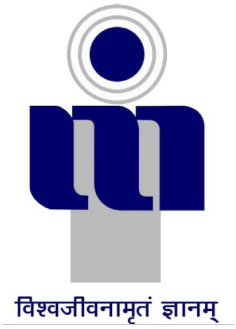
"Analysis of Covid-19 and Covid-19 Detection from Lung X-rays"

A minor project report,
submitted in partial fulfillment of the requirement for the award of
B.Tech. degree in computer science and engineering

by

Devicharith (2018BCS-019)



ABV INDIAN INSTITUTE OF INFORMATION
TECHNOLOGY AND MANAGEMENT
GWALIOR-474 015

2020

CANDIDATE DECLARATION

I hereby certify that the work, which is being presented in the report, entitled **Analysis of Covid-19 and Covid-19 Detection fromLung X-rays**, in partial fulfillment of the requirement for the award of the Degree of **Bachelor of Technology** and submitted to the institution is an authentic record of my work carried out during the period *July 2020* to *September 2020* under the supervision of **Dr. Gaurav Agrawal**. I also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Date: 9-9-2020 Signature of the Candidate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: 9-9-2020 Signatures of the Research Supervisors

ABSTRACT

The outbreak of Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome (SARS); in response to the rapidly increasing number of cases of the emerging disease, this Analysis attempts to provide a comprehensive review of the Covid-19 Virus. I hope that this Analysis helps in understanding and taking some precautions of the disease as the Analysis provides the significant symptoms of COVID-19 and what age groups are mostly getting affected and Analysis on the number of cases through a period for each state. So, in this particular situation, one primary thing that needs to be done is manual testing, so that the actual situation can be understood and appropriate decisions can be taken. But the drawbacks of manual testing include the availability of testing kits, which are costly and inefficient blood tests; a blood test takes hours to generate the result. So, the idea is to overcome these limitations using the Deep Learning technique for efficient treatment. The faster we produce the results, the fewer cases in the city, that's why we can use CNN to get our job done.

Keywords: Coronavirus, pneumonia, outbreak, SARS-CoV-2, COVID-19

ACKNOWLEDGEMENTS

I am highly indebted to Dr Gaurav Agrawal, and am obliged for giving me the autonomy of functioning and experimenting with ideas. I would like to take this opportunity to express my profound gratitude to her not only for her academic guidance but also for her personal interest in my project and constant support coupled with confidence-boosting and motivating sessions which proved very fruitful and were instrumental in infusing self-assurance and trust within me. The nurturing and blossoming of the present work are mainly due to her valuable guidance, suggestions, astute judgment, constructive criticism,and an eye for perfection. My mentor always answered a myriad of my doubts with smiling graciousness and prodigious patience, never letting me feel that we are novices by always lending an ear to my views, appreciating and improving them, and by giving me a free hand in my project. It's only because of her overwhelming interest and helpful attitude, the present work has attained the stage it has.Finally, I am grateful to our Institution and colleagues whose constant encouragement served to renew my spirit, refocus my attention and energy, and helped me in carrying out this work.

(Devi charith)

TABLE OF CONTENTS

ABSTRACT **ii**

LIST OF FIGURES **v**

LIST OF TABLES **vi**

ABBREVIATIONS **viii**

1 Introduction **1**

1.1 BACKGROUND INFORMATION 1
1.2 MOTIVATION 1
1.3 OBJECTIVES 2
1.4 SALIENT FEATURES 2
1.5 LITERATURE SURVEY 3

2 Methodology **4**

2.1 SYSTEM ARCHITECTURE 4
2.1.1 Convolution 4
2.1.2 Max_pooling 6
2.1.3 Flattening 6
2.1.4 Full Connection 7
2.2 DATA PREPROCESSING 8
2.2.1 BGR-RGB 9
2.2.2 Scaling 9
2.2.3 Splitting of Data into Train, Test, validation, and Data Augmentation 9
2.3 IMPLEMENTATION DETAILS 9
2.3.1 Dataset 9
2.3.2 Tools and Hardware 9
2.3.3 Hardware_Setup 10

3 Results and Discussion **11**
3.1 Analysis . . . 11
3.1.1 World Update . . . 11
3.1.2 India Updates . . . 12
3.1.2.1 Age Group Analysis . . . 12
3.1.2.2 Gender Analysis . . . 13
3.1.3 STATE-WISE ANALYSIS . . . 14
3.1.4 Major Symptoms . . . 15
3.1.5 CNN-MODEL . . . 16

4 Conclusion **19**
4.1 ADVANTAGES: . . . 19
4.2 LIMITATIONS: . . . 19
4.3 FUTURE SCOPE: . . . 20

References **20**

LIST OF FIGURES

2.1 from A-Z machine learning udemy 4
2.2 Convolution2 from A-Z machine learning udemy 5
2.3 max_pool from A-Z machine learning udemy 6
2.4 Flatten from A-Z machine learning udemy 7
2.5 System Architecture 7
2.6 Non-covid 8
2.7 covid 8

3.1 Worldwide Analysis 11
3.2 Age_Group Analysis 12
3.3 Missing Gender Analysis 13
3.4 Gender Analysis 13
3.5 state wise-1 14
3.6 state wise-2 14
3.7 state wise-3 15
3.8 Accuracy metrics for VGG-16 16
3.9 Loss metrics for VGG-16 17
3.10 Accuracy metrics for Inception-16 17
3.11 Loss metrics for Inception-16 18

LIST OF TABLES

1.1 Related Work 3

2.1 Ratio of covid and non-covid images 9

3.1 Symptoms 15
3.2 Accuracy Table 18

ABBREVIATIONS

ANN	Artificial Neural Network
BP	Back Propagation
CNN	Convolutional Neural Network
MERS	Middle East Respiratory Syndrome
SARS	Severe Acute Respiratory Syndrome

CHAPTER 1

Introduction

On December 31, 2019, a great number of cases of pneumonia of unknown cause, in the city of Wuhan in China, was reported to the World Health Organisation. In January 2020, an anonymous new virus was identified and named the 2019 novel coronavirus.

1.1 BACKGROUND INFORMATION

Coronavirus is a family of viruses that affect the respiratory system of a person.Respiratory diseases can be the common cold to more severe diseases as SARS and MERS.

- Middle East Respiratory Syndrome (MERS-CoV)
- Severe Acute Respiratory Syndrome (SARS-CoV)

Corona viruses got its name because of the way they look under a microscope.The virus consists of a genetic material surrounded by an envelope with protein spikes, which appears like a crown. The word Corona means "crown" in Latin.
Convolutional Neural Network is a special Neural Network used for image recognition. CNN are much better than ANN because after learning a certain pattern in a picture, a convolutional network can recognize it anywhere, CNN takes advantage of local spatial coherence of images [2].

1.2 MOTIVATION

As humans we can't differentiate an image without having a label on them but a cnn can distinguish it.
We can distinguish the x-ray image using a deep-learning method calledConvolution-Neural-Networks, which stores the features of the different labelimages.

Limitation:

- The amount of time required for the result and the cost for manual testing
- And the limited availability of test kits.

So we need a technique that provides the result in a faster way and reduces the cost to check if the person is affected by covid-19 or not. Still, this model can return good accuracy and can be further improved.

1.3 OBJECTIVES

- The objective of the project is to do an analysis on Covid-19.
- Later develop and train a deep learning model on the chest X-Ray dataset to classify the given x-ray images of affected patients vs healthy people
- And train some pre-trained models like alex-net and VGG-16 and compare the accuracies of the different models.

1.4 SALIENT FEATURES

- After learning a particular pattern in an image, a convnet can recognize the pattern no matter where it is in the image. AN Artificial Neural Network will have to learn the same pattern if it appears in a different location even if it's the same pattern. This makes convnets data efficient when processing images 11they need fewer training samples to learn representations that have generalization power[1].

- They can learn spatial hierarchies of patterns. The first convolution layer will learn small local patterns, a second convolution layer with max_pool and non_linearity will learn larger patterns made of the features of the first layers,and so on. This allows convnets to efficiently learn all the features.[1]

1.5 LITERATURE SURVEY

Table 1.1: Related Work

S.No.	Author	Title	Year	Publisher	Work
1	François Cholle	Deep Learning with python	2017	Google	Understanding neural networks, CNN and Back-propagation
2	C. Szegedy et al	Going deeper with convolutions," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition	2015	IEEE	Inception-V2 is a combined architecture proposed by that uses the idea of inception blocks and residual layers together
3	Taban Majeed,Rasber Rashid,Dashti Ali, and Aras Asaad	Covid-19 detection using CNN transfer learning from X-ray Images	2020	doi.org	Quantitative analysis to evaluate 12 off-the-shelf convolutional neural networks (CNNs) for the purpose of COVID-19 X-ray image analysis
4	T. Ai et al	Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease (COVID-19)	2020	pubs.rnsa.org	Review studies devoted to the use of radiography images to aid and complement PCR in China diagnosing COVID-19 case.
5	H. S. Maghdid, A.T. Asaad,K. Z. Ghafoor,A. S. Sadiq, and M. K. Khan	"Diagnosing COVID-19 Pneumonia from X-Ray and CT Images using Deep Learning and Transfer Learning Algorithms"	2020	arxiv.org	A model of 16layers that can detect covid-19 using ct scans and x-rays on small datasets.
6	K. Simonyan and A. Zisserman.	Very Deep Convolutional Networks for Large-ScaleImage Recognition	2015		Brief knowledge about vgg16.

CHAPTER 2

Methodology

The Methodology includes 4 steps:

- Convolution operation
- Pooling
- Flattening
- Full Connection

2.1 SYSTEM ARCHITECTURE

2.1.1 Convolution

As we are dealing with images in computer terms, an image is just a 2d matrix with pixel values between 0 and 255 including them where 0 represents brightness, and 255 represents black, so a black and white image is nothing but a matrix with pixel values 0's and 255's

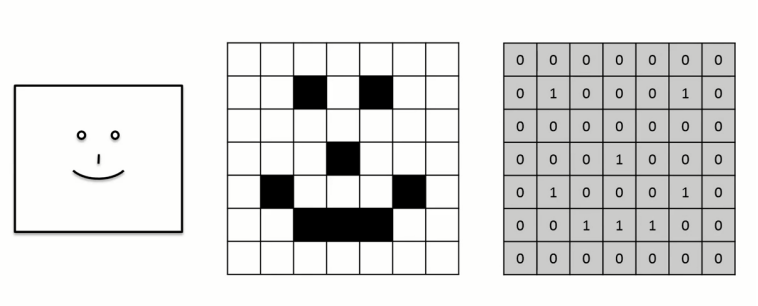


Figure 2.1: from A-Z machine learning udemy

As we can see the image above the rightmost matrix represents the The leftmost image where the pixel value 1 represents black pixel and white as 0. We've got an input

image as we discussed that's how we're going to look at images just ones and zeros as we can see the input image it's the smiling image we looked before.

Figure 2.2 describes about Convoultion.

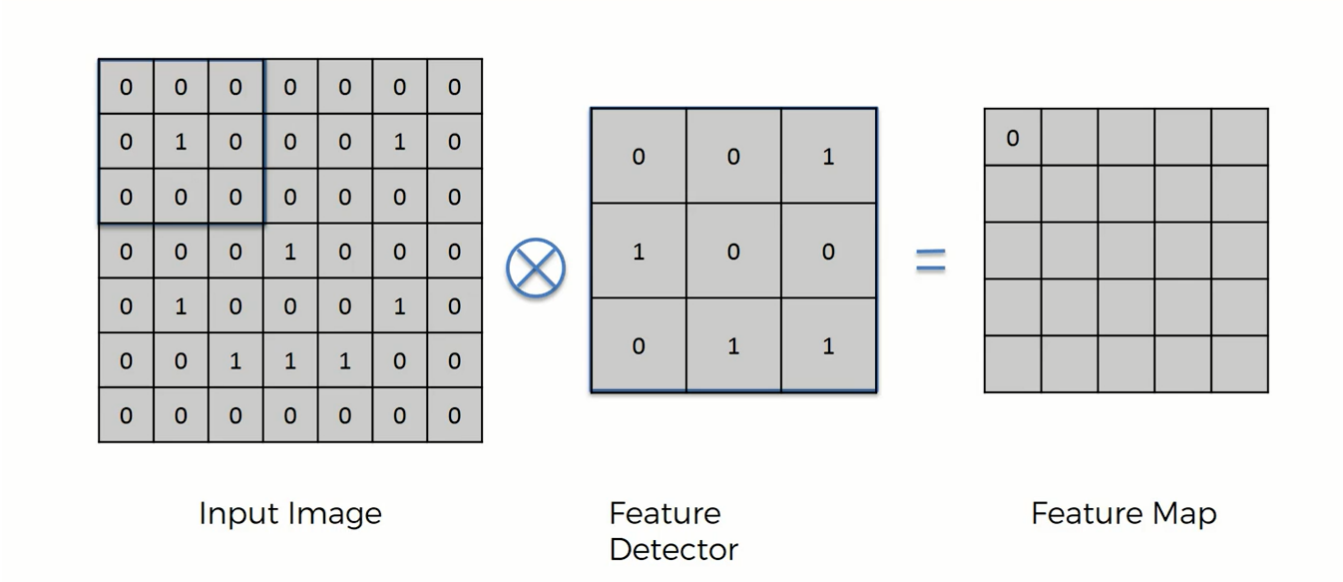


Figure 2.2: Convolution2 from A-Z machine learning udemy

This is called Feature map and sometimes called convolved feature.

And that's a very important function of the feature detector of this whole convolution step is to make the image smaller because it'll be easier to process and it'll be just faster.

we are losing some part of the information but at the same time, we are storing the features of the image. That's what a feature detector does for us; it detects the important features for us.We get the max value when the pattern matches in the image.

We don't just have a single feature detector we need to detect as many as features as possible from the input image to classify it so we have different feature detectors for an input image for the feature map we used a feature detector similar to the one we just saw for the next one we use a different feature detector[1].

2.1.2 Max_pooling

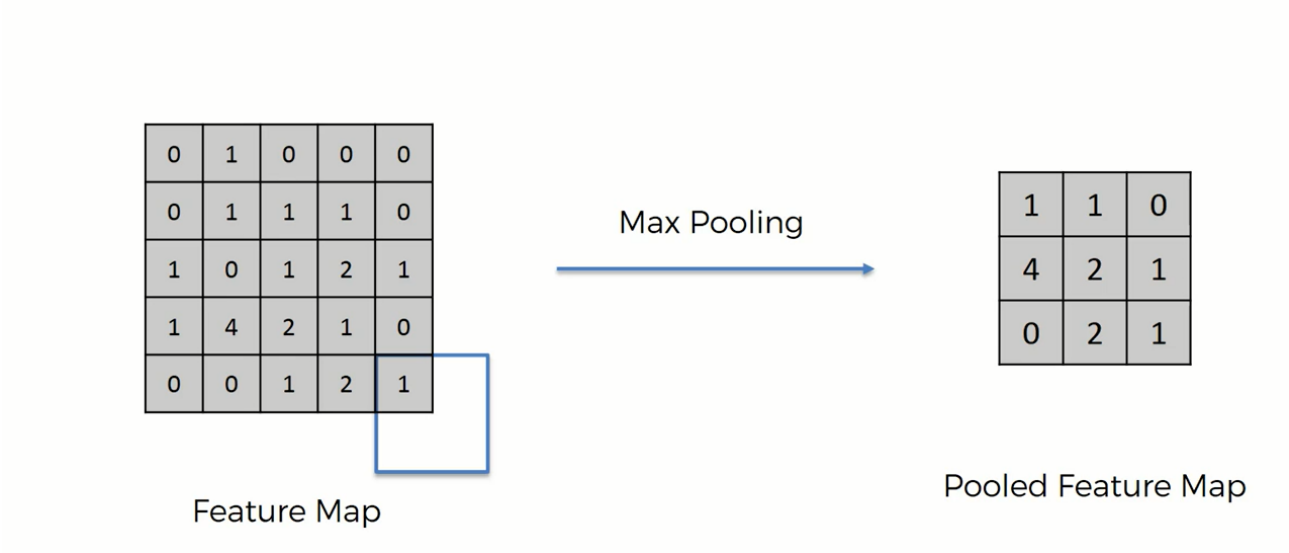


Figure 2.3: max_pool from A-Z machine learning udemy

let we want to identify a cat in an image so in an image, the cheetah can look in one direction in another image it can look in another direction it can be in one part in an image it can be in another position so at the end we need to identify it like a cheetah no matter what direction it's looking or where it is in the image.

One of the distinct features of the cheetah is the tears that are going from its eyes to the side of its nose so if the neural net is trying to find the feature it should find it no matter where it is in the image and what direction the cheetah is looking.

There are several different types of pooling methods like mean pooling Max pooling But for now, we're just applying Max pooling so we take a box of two by two pixels[1].

2.1.3 Flattening

We just flatten the max-pooling matrix into a vector so that we can use an input for the neural network later on for classification.

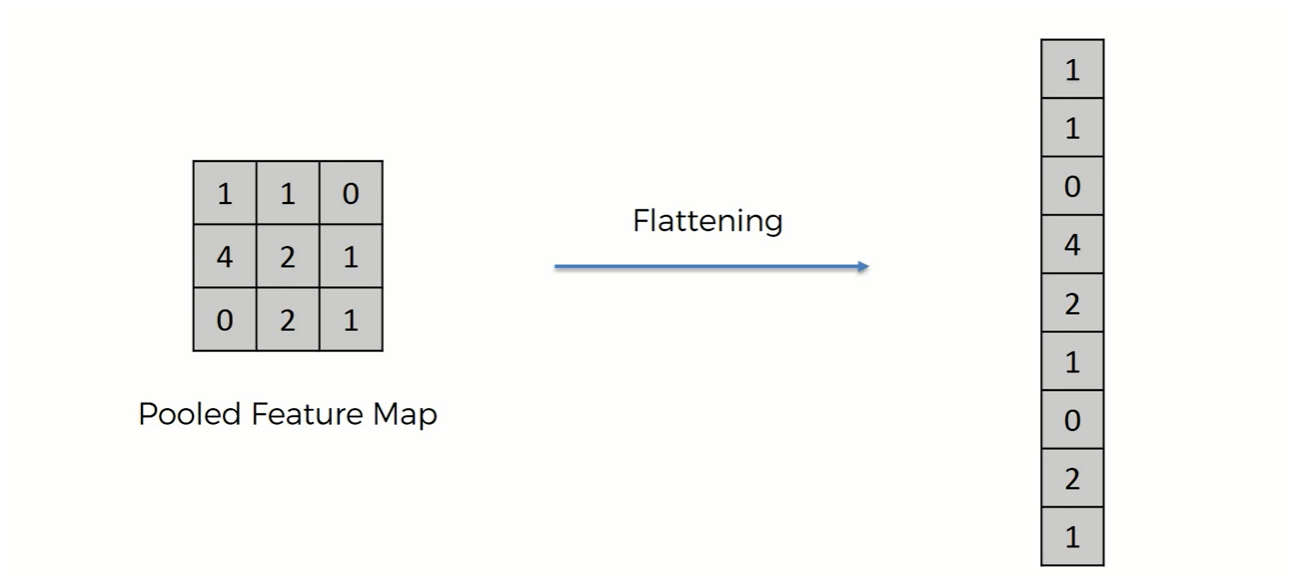


Figure 2.4: Flatten from A-Z machine learning udemy

2.1.4 Full Connection

The flatten max pooled matrix is the input for artificial neural networks.Initially the neurons in the neural network have random weights and the weights are changed accordingly to maximize the output accuracy.

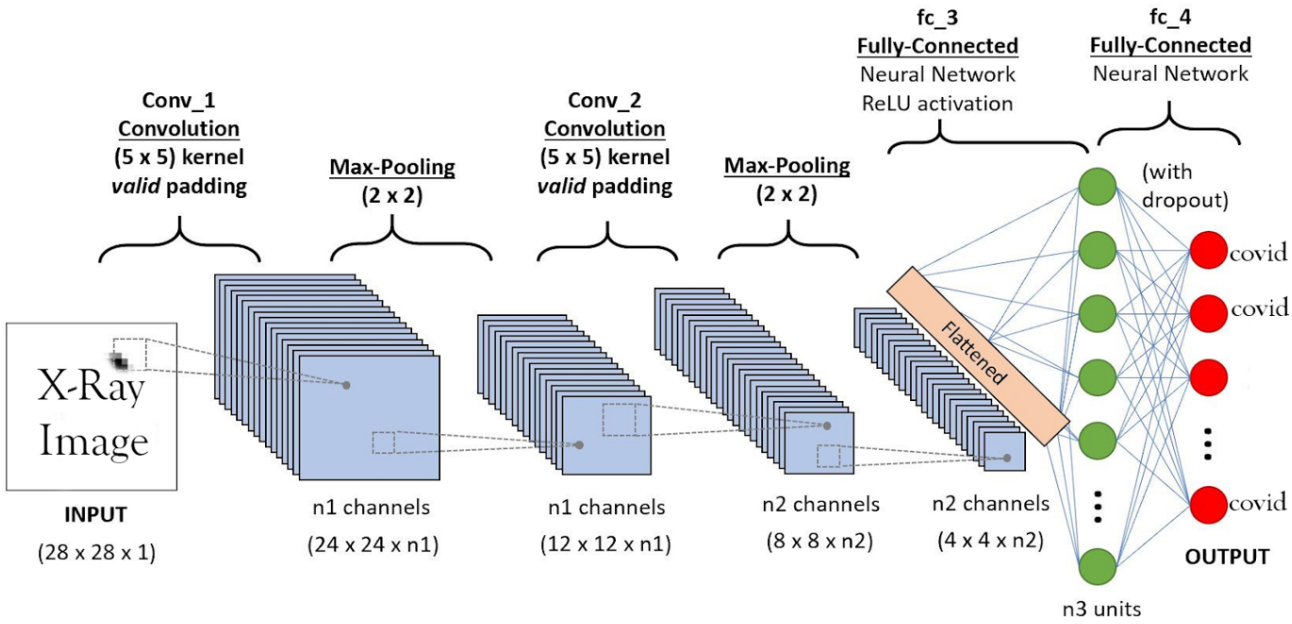


Figure 2.5: System Architecture

2.2 DATA PREPROCESSING

These image paths are stored in a dataset path_images in a column and labels of the image and then we iterate a loop over the dataset and classify these images into their respective directories.

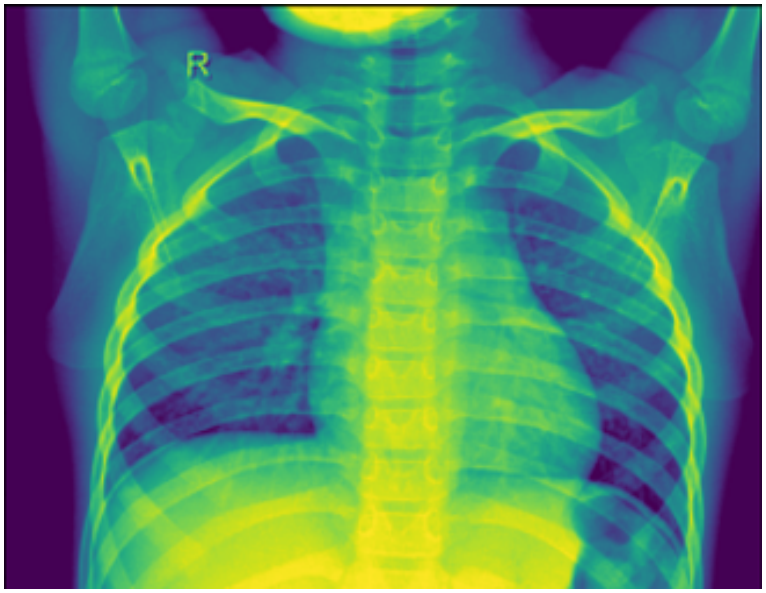


Figure 2.6: Non-covid

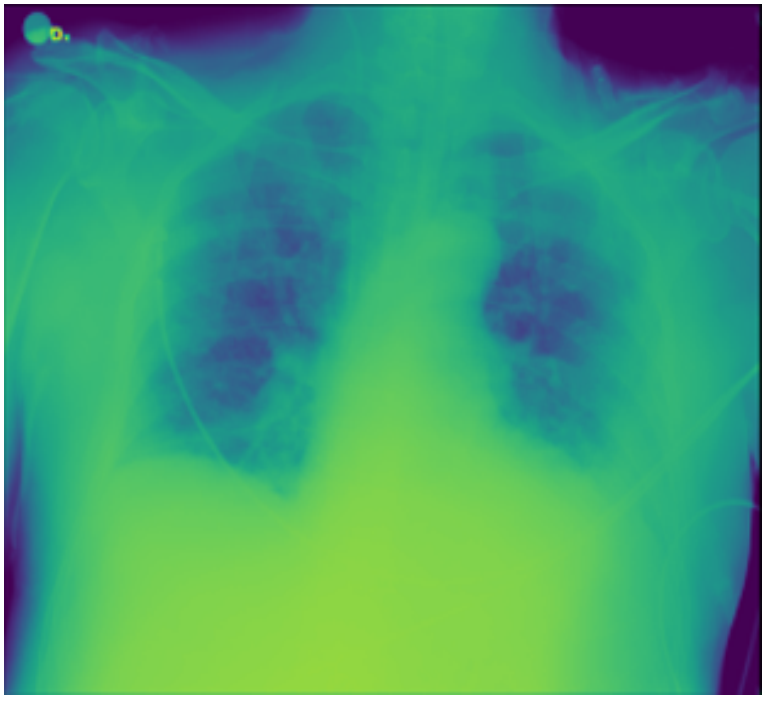


Figure 2.7: covid

2.2.1 BGR-RGB

As we are making use of a pre-trained model VGG-16. VGG-16 takes an RGB image. which are of size 224*224 as most of the images are BGR we convert them to RGB using the cv2 module.

2.2.2 Scaling

Scaling the pixel values are between 0 and 255 we divide each pixel value by255 so that we scale them down to 0 and 1.

2.2.3 Splitting of Data into Train, Test, validation, and Data Augmentation

Total data in the datasets are divided into Test, Train data accordingly.
Train Data: 80% or 0.75
Validation Data: 10% or 0.1
Test Data: 10% or 0.1

Table 2.1: Ratio of covid and non-covid images

	Training	Test
non-covid	1072	269
covid	113	28

As we have few images data-augmentation plays a key role here.Data-augmentation is a technique to increase the diversity of your training set by applying random transformations such as image rotation and flip images in the vertical and horizontal direction.

2.3 IMPLEMENTATION DETAILS

2.3.1 Dataset

- A dataset containing information regarding COVID 19 in India.
- The dataset is a collection of X Rays of Chest with X Rays of bothHealthy lungs and Pneumonia Lungs i.e Lungs affected by Corona.

2.3.2 Tools and Hardware

- Python

- Numpy
- Pandas
- Keras or Tensorflow
- Seaborn
- sklearn
- CPU or GPU(recommended)

2.3.3 Hardware_Setup

- If your computer does not have GPU(GPU is recommended for faster Computation) we can use GPUs that are provided by Google or Kaggle.
- Kaggle is recommended as the datasets are available on kaggle. Datasets can be loaded with ease if you are using Kaggle kernel.
  If you want to do it on Google Colab you need to upload the datasets to your Google Drive first.
- In a kaggle kernel we can use the GPU thats provided for us by Select the Settings tab.
  Then select the checkbox for Enable GPU. Verify the GPU is attached to your kernel in the console bar, where it should show GPU ON next to your resource usage metrics.
  And turn on the internet to download the relvent libraries.

CHAPTER 3

Results and Discussion

3.1 Analysis

3.1.1 World Update

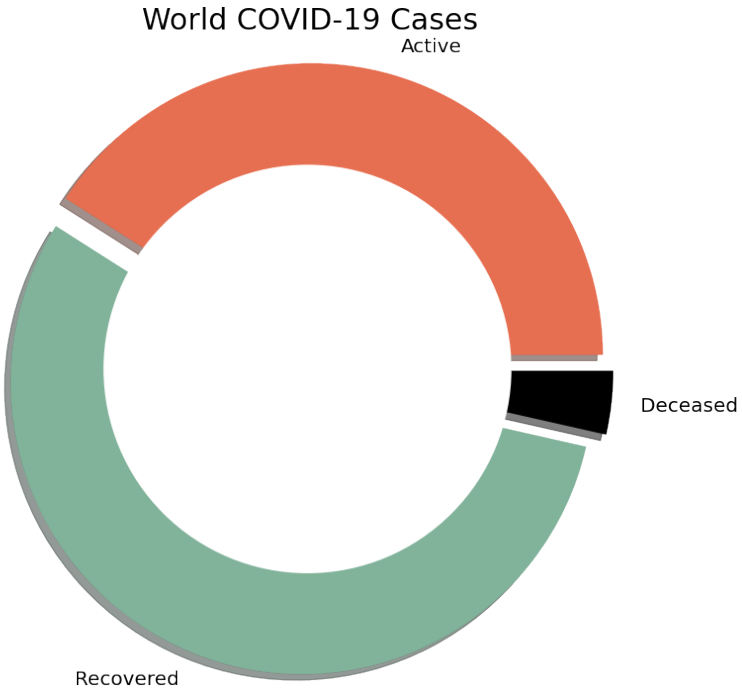


Figure 3.1: Worldwide Analysis

Figure 3.1 The analysis is made from data that is available till:

- 3rdAug202040.991% of the world population is still suffering from Covid-19.
- 55.481% of the world population recovered from Covid-19.
- And 3.5% deceased.

3.1.2 India Updates

3.1.2.1 Age Group Analysis

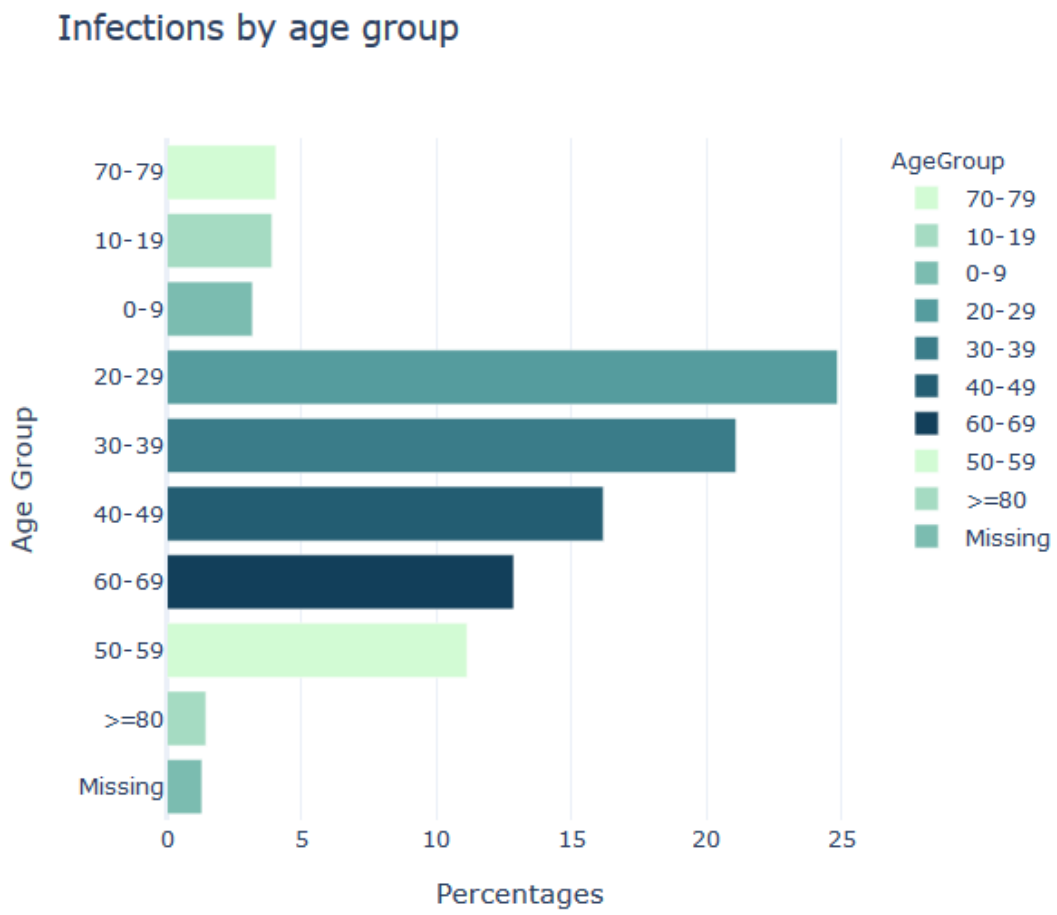


Figure 3.2: Age_Group Analysis

According to Figure 3.2 The most affected age groups between 20-60. And the least affected are ages between 0-19 and >60 which makes upto 17%.

3.1.2.2 Gender Analysis

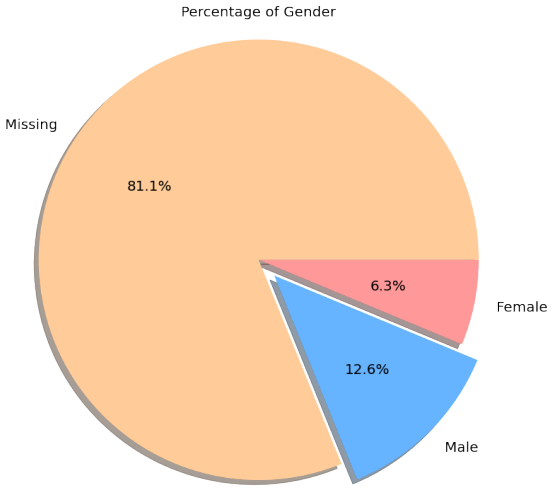


Figure 3.3: Missing Gender Analysis

Figure 3.3 Most of the data regarding Gender is so if a complete data is available the pie chart changes according to the data. Where 81% missing,Male:12.6%,women:6.3%

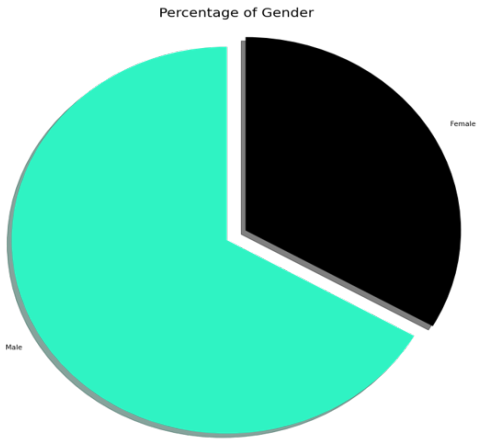


Figure 3.4: Gender Analysis

Figure 3.4 Neglecting the missing data male is the most affected compared gender.

3.1.3 STATE-WISE ANALYSIS

	State/UnionTerritory	Confirmed	Deaths	Cured	Active	Death Rate (per 100)	Cure Rate (per 100)
22	Maharashtra	431719	15316	266883	149520	3.550000	61.820000
32	Tamil Nadu	251738	4034	190966	56738	1.600000	75.860000
1	Andhra Pradesh	150209	1407	76614	72188	0.940000	51.000000
11	Delhi	136716	3989	122131	10596	2.920000	89.330000
18	Karnataka	129287	2412	53648	73227	1.870000	41.500000
39	Uttar Pradesh	89048	1677	51334	36037	1.880000	57.650000
41	West Bengal	72777	1629	50517	20631	2.240000	69.410000
35	Telengana	64786	530	46502	17754	0.820000	71.780000
13	Gujarat	62463	2464	45699	14300	3.940000	73.160000
4	Bihar	54240	309	34994	18937	0.570000	64.520000
30	Rajasthan	42646	690	29977	11979	1.620000	70.290000
3	Assam	41726	101	31442	10183	0.240000	75.350000
14	Haryana	35758	428	29080	6250	1.200000	81.320000
27	Odisha	33479	187	21274	12018	0.560000	63.540000
21	Madhya Pradesh	32614	876	22969	8769	2.690000	70.430000
19	Kerala	24742	81	13775	10886	0.330000	55.670000

Figure 3.5: state wise-1

16	Jammu and Kashmir	20972	388	12871	7713	1.850000	61.370000
29	Punjab	17063	405	11075	5583	2.370000	64.910000
17	Jharkhand	11686	113	4513	7060	0.970000	38.620000
7	Chhattisgarh	9385	55	6610	2720	0.590000	70.430000
5	Cases being reassigned to states	9265	0	0	9265	0.000000	0.000000
40	Uttarakhand	7447	83	4330	3034	1.110000	58.140000
12	Goa	6193	48	4438	1707	0.780000	71.660000
37	Tripura	5233	23	3463	1747	0.440000	66.180000
28	Puducherry	3606	51	2198	1357	1.410000	60.950000
23	Manipur	2756	6	1699	1051	0.220000	61.650000
15	Himachal Pradesh	2634	14	1502	1118	0.530000	57.020000
26	Nagaland	1831	5	640	1186	0.270000	34.950000
2	Arunachal Pradesh	1673	3	969	701	0.180000	57.920000
20	Ladakh	1462	7	1108	347	0.480000	75.790000
9	Dadra and Nagar Haveli and Daman and Diu	1145	2	725	418	0.170000	63.320000
6	Chandigarh	1079	18	683	378	1.670000	63.300000
24	Meghalaya	856	5	252	599	0.580000	29.440000

Figure 3.6: state wise-2

31	Sikkim	650	1	269	380	0.150000	41.380000
0	Andaman and Nicobar Islands	636	7	226	403	1.100000	35.530000
25	Mizoram	468	0	253	215	0.000000	54.060000
38	Unassigned	77	0	0	77	0.000000	0.000000
8	Dadar Nagar Haveli	26	0	2	24	0.000000	7.690000
10	Daman & Diu	2	0	0	2	0.000000	0.000000

Figure 3.7: state wise-3

Figure 3.7 Of all the states and Union Territories, Only Lakshadweep has not reported any case so far the remaining 35 have reported Covid-19 cases so far.On August 3, the total number of cases reached 19,690,596 in which 11,68,023have recovered and 38,135 have died so far.

The most affected states according the analysis are in the order

- Maharashtra
- TamilNadu
- AndhraPradesh
- Delhi

Till 3rd Aug. These states mentioned above also have the most number of death rates. Makes upto 60 percent of the country's Covid-19 toll as you can see from above DataFrame.

Cure-rate in Delhi, Haryana, Tamilnadu are high when compared to otherstates.

3.1.4 Major Symptoms

Table 3.1: Symptoms

Symptoms	Training
Fever	88.1%
Dry-cough	67.7%
Fatigue	38.1%
Difficult of Breath	18.6%
Sore throat	13.9%
Headache	13.3%

3.1.5 CNN-MODEL

The performance metrics used to evaluate the model are the accuracy and the loss. our main aim was to increase the accuracy at the same time decrease the loss.//

- The figure shows the comparison between accuracy and epochs, and also between loss and epochs. From the figure the accuracies i.e., both the training and validation accuracies are increasing from the first epoch itself which tells that the model is learning.
- TensorFlow Framework was used. The dataset is taken from the Kaggle.

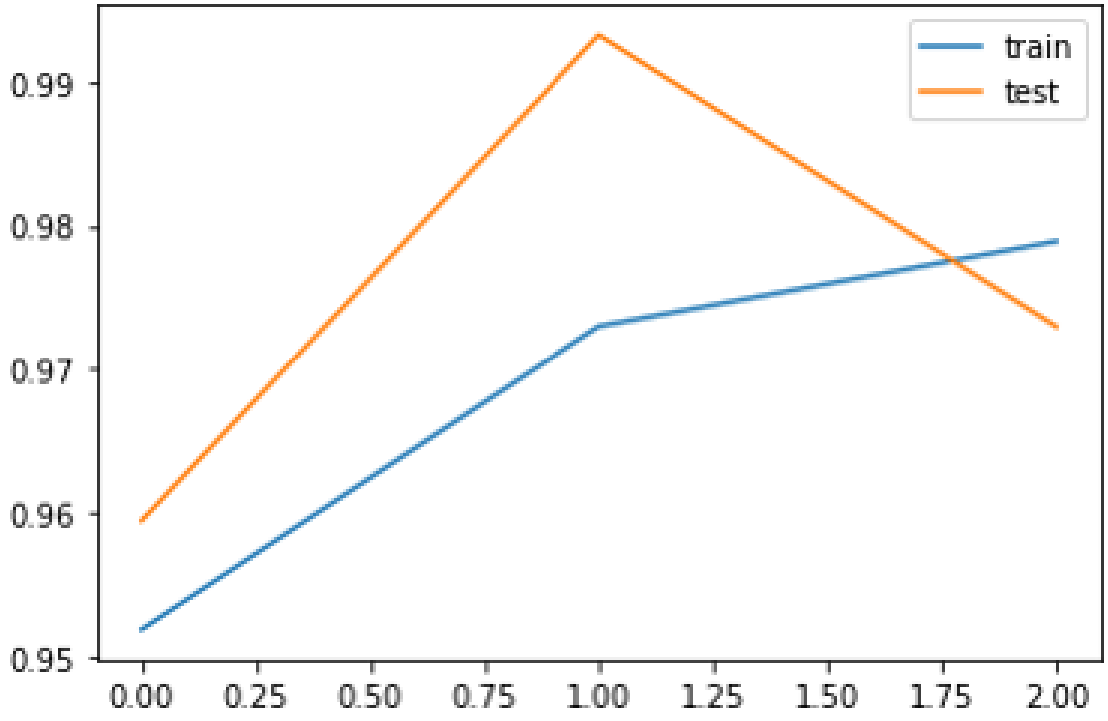


Figure 3.8: Accuracy metrics for VGG-16

From the Figure 3.8 it can be observed that the model's training accuracy and validation accuracy are increasing for each epoch thus the model is neither over fitting nor underfitting.

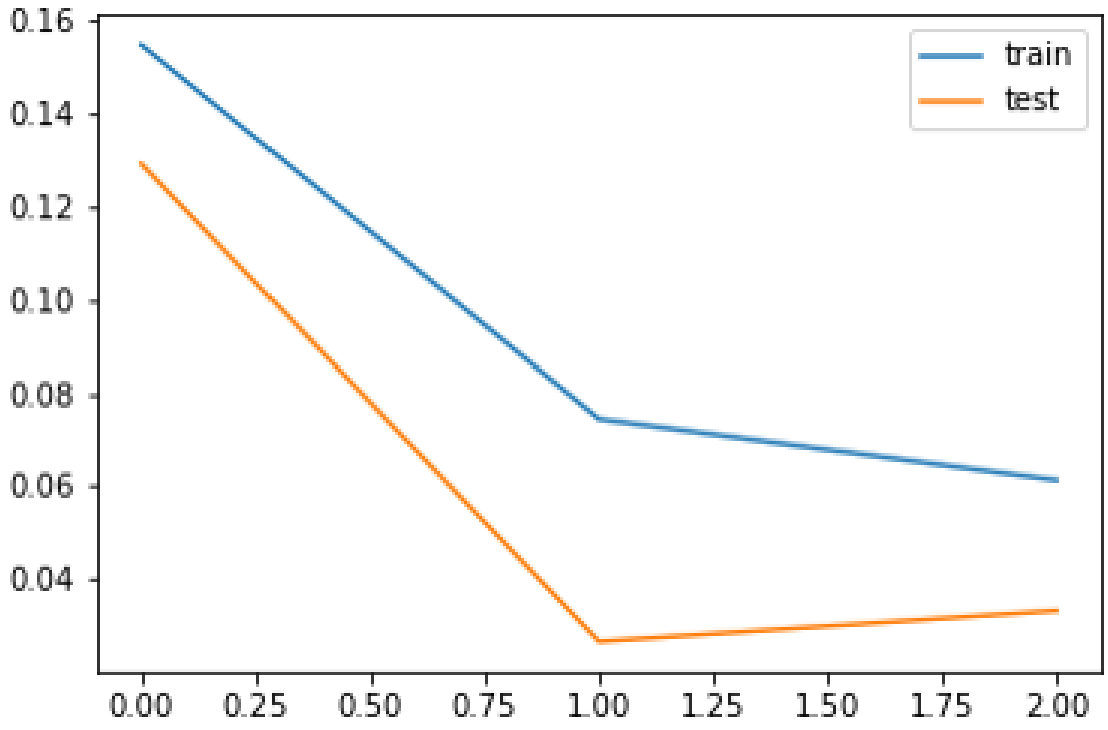


Figure 3.9: Loss metrics for VGG-16

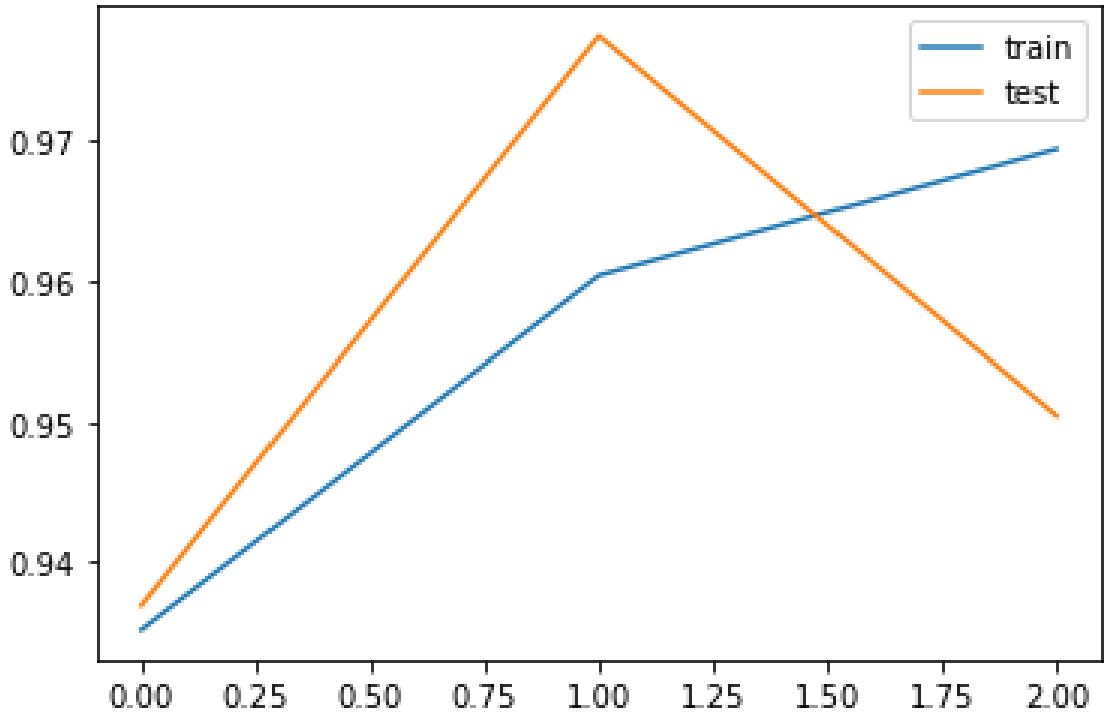


Figure 3.10: Accuracy metrics for Inception-16

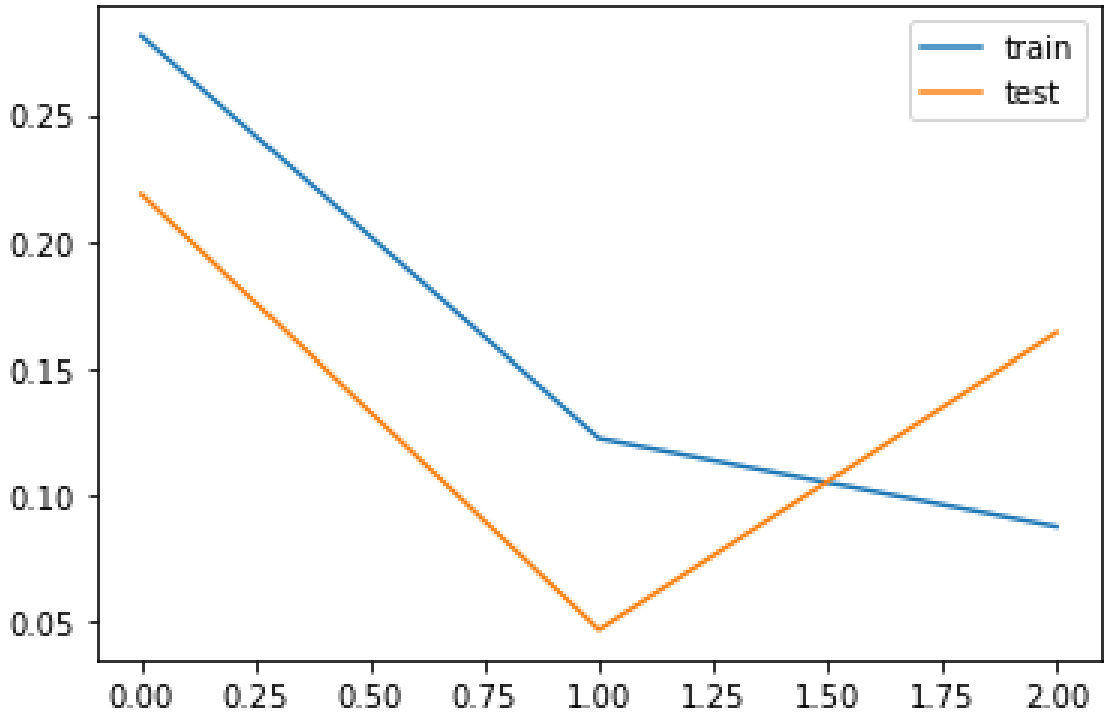


Figure 3.11: Loss metrics for Inception-16

These metrics are for 3 epochs and a learning rate of 0.001. If the no.of are increased overfitting is observed we used just 3 epochs as the available data is less.

Table 3.2: Accuracy Table

Model	Train Accuracy	Test Accuracy
VGG-16	96.572	97.323
InceptionV3	95.65	93.63
CNN from Scratch	70.98	65.39

Note: These results are for 1000 normal x-ray images and 141 COVID-images these metrics may change if the no.of input images change.

CHAPTER 4

Conclusion

This report provides a critical analysis for 3 CNN-Architectures, proposed originally for image analysis. These CNN-Architectures are used to differentiate COVID-19 disease vs healthy based on chest X-ray images.
We also proposed a simple CNN architecture but gave us low accuracy when compared to other standard CNN architectures it was observed that a learning rate of 0.01 achieved good accuracy the number of epochs is restricted to 3 as any other value greater than that over fitting is observed.
And the analysis on the COVID-19 covering what age groups are mostly getting affected, Major symptoms, state-wise analysis with cure rates, and death rates. The result of the CNN that we built from scratch and compared with other deep learning CNN models such as VGG16, and InceptionV3.
The pre-trained models came up with higher accuracy when compared to the model that we have built from scratch. VGG-16 and InceptionV3showed the same metrics.

4.1 ADVANTAGES:

1. The amount of time required for the result and the cost is less when compared to manual testing.

2. And the limited availability of test kits.

3. And can be used for different x-ray models for classification.

4.2 LIMITATIONS:

- Number of x-rays we have are really less. For better accuracy we need more images.

- And one more important limitation is that not every persons lungs get affected by covid-19.

4.3 FUTURE SCOPE:

The future work may be:

1. Removing other unnecessary noise in the image such as text writing and medical devices marked on chest X-rays for a better vision and understanding.
2. And the collection of more COVID Images will improve the models accuracy.

REFERENCES

[1] Deep learning with François Chollet

[2] Taban Majeed, Rasber Rashid, Dashti Ali, and Aras Asaad, *"Problems of Deploying CNN Transfer Learning to Detect COVID-19 from Chest X-rays"*, May 19, 2020.

[3] C. Szegedy et al.*"Going deeper with convolutions"*, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol.07-12-June-2015, pp. 1–9

[4] T. Ai et al,*"Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease(COVID-19) in China: A Report of 1014 Cases"*, Radiology, p. 200642, Feb. 2020

[5] K. Simonyan and A. Zisserman,*"Very Deep Convolutional Networks for Large-ScaleImage Recognition"*, 2015.

[6] H. S. Maghdid, A. T. Asaad, K. Z. Ghafoor, A. S. Sadiq, and M. K. Khan,*"Diagnosing COVID-19 Pneumonia from X-Ray and CT Images using Deep Learning and Transfer Learning Algorithms"*, arXiv, Mar. 2020.