1 Score	AUC
Random Forest	0.75	0.69	0.75
Decision Tree	0.69	0.62	0.68

Logistic Regression	0.62	0.51	0.60
XGBoost	0.75	0.70	0.75
Light GBM	0.75	0.70	0.75
SVM	0.71	0.63	0.70
K Means	0.62	0.60	0.63

The plot for comparing the accuracies of different models is shown below

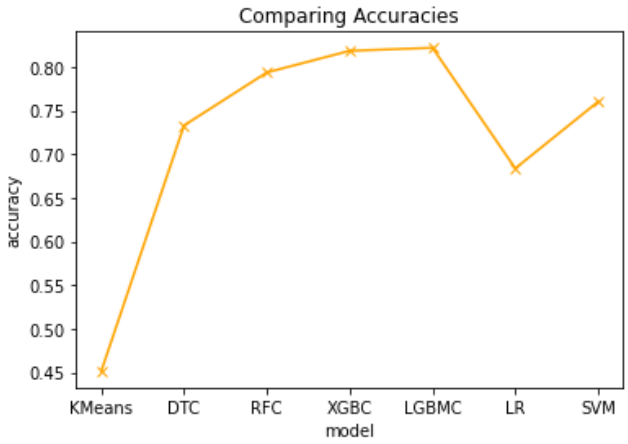


Figure 8: Comparing accuracy

7.5 Comparative Analysis

The plot for comparing the accuracies of models trained on masked and unmasked dataset is shown below

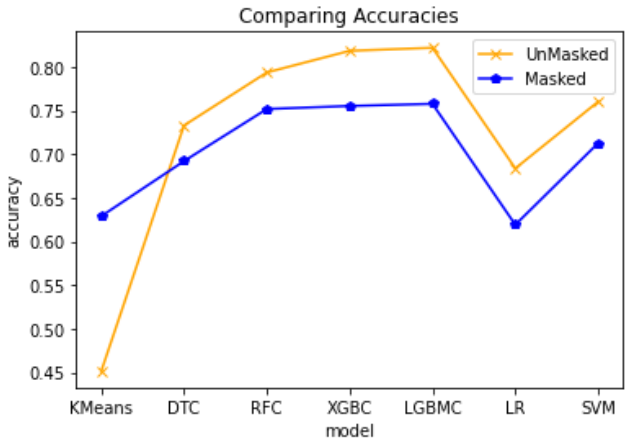


Figure 9: Accuracy comparison

As it is seen that the accuracies using both datasets is quite similar though slightly lower in masked images because creation of masked images may sometimes lead to blackening of some portions of the lungs which hinders the information required for the prediction of covid-19.

8 Deep Learning Models

8.1 Creation of dataset

The imported dataset was then converted into tensors of batch size 32 each and they were then used for training and validating the models.

8.2 VGG-19

Very Deep Convolutional Networks designed to carry out Large-Scale Image Recognition baseline model is used in our multi-class classification as it attains a significant accuracy on image classification and localization tasks. Due to its inherent strength in processing X-Ray image recognition, we used it for our experiments. We implemented the Transfer Learning technique on the VGG19 (baseline) and customized the VGG-19 model during the training on X-Ray datasets. Pre-trained weights of the ImageNet dataset were used. We used discriminative learning rates to preserve the lower-level features and regulate the higher-level features for optimum results

VGG19 is a variant of VGG model which in short consists of 19 layers (16 convolution layers, 3 Fully connected layer, 5 MaxPool layers and 1 SoftMax layer).

To reduce the time and space complexity we removed some of the convolution layers and used the 3 fully connected layer, 5 MaxPool layers and 1 SoftMax layer of 2 Neurons as it is binary classification problem.
The model was compiled using ”adam” optimiser and ”categorical cross entropy” as loss function.

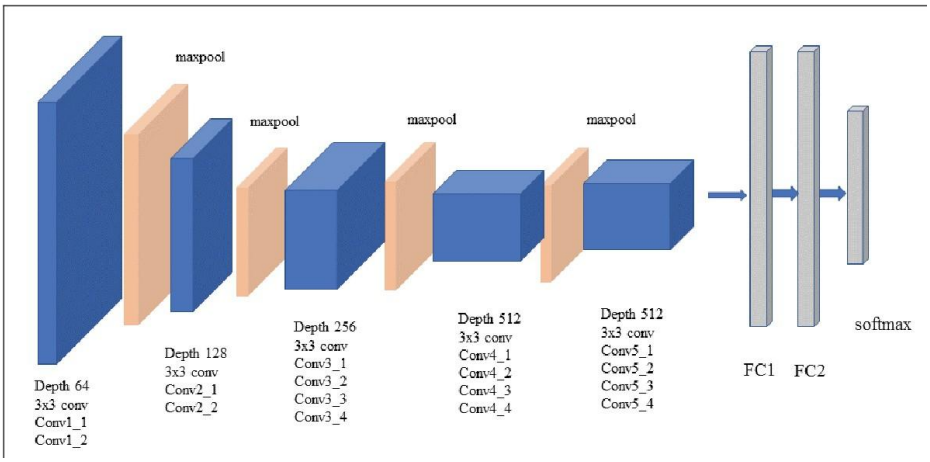
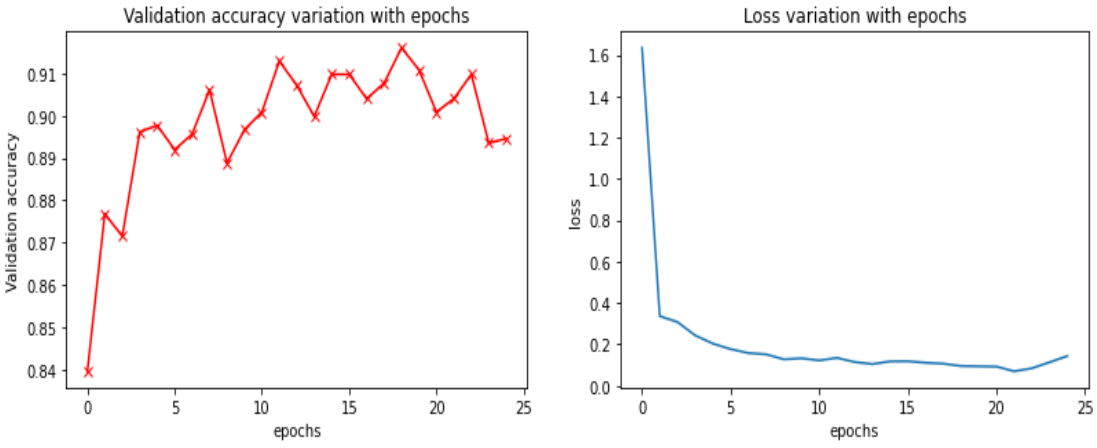


Figure 10: The figure illustrates VGG-19 Architecture tailored to detect COVID-19 in Chest X-Ray Imaging

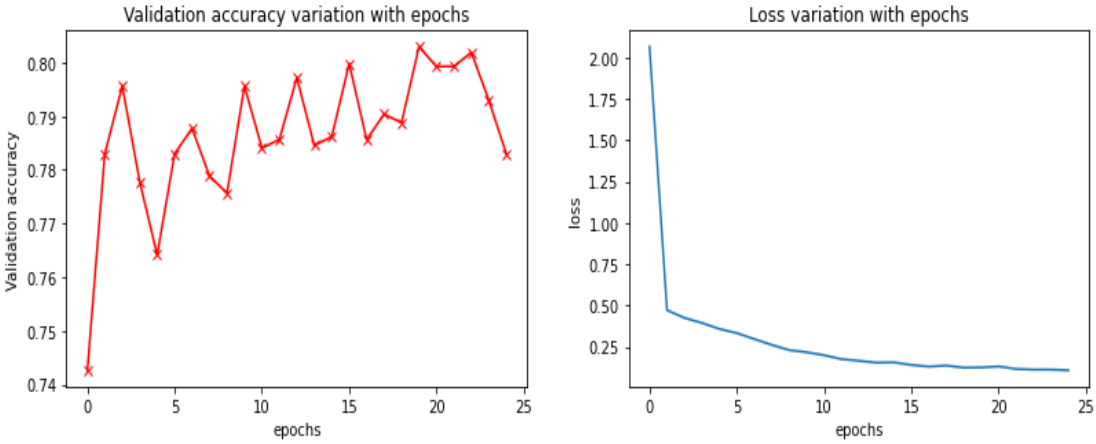
8.2.1 Applying on unmasked dataset

Validation accuracy and loss obtained on running 25 epochs is shown below



8.2.2 Applying on masked dataset

Validation accuracy and loss obtained on running 25 epochs is shown below



8.2.3 Comparative Analysis

As it is seen that the accuracies using both datasets is quite similar though slightly lower in masked images because creation of masked images may sometimes lead to blackening of some portions of the lungs which hinders the information required for the prediction of covid-19.

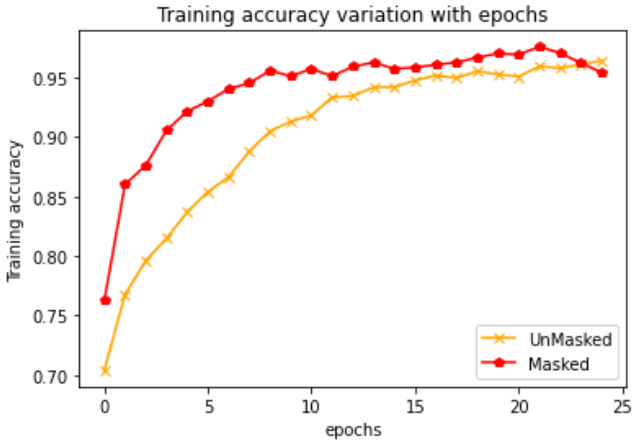


Figure 11: Comparison of accuracy

8.3 ResNet 50

ResNet is known to be a better deep learning architecture as it is relatively easy to be optimized and can attain higher accuracy. Furthermore, there is always a problem of vanishing gradient, which is resolved using the skip connections in the network. As the number of layers in the deep network architecture increases, the time complexity of the network increases. This complexity can be reduced by utilizing a bottleneck design. As a consequence, ResNet50 is a preferred pretrained model using weights from imagenet to build up our framework.

The model is compiled using stochastic gradient descent optimizer and sparse categorical cross entropy as loss function

Experiments on masked and unmasked dataset and its comparative analysis are shown in colab file.

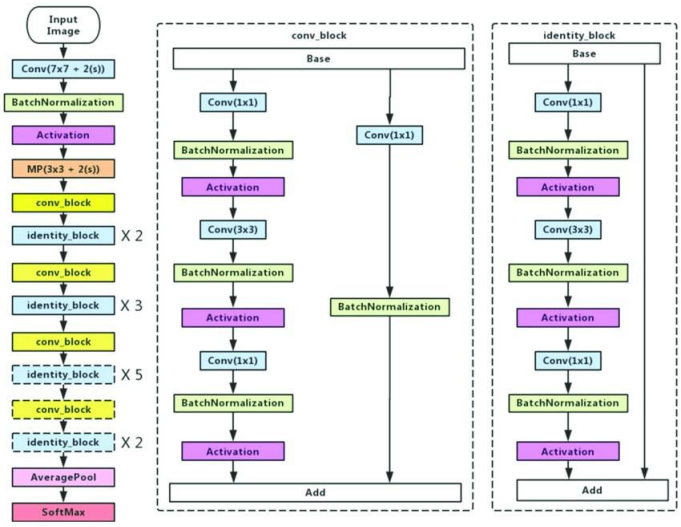


Figure 12: ResNet50 Architecture

8.4 Efficient Net B3

The EfficientNet network is based on a novel scaling strategy for CNN models. It employs a straightforward compound coefficient that is quite successful. Unlike existing approaches that scale network parameters such as width, depth and resolution, EfficientNet evenly scales each dimension with a given set of scaling factors. Scaling individual dimensions increases model performance in practice,

but balancing all network dimensions in relation to available resources significantly enhances overall performance.

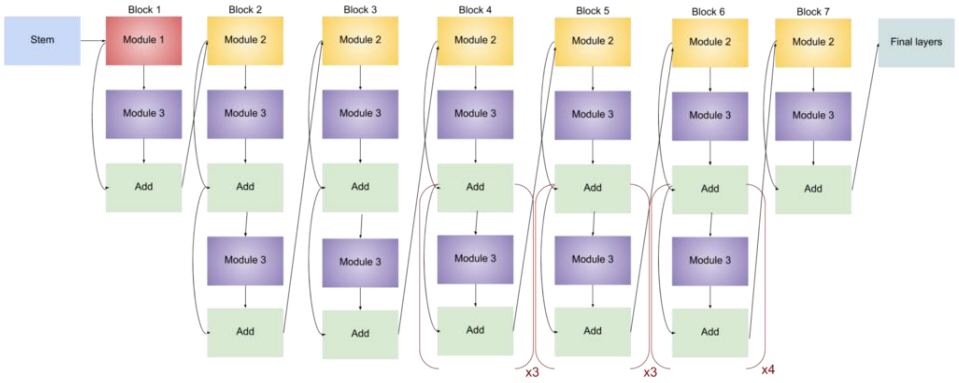


Figure 13: EfficientNet B3 Architecture

Pre trained weights from imagenet were used to load the model and then is compiled using stochastic gradient descent optimizer and sparse categorical cross entropy as loss function

Experiments on masked and unmasked dataset and its comparative analysis are shown in colab file.

8.5 Comparative Analysis

Models report			
Deep Learning Model	Training Accuracy	Validation Loss	Validation Accuracy
VGG 19	0.975	0.38	0.91
ResNet 50	0.88	0.79	0.83
EfficientNet B3	0.984	0.12	0.96

9 Major takeaways from the project

- **EDA**: Learned the complete process of EDA (Exploratory Data Analysis).

- **Data analysis**: Learned to withdraw some insights from the dataset both mathematically and by visualizing it.
- **Data visualization**: Learned to visualise the data and get better insight from it.
- **Models comparison**: Application of different models and their comparative analysis.
- **Optimization methods**: Application of different optimising methods to increase the accuracy of the models.
- **Image dataset**: Learned to work with large image dataset keeping time and space constraint in mind.
- **Covolutional Neural Networks**: Application of different CNN models on image datasets
- **Segmentation** : Application of image segmenation for optimizing the accuracies

10 References

- COVID-19 detection in X-ray images using convolutional neural networks
- U-Net: Convolutional Networks for Biomedical Image Segmentation
- Efficient Net
- Dataset Import using Keras tensorflow
- Identification of COVID-19 samples from chest X-Ray images using deep learning