# Covid – 19 Prediction

**By**

G.Jai Surya Gowd – 17MIS1122

K.Sai Prakash – 17MIS1152

M.Sri Teja – 17MIS1164

S.Pavan Kalyan – 17MIS1131

**Under The Guidance**

Prof. Bhuvaneswari Anbalagan

**Fall Sem 2020 -21**


VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

## Abstract :

The COVID-19 pandemic is causing a major outbreak in all around the world, having a severe impact on the health and life of many people in world. One of the crucial step in fighting COVID-19 is the ability to detect the infected patients early enough, and put them under special care. To find the patient is effected with corona first we have to know the symptoms of the corona. By using Symptoms we can perform Machine Learning algorithms on the dataset of the Symptoms dataset to find the corona is positive or negative. We use machine learning algorithms like Logistic regression and Decision tree algorithms to predict the corona. And we can also predict the corona virus by X-ray images of lungs. We can do this by using deep learning methods like Convolution Neural Network(CNN).By this we can detect the infected patients early.

*Keywords : Covid-19, Symptoms, Machine Learning, Logistic regression, Decision tree, X-ray, Deep learning, CNN.*

## Introduction :

The coronavirus disease pandemic emerged in Wuhan, China in December 2019 and became a serious public health problem worldwide. Til now, there is no specific drug or vaccine has been found for corona. COVID-19 causes lighter symptoms in about 99% of cases, according to early data, while the rest is severe or critical. As of 4th October 2020, the total number of worldwide cases of Coronavirus is 35,248,330. Of these, 1,039,541 (4%) people were deaths and 26,225,235 (96%) were recovered. The number of active patients is 7,983,554. Of these, 7,917,287 (99%) had mild disease while 66,267 (1%) had more severe disease. Nowadays the world is struggling with the COVID-19 epidemic. Deaths from pneumonia developing due to the SARS-CoV-2 virus are increasing day by day.

X-ray is one of the most important methods used for the diagnosis of pneumonia worldwide. X-ray is a fast, cheap and common clinical method. The X-ray gives the patient a lower radiation dose compared to computed tomography (CT) and magnetic resonance imaging (MRI). However, making the correct diagnosis from X-ray images requires expert knowledge and experience. It is much more difficult to diagnose using a chest X-ray than other imaging modalities such as CT or MRI.

By looking at the chest X-ray, COVID-19 can only be diagnosed by specialist physicians. The number of specialists who can make this diagnosis is less than the number of normal doctors. Even in normal times, the number of doctors per person is insufficient in countries around the world. According to data from 2017, Greece ranks first with 607 doctors per 100,000 people. In other countries, this number is much lower.

In case of disasters such as COVID-19 pandemic, demanding health services at the same time, collapse of the health system is inevitable due to the insufficient number of hospital beds and health personnel. Also, COVID-19 is a highly contagious disease, and doctors, nurses, and caregivers are most at risk. Early diagnosis of pneumonia has a vital importance both in terms of slowing the speed of the spread of the epidemic by quarantining the patient and in the recovery process of the patient.

Doctors can diagnose pneumonia from the chest X-ray more quickly and accurately thanks to computer-aided diagnosis (CAD). Use of artificial intelligence methods are increasing due to its ability to cope with enormous datasets exceeding human potential in the field of medical services. Integrating CAD methods

into radiologist diagnostic systems greatly reduces the workload of doctors and increases reliability and quantitative analysis. CAD systems based on deep learning and medical imaging are becoming more and more research fields.

**Literature Survey**

| S.NO | Title & Author | Year & Publication | Techniques | Limitations |
|---|---|---|---|---|
| 1. | **Identifying pneumonia in chest X-rays: A deep learning approach**<br>**AK Jaiswal, P Tiwari, S Kumar, D Gupta, A Khanna** | **2019**<br>***Elviser*** | **Deep learning techniques** | • **Cloud Based System**<br>• **Information about** |
| 2. | **Diagnosis of ambulatory community-acquired pneumonia**<br>**D Lieberman, P Shvartzman, I Korsonsky** | **2003**<br>**Taylor & Francis** | **Physical test** | • **Counting on arbitatry paients** |
| 3. | **A transfer learning method with deep residual network for pediatric pneumonia diagnosis**<br>**G Liang, L Zheng** | **2020**<br>**Elsevier** | **Deep learning**<br>**Residual network** | **children pneumonia classification task** |
| 4. | **Guidelines for the initial management of adults with community-acquired pneumonia: diagnosis, assessment of severity, and initial antimicrobial therapy.**<br>**MS Niederman, JB Bass Jr, GD Campbell** | **1993**<br>**europepmc.org** | **Web application** | • **Search Doctor**<br>• **Online Appointment** |

| | | | | |
|---|---|---|---|---|
| **5.** | **Intelligent and effective heart disease prediction system using weighted associative classifiers**<br>**J Soni, U Ansari, D Sharma , S Soni** | **2011**<br>**academia.edu** | **Android Application** | • **Patient records**<br>• **Order medicine** |
| **6.** | **Clinical utility of different lipid measures for prediction of coronary heart disease in men and women**<br>**E Ingelsson, EJ Schaefer, JH Contois, JR McNamara** | **2007**<br>**jamanetwork.com** | **Web application** | • **Patient monitoring**<br>• **Measuring lipids** |
| **7.** | **A real time patient monitoring system for heart disease prediction using random forest algorithm**<br>**S Sreejith, S Rahul, RC Jisha** | **2016**<br>**Springer** | **Random forest algorithm** | • **Monitor patient health** |
| **8.** | **Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network**<br>**S Radhimeenaksh** | **2016**<br>**IEEE** | **Data mining techniques and arittifical network** | • **Mointoring the patient**<br>• **Patient records about their previous diseases** |

# Architecture and Algorithms :

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.[3] Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

## Logistic Regression

Logstic regresion is a superised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent riable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Input values (x) are combined linearly using eights or cefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key dference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the bis or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an aociated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's).

Written another way, we are modeling the probability that an input (X) belongs to the default class (Y=1), we can write this formally as:

P(X) = P(Y=1|X)

We're predicting probabilities? I thought lgistic regression was a classification algorithm?

Note that the probability prediction must be tansformed into a binary values (0 or 1) in order to actually make a probability prediction. More on this later when we talk about making predictions.

Logistic regression is a linear method, but the predictions are transformed using the logistic function. The impact of this is that we can no longer understand the predictions as a linear combination of the inputs as we can with linear regression, for example, continuing on from above, the model can be stated as:

p(X) = e^(b0 + b1*X) / (1 + e^(b0 + b1*X))

I don't want to dive into the math too much, but we can turn around the above equation as follows (remember we can remove the e from one side by adding a natural logarithm (ln) to the other):

ln(p(X) / 1 – p(X)) = b0 + b1 * X

This is useful because we can see that the calculation of the output on the right is linear again (just like linear regression), and the input on the left is a log of the probability of the default class.

This ratio on the left is called the odds of the default class (it's historical that we use odds, for example, odds are used in horse racing rather than probabilities). Odds are calculated as a ratio of the probability of the event divided by the bability of not the event, e.g. 0.8/(1-0.8) which has the odds of 4. So we could instead write:

ln(odds) = b0 + b1 * X

Because the odds are log transformed, we call this left hand side the log-odds or the probit. It is possible to use other types of functions for the transform (which is out of scope_, but as such it is common to refer to the transform that relates the linear regression equation to the probbilities as the link function, e.g. the probit link function.

We can move the exponent back to the right and write it as:

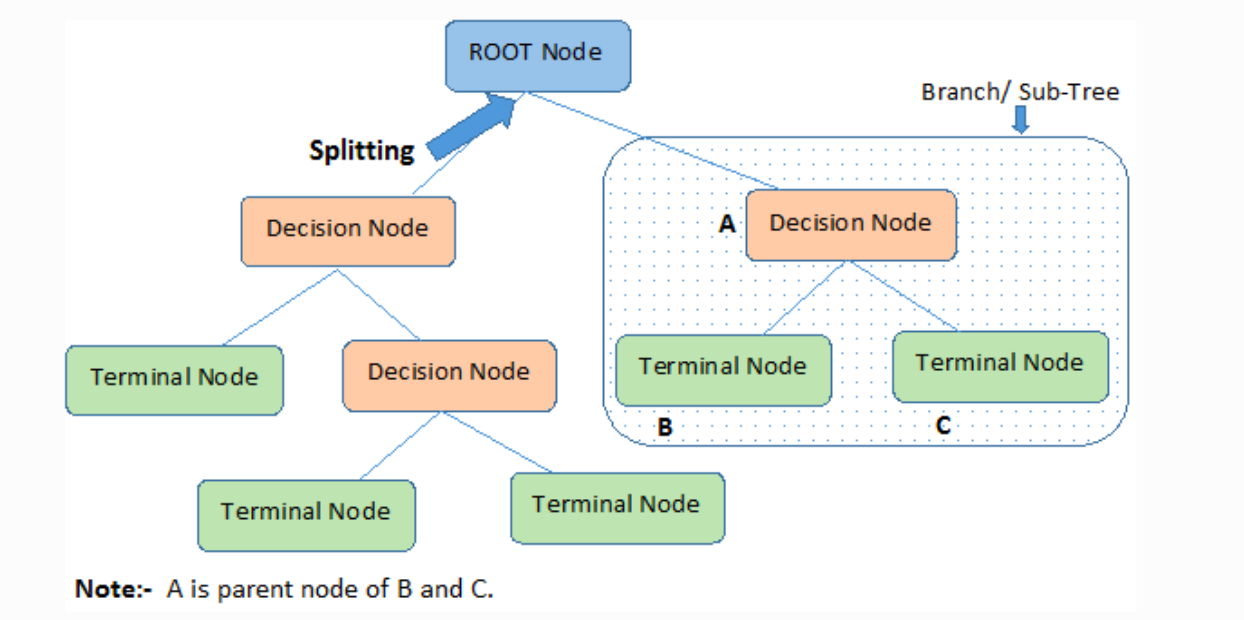
odds = e^(b0 + b1 * X)

All of this helps us understand that indeed the model is still a linear combination of the inputs, but that this linear combination relates to the log-odds of the default class.

**Decision Tree Algorithm**

Decisin Tre algrithm belong to the family of supervised learning algorithms. Unlike other supervised learning algorithms. The decision tre agoithm can be used for solving regression and classification problems too.

The goal of using a Decisn Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data.
In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.




While making decision tree, at each node of tree we ask different type of questions. Based on the asked question we will calculate the information gain corresponding to it.

**Information Gain**

Information gain is used to decide which feature to split on at each step in building the tree. Simplicity is best, so we want to keep our tree small. To do so, at each step we should choose the split that results in the purest daughter nodes. A commonly used measure of purity is called information. For each node of the tree, the information value measures how much information a feature gives us about the class. The split with the highest information gain will be taken as the first split and the process will continue until all children nodes are pure, or until the information gain is 0.

**Deep learning**

Deep learning is a sub-branch of the machine learning field, inspired by the structure of the brain. Deep learning techniques used in recent years continue to show an impressive performance in the field of medical image processing, as in many fields. By applying deep learning techniques to medical data, it is tried to draw meaningful results from medical data. Deep learning models have been used successfully in many areas such as classification, segmentation and lesion detection of medical data. Analysis of image and signal data obtained with medical imaging techniques such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT) and X-ray with the help of deep learning models. As a result of these analyzes, detection and diagnosis of diseases such as diabetes mellitus, brain tumor, skin cancer and breast cancer are provided with convenience.

A convolutional neural network (CNN) is a class of deep neural networks used in image recognition problems. Coming to how CNN works, the images given as input must be recognized by computers and converted into a format that can be processed. For this reason, images are first converted to matrix format. The system determines which image belongs to which label based on the differences in images and therefore in matrices. It learns the effects of these differences on the label during the training phase and then makes predictions for new images using them. CNN consists of three different layers that are a convolutional layer, pooling layer, and fully connected layer to perform these operations effectively. The feature extraction process takes place in both convolutional and pooling layers. On the other hand, the classification process occurs in fully connected layer. These layers are examined sequentially in the following. 3.2.1 Convolutional Layer Convolutional layer is the base layer of CNN. It is responsible for determining the features of the pattern. In this layer, the input image is passed through a filter. The values resulting from filtering consist of the feature map. This layer applies some attern to extract low- and high-level features in the pattern . The kernel is a 3x3 or 5x5 shaped matrix to be transformed with the input pattern matrix. Stride parameter is the number of steps tuned for shifting over input matrix.

The output of convolutional layer can be given as:

$$x_j^l = f\left(\sum_{a=1}^{N} w_j^{l-1} * y_a^{l-1} + b_j^l\right)$$

where xj l is the j-th feature map in layer l, wj l-1 indicates j-th kernels in layer l-1, ya l-1 represents the a-th feature map in layer l-1, bj l indicates the bias of the j-th feature map in layer l, N is number of total features in layer l-1, and (*) represents vector convolution process.

**Pooling Layer**

The second layer after the convolutional layer is the pooling layer. Pooling layer is usually applied to the created feature maps for reducing the number of feature maps and network parameters by applying corresponding mathematical computation. In this study, we used maxpooling and global average pooling. The max-pooling process selects only the maximum value by using the matrix size specified in each feature map, resulting in reduced output neurons. There is also a global average pooling layer that is only used before the fully connected layer, reducing data to a single dimension. It is connected to the fully connected layer after global average pooling layer. The other intermediate layer used is the dropout layer. The main purpose of this layer is to prevent network overfitting and divergence [44].

**Fully Connected Layer**

Fully connected layer is the last and most important layer of CNN. This layer functions like a multi-layer perceptron. Rectified Linear Unit (ReLU) activation function is commonly used on fully connected layer, while Softmax activation function is used to predict output images in the last layer of fully connected layer.

Mathematical computation of these two activation functions are as follow:

$$ReLU\ (x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

$$Soft\max(x_i) = \frac{e^{x_i}}{\sum_{y=1}^{m} e^{x_y}}$$

where xi and m represent input data and the number of classes, respectively. Neurons in a fully connected layer have full connections to all activation functions in previous layer.

**Pre-Trained Models**

In the analysis of medical data, one of the biggest difficulties faced by researchers is the limited number of available datasets. Deep learning models often need a lot of data. Labeling this data by experts is both costly and time consuming. The biggest advantage of using transfer learning method is that it allows the training of data with fewer datasets and requires less calculation costs. With the transfer learning method, which is widely used in the field of deep learning, the information gained by the pre-trained model on a large dataset is transferred to the model to be trained.

**ResNet50**

Residual neural network (ResNet) model is an improved version of convolutional neural network (CNN). ResNet adds shortcuts between layers to solve a problem. Thanks to this, it prevents the distortion that occurs as the network gets deeper and more complex. In addition, bottleneck blocks are used to make training faster in the ResNet model [45]. ResNet50 is a 50- layer network trained on the ImageNet dataset. ImageNet is an image database with more than 14 million images belonging to more than 20 thousand categories created for image recognition competitions [46].

**InceptionV3**

InceptionV3 is a kind of convolutional neural network model. It consists of numerous convolution and maximum pooling steps. In the last stage, it contains a fully connected neural network [47]. As with the ResNet50 model, the network is trained with ImageNet dataset. Inception-ResNetV2 The model consists of a deep convolutional network using the Inception-ResNetV2 architecture that was trained on the ImageNet-2012 dataset. The input to the model is a 299×299 image, and the output is a list of estimated class probabilities [48].

ResNet101 & ResNet152

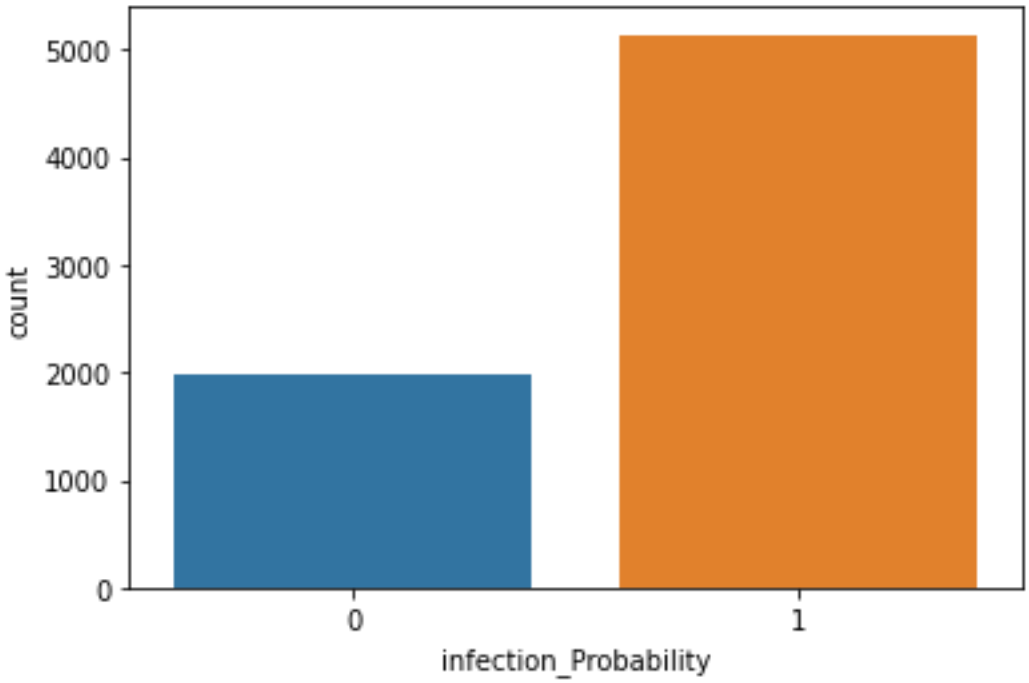
ResNet101 and ResNet152 consist of 101 and 152 layers respectively due to stacked ResNet building blocks. You can load a pretrained version of the network trained on more than a million images from the

ImageNet database [46]. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224x224.

**Experimental Setup**

Python programming language was used to train the proposed deep transfer learning models. All experiments were performed on Google Colaboratory (Colab) Linux server with the Ubuntu 16.04 operating system using the online cloud service with Central Processing Unit (CPU), Tesla K80 Graphics Processing Unit (GPU) or Tensor Processing Unit (TPU) hardware for free. CNN models (ResNet50, ResNet101, ResNet152, InceptionV3 and InceptionResNetV2) were pre-trained with random initialization weights by optimizing the cross-entropy function with adaptive moment estimation (ADAM) optimizer ($\beta 1 = 0.9$ and $\beta 2 = 0.999$). The batch size, learning rate and number of epochs were experimentally set to 3, 1e-5 and 30, respectively for all experiments. The dataset used was randomly split into two independent datasets with 80% and 20% for training and testing respectively.

**Results :**



```
lr_acc = accuracy_score(y_test,lr_pred)
lr_acc
```

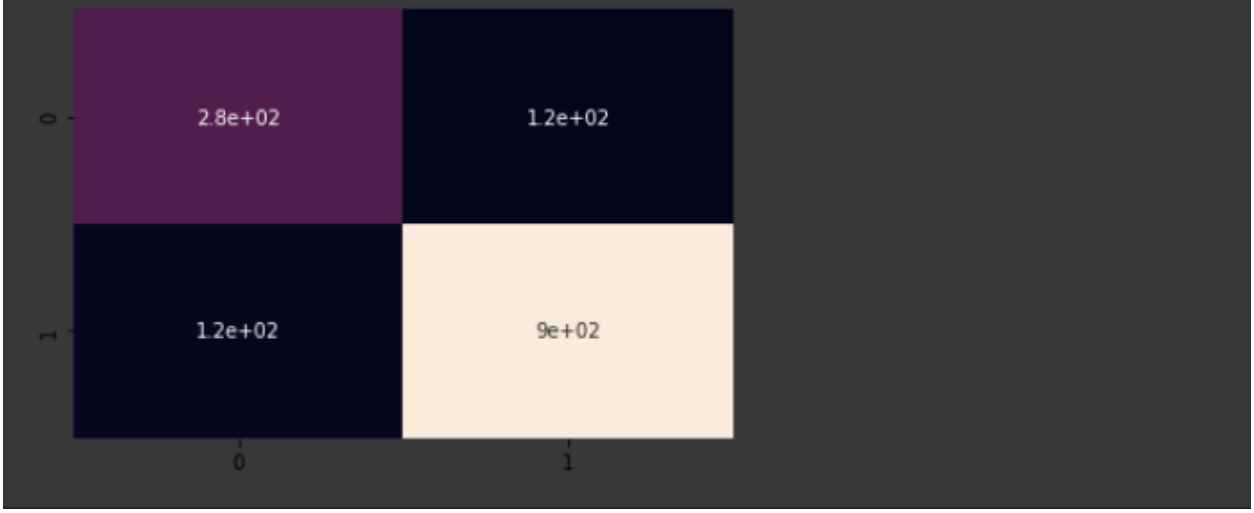
```
0.8021126760563381
```

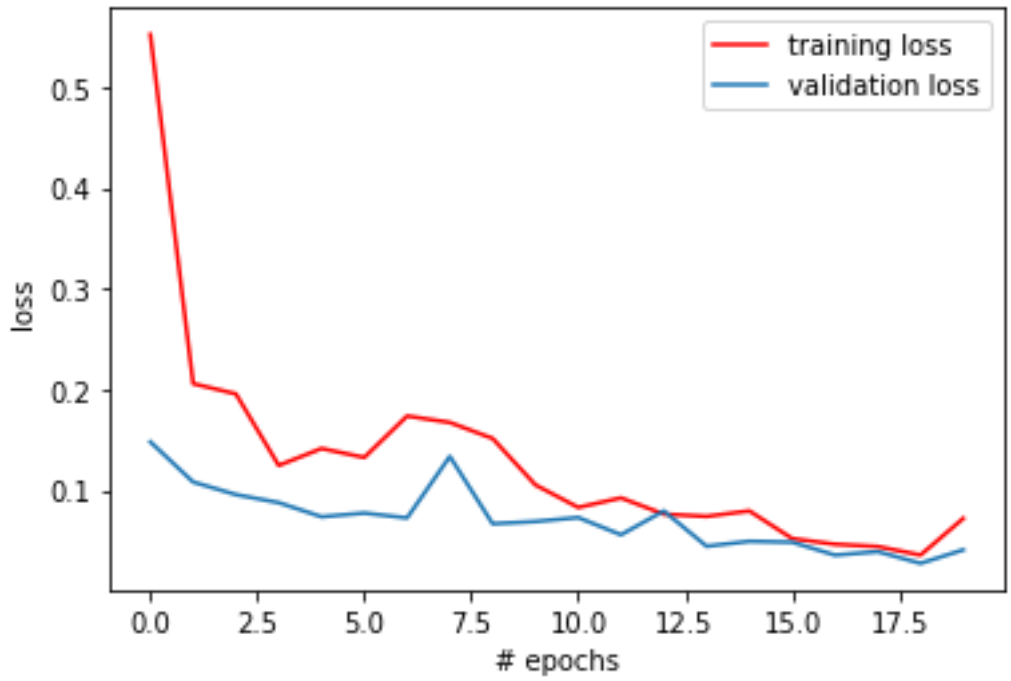
```
lr_df = pd.DataFrame(data=lr_cm,columns=['0','1'],index=['0','1'])
lr_df
sns.heatmap(lr_df,annot=True,cbar=False)
plt.show()
```

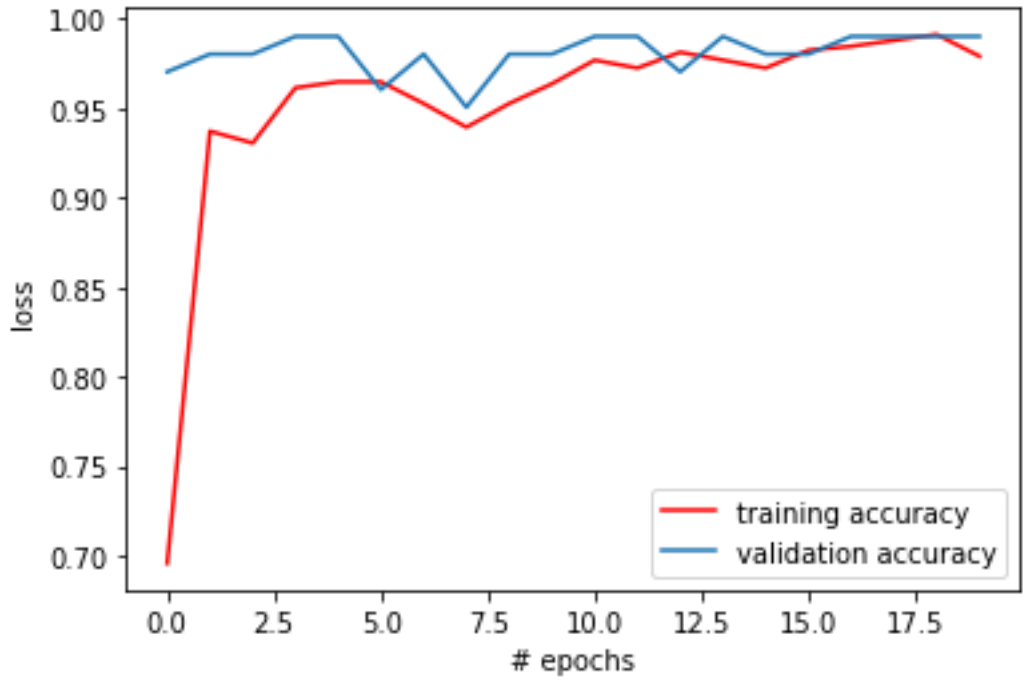

```
tr_acc = accuracy_score(y_test,tr_pred)
tr_acc
```

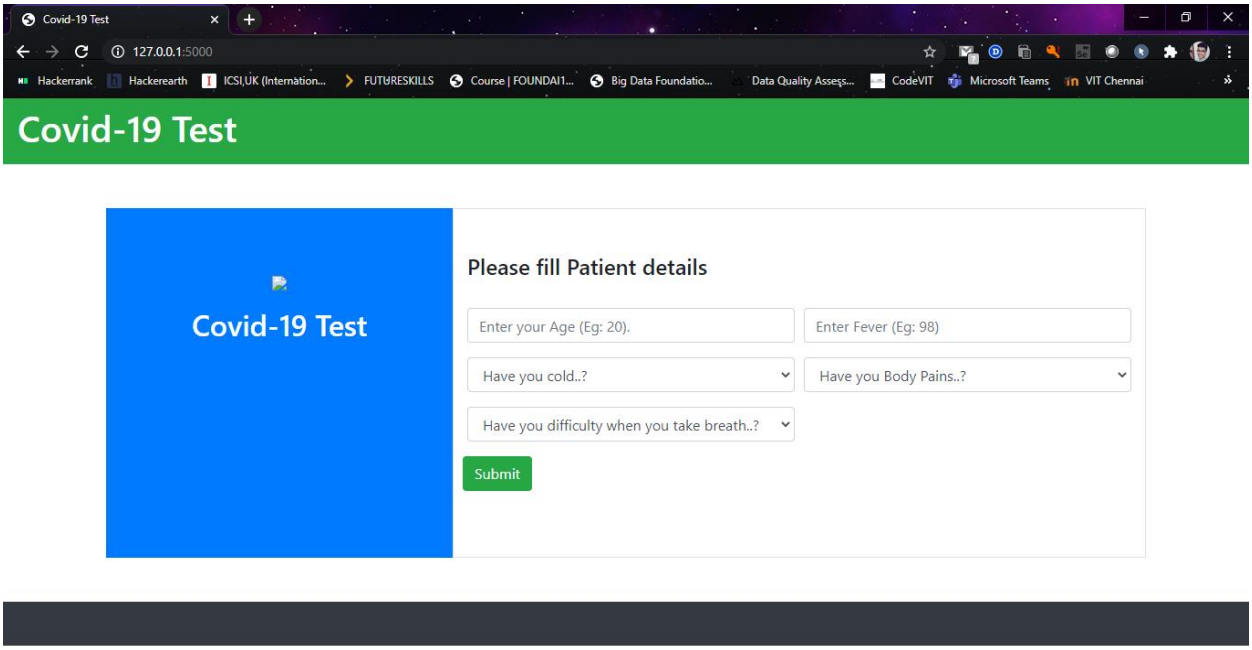
```
0.8309859154929577
```

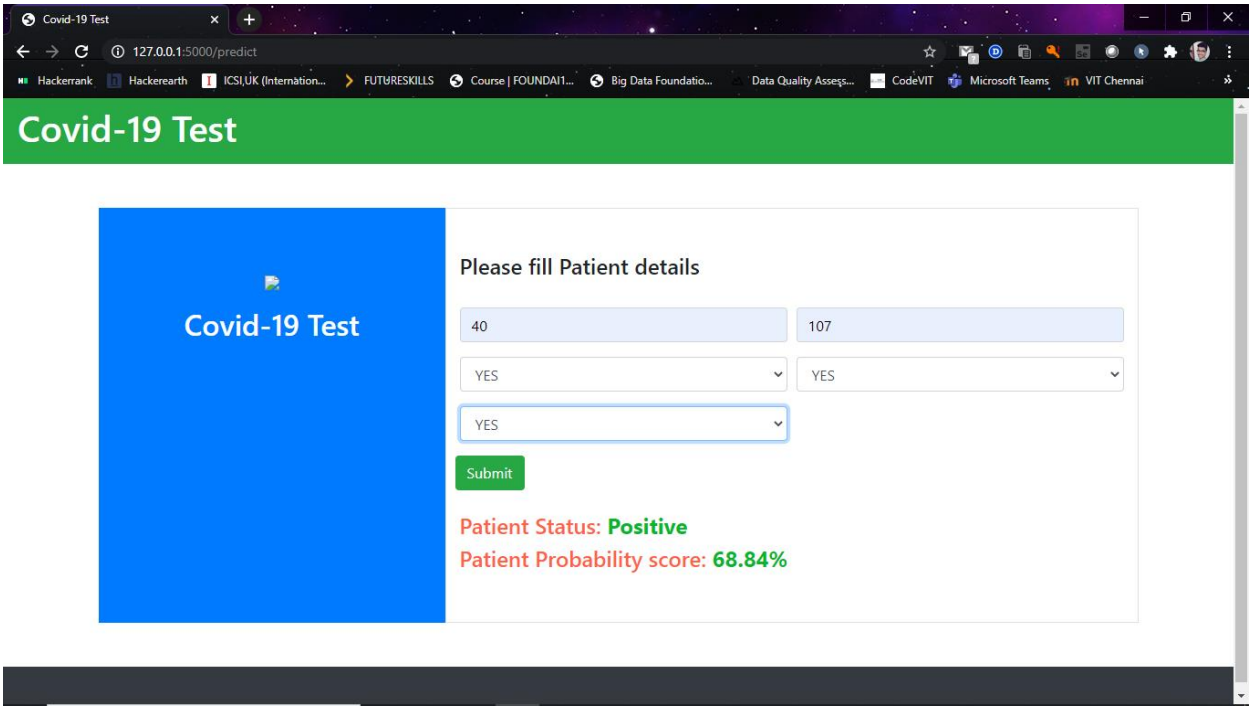
```
tr_df = pd.DataFrame(data=tr_cm,columns=['0','1'],index=['0','1'])
tr_df
sns.heatmap(tr_df,annot=True,cbar=False)
plt.show()
```

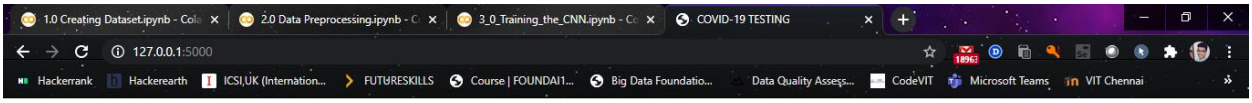



training loss
validation loss
loss
# epochs


training accuracy
validation accuracy
loss
# epochs


Covid-19 Test
127.0.0.1:5000
Hackerrank
Hackerearth
ICSI,UK (Internation...
FUTURESKILLS
Course | FOUNDAI1...
Big Data Foundatio...
Data Quality Assess...
CodeVIT
Microsoft Teams
VIT Chennai
Covid-19 Test
Covid-19 Test
Please fill Patient details
Enter your Age (Eg: 20).
Enter Fever (Eg: 98)
Have you cold..?
Have you Body Pains..?
Have you difficulty when you take breath..?
Submit


Covid-19 Test
127.0.0.1:5000/predict
Hackerrank
Hackerearth
ICSI,UK (Internation...
FUTURESKILLS
Course | FOUNDAI1...
Big Data Foundatio...
Data Quality Assess...
CodeVIT
Microsoft Teams
VIT Chennai
Covid-19 Test
Covid-19 Test
Please fill Patient details
40
107
YES
YES
YES
Submit
Patient Status: Positive
Patient Probability score: 68.84%

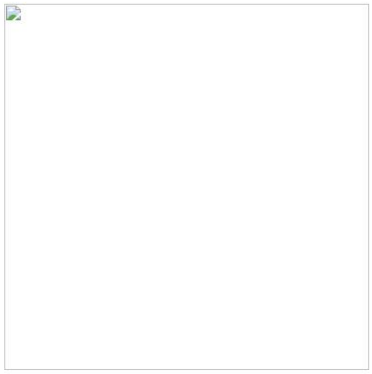


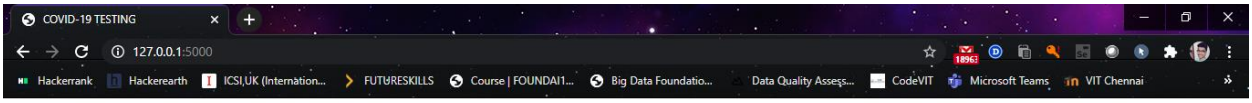

# COVID-19 TESTING USING X-RAY IMAGES

Choose File No file chosen Predict

PREDICTION: ...

PROBABILITY: ...



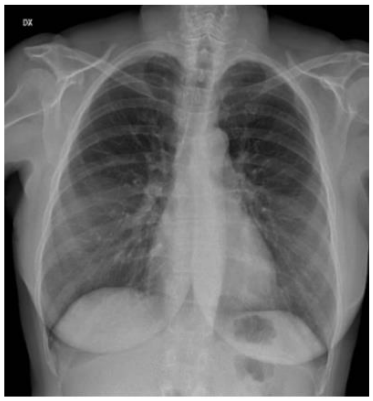


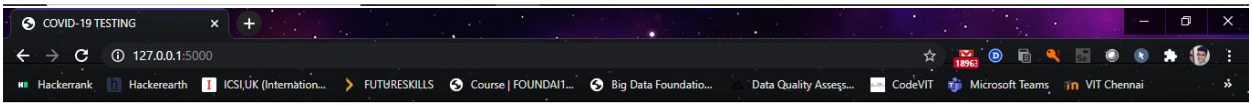

# COVID-19 TESTING USING X-RAY IMAGES

Choose File F051E018-...9E960B.jpeg Predict

PREDICTION: Covid19 Negative

PROBABILITY: 0.74



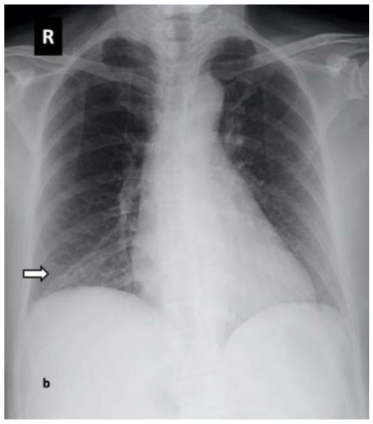



# COVID-19 TESTING USING X-RAY IMAGES

Choose File 1.CXRCTTh...01-fig2b.png Predict

PREDICTION: Covid19 Positive

PROBABILITY: 1.00



## Conclusion

Early prediction of COVID-19 patients is vital to prevent the spread of the disease to other people. In this study, we proposed a deep transfer learning based approach using Chest X-ray images obtained from normal, COVID-19, bacterial and viral pneumonia patients to predict COVID-19 patients automatically. In the light of our findings, it is believed that it will help radiologists to make decisions in clinical practice due to the higher performance. In order to detect COVID-19 at an early stage, this study gives insight on how deep transfer learning methods can be used. In subsequent studies, the classification performance of different CNN models can be tested by increasing the number of COVID-19 Chest X-ray images in the dataset

## References :


Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications*, *68*(16).

Austin, P. C., Tu, J. V., & Lee, D. S. (2010). Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *Journal of clinical epidemiology*, *63*(10), 1145-1155.

Wright, R. E. (1995). Logistic regression.

Jin, C., De-Lin, L., & Fen-Xiang, M. (2009, July). An improved ID3 decision tree algorithm. In *2009 4th International Conference on Computer Science & Education* (pp. 127-130). IEEE.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

M. Chowdhury, T. Rahman, A. Khandakar, *et al.***Can AI help in screening viral and COVID-19 pneumonia?**
arXiv preprint arXiv:2003.13145 (2020)

Malki, Z., Atlam, E. S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., & Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, Solitons & Fractals, 138, 110137.

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188.

Jaiswal, Amit Kumar, et al. "Identifying pneumonia in chest X-rays: A deep learning approach." Measurement 145 (2019): 511-518.

Lieberman, David, et al. "Diagnosis of ambulatory community-acquired pneumonia." Scandinavian journal of primary health care 21.1 (2003): 57-60.

Liang, G., & Zheng, L. (2020). A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Computer methods and programs in biomedicine, 187, 104964.

Niederman, M. S., et al. "Guidelines for the initial management of adults with community-acquired pneumonia: diagnosis, assessment of severity, and initial antimicrobial therapy. American Thoracic Society. Medical Section of the American Lung Association." The American review of respiratory disease 148.5 (1993): 1418-1426.

Luo, J., Rizvi, H., Preeshagul, I. R., Egger, J. V., Hoyos, D., Bandlamudi, C., ... & Chaft, J. E. (2020). COVID-19 in patients with lung cancer. Annals of Oncology, 31(10), 1386-1396.

Chacón-Aguilar, R., Osorio-Cámara, J. M., Sanjurjo-Jimenez, I., González-González, C., López-Carnero, J., & Pérez-Moneo, B. (2020, April). COVID-19: fever syndrome and neurological symptoms in a neonate. In Anales de pediatria. Elsevier.

COVID, T. C. (2020). Characteristics of Health Care Personnel with COVID-19-United States, February 12-April 9, 2020.

Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., & Cheng, X. (2020). Artificial intelligence and machine learning to fight COVID-19.

Jnr, B. A. (2020). Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. Journal of Medical Systems, 44(7), 1-9.

Li, N., & Yu, X. (2020). Outbreak and regression of covid-19 epidemic among chinese medical staff. Risk Management and Healthcare Policy, 13, 1095.

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Acharya, U. R. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in Biology and Medicine, 103792.