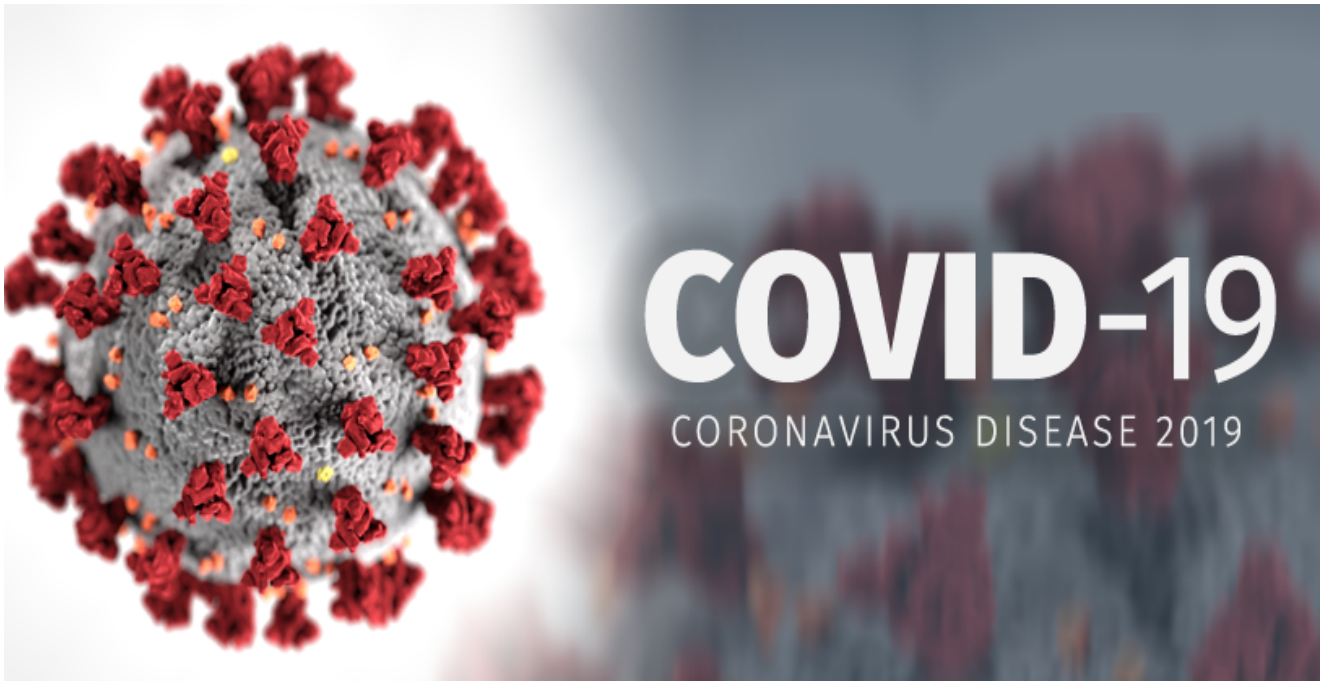
Capstone Project Report

Prediction of Covid-19 using Chest images



Mritunjay Ashish, MT20102, IIITD
Naveen, MT20078, IIITD
Shubham Sharma, MT20315, IIITD
Sumit Patiyal, PhD Scholar, CCB, IIITD
Dr. G.P.S. Raghava, Professor & HOD CCB, IIITD

Abstract ----------
Globally, the new coronavirus illness 2019 (COVID-19) represents a public health emergency. Every day, the number of sick persons and deaths grows, putting immense strain on our social and healthcare systems. COVID-19 cases must be diagnosed quickly to battle the virus and relieve the burden on the healthcare system.
The chest X-ray was the first imaging procedure that was used in the treatment of COVID-19. The use of radiological imaging to emphasise the performance of chest X-rays is common. COVID-19 has been found in patients with abnormal results on chest X-rays, according to new research. Machine learning algorithms for detecting COVID-19 using chest X-rays are included in several studies on this issue.
There are right now 269,993,520 affirmed cases in 222 nations and regions . The casualty rate is as yet being evaluated. Many researchers have worked on covid prediction in recent years, employing various machine learning techniques such as Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbour (KNN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Xgboost (XGB), Naive Bayes, and others. In this paper, we describe a model that combines data preprocessing and scaling approaches to clean the dataset, different classifiers applied to this dataset, and obtained the highest accuracy of 99.86 percent and Roc AUC of 99.86 percent by applying K-Nearest Neighbour over the test data.

Introduction ----------

Covid is a positive single-stranded substantial RNA infection that infects both humans and animals. Covid was first documented in 1966 by Tyrell and Bynoe [1], who discovered the infections in patients with ordinary colds.
The four subfamilies of coronaviruses are alpha, beta, gamma, and delta. Gamma and delta coronaviruses are derived from pigs and birds, whereas alpha and beta coronaviruses are derived from highly developed species such as bats. The genome is around 26 to 32 kb in size. One of the seven subtypes of Covids that may infect humans, beta-coronaviruses, can cause serious illness and death, whereas alpha-coronaviruses cause asymptomatic or suggestive contaminations. SARSCoV2 is a beta-coronavirus of the B gene family that has been associated to SARS-CoV infection[2][3].
Contamination can proceed to serious infection in around 75% of patients, with dyspnoea and substantial chest symptoms consistent with pneumonia, as revealed on confirmation by processed tomography [4].

Dense communities are extremely vulnerable, and because to increased migration between China and Africa, Africa is likely the most vulnerable region. Few African governments have enough analytical constraints and specific triggers in place to cope with such occurrences. The World Health Organization has selected 13 high-priority nations (Algeria, Angola, Ivory Coast, DRC, Ethiopia, Ghana, Kenya, Mauritius, Nigeria, South Africa, Tanzania, Uganda, and Zambia) with direct or major transit linkages to China.

Subsequent studies show that people over the age of 60 are at higher risk than children, who may be less prone to pollution or, if they do, may experience milder side effects or even a silent illness[5].

Similar to previous infections, SARSCoV2 infects lung alveolar epithelial cells by receptor-mediated endocytosis utilizing the angiotensin-converting compound II (ACE2) as a section receptor[3].

Immunoglobulins, remdesivir, arbidol hydrochloride in combination with interferon atomization, oseltamivir, ritonavir and addition of oseltamivir, lopinavir in addition to ritonavir, lopinavir in addition to ritonavir, treatment of undifferentiated mesenchymal organisms, darunavir in addition to cobicist [7].

Coronavirus is the primary source of death worldwide; subsequently, Coronavirus expectation is a significant metric that, whenever done early, can save many lives. A few explorations looking at the viability of prescient information mining draws near, and other machines learning advances to figure different sicknesses have been led. Lately, a few specialists have been dealing with Coronavirus forecasts and deciding the best way to deal with expect the condition. Nonetheless, more examination and patient information from emergency clinic records become accessible as time passes. Many open wellsprings of admittance to patient information and assessments might be used to make the legitimate determination of patients and identify this illness before it becomes deadly utilizing different PC innovations. We as a whole realize that information mining and AI advancements are the best indicators of Coronavirus sickness. In our review, we have likewise utilized an assortment of AI strategies to build the exactness of our forecasts.

Literature Review ----------

S. Tabik et al[8] proposed that there is a scarcity of image datasets, and in the early stage of covid-19, there is also a deficiency of datasets. In this paper, the first and essential aim of the author is to design high-quality datasets. In this paper, the author had applied a deep learning technique called covidnet. In this paper, the authors had used the image dataset, which divides into three classes, i.e., standard images, pneumonia images, and covid images. The authors had also described their dataset named as COVIDGR dataset. The authors also said that good advice about the dataset always makes a tiny and intelligent dataset rather than an extensive dataset. A small dataset will save the computation power and give the result faster. In this paper, the author had collected the images from the radiologist. The authors also describe the cxr idea, which is images of positive covid-19 patients. The COVIDGR dataset is classified into two classes, positive and negative, and the author had used a total of 852 appearances for both levels. In the paper, the authors give the new method called COVID-SDNet, based on the CNN classifier. The feature extraction used by the author in this paper is resnet50 with imagenet. The authors had given the three levels in the new methodology, including cleaning the dataset and deep learning methods. The three levels offered by the authors are segmentation, transformation, and fusion of CNN. Finally, this paper helps to detect the disease in calm and uncompromising patients[8].

A. Mohammed et al.[9] proposed an approach called resnext+, which will be used for prediction purposes. The author had also used the Long Short Term Memory model, which will be used for the slices' dependency. There are many ways to diagnose covid-19 in which x-ray is one of them, which increases the availability of deep learning-based models. In this paper, the author had used chest images for the diagnosis of the disease. The main center of attraction of the chest ct scans is the detection of lungs. The proposed model of the paper is R-CNN, and the feature extraction used is the VGG16. According to the doctors, the infection in covid-19 and pneumonia patients' lungs is nearly similar, so transfer-learning is the best technique. The author has classified the images into three categories - pneumonia images, covid-19 infected images, and healthy images. This paper's approach is that the author first takes the CT images and passes those images through Resnext+, and the features that come out will die in the LSTM model, which gives good results.

L. Sun et al.[10] propose an adaptive selection of the features with guided deep forest, which will be used to classify covid-19 disease. Authors had applied this technique to remove the duplicacy of features. The above-proposed model contains an N random forest followed by a characteristic unit of choices. The authors proposed an approach called AFS-DF that used many machine learning techniques like support vector

machines, Random forest, Logistic regression, and neural networks. The model that the author proposes is best because it gives outstanding results than all existing methods.

W. Xie, C. Jacobs[11] proposes a new approach called a relational approach (RTSU-NET). The authors had applied this approach to 470 covid-19 patients. Authors have mentioned that every human being contains five lung lobes, and authors had worked on the lung. The author has used two levels of CNN. This paper consists of mainly three steps - the first one is non-local NN. The second is framework resolution, which comprises many stages. The third is the relational model followed by two CNN approaches. The relational technique is very robust which takes less than one minute to train the CT scan. So, this approach helps the severe patients of covid-19 and also produces effective results.

E. -S. M. El-Kenawy[12] proposed the algorithm for the classification of covid-19. The model given by the author runs in three levels. In the first level, the extraction of features takes place by using Alexie. In the second level, the whale algorithm is applied, followed by the first step. In the last and final step, the author had applied the prediction classifiers, i.e., SVM, KNN, DT, etc., which displays meaningful conflict. The author employed two datasets to do the above-proposed model. This dataset consists of covid and non-covid CT images. The classifier approached by the authors gives excellent results. The authors have also used the 3D oculus architecture, which we can consider an effective tool of this paper. Last but not least, the author has applied three frameworks for this paper and gives good results in comparison to the previous work of this paper.

Z. Han et al.[13] proposed this paper in 2020 when there is a significant effort already done in this field. In this era, the authors have proposed a 3D-based attention model. In this paper, the author has taken 460 images classified into three classes: covid images, pneumonia images, and non-pneumonia images. The author has divided the images into three classes because the lungs' infection in covid and pneumonia is quite the same. The lungs' infection is the same, so even radiologists cannot tell the difference between them. According to the input, the authors have divided the approach into three levels. The first level is based on segmentation used to detect the infection and used for further usage. In the second level, the author has applied the slicing procedure for using 2D models. In the third level, the authors have used the 3D CNN, which will decide the precise way. In this paper, the author had mainly used the MIL method, which uses supervised learning techniques. After the MIL method, the author had used the attention structure. The combination of the above methods gives good results.

S. Rajaraman[14] proposed a method that takes the chest x-ray images as input and predicts covid as output. In the intermediate stages, the authors used different models like modularity-specific models, task-specific models, pruned models, and ensemble learning

techniques for performance evaluation. The authors used the custom and pre-trained CNN model trained on large-scale chest x-ray images concerning the deep learning models. The author's technique is primarily separated into four phases. The first stage instructs us on how to detect covid. The modularity transfer was applied in the second stage, which will fine-tune using the chest x-ray dataset pictures. The author had employed model pruning in the third phase. In the fourth and last step, the author used the ensemble classification technique this proposed method of authors to detect the covid suspects by using chest radiography.

Proposed Methodology ----------

The proposed method worked on covid image dataset and predicted whether a x-ray image is of covid patient or not.First, we have to analyze the dataset; then, we apply preprocessing on the dataset by resizing and loading the covid and non covid images; Then we apply a distance matrix between all points in the circular probe and the external circular boundary of the images. Then we split the data set into training and test data in a ratio of 80:20. After splitting the dataset, we perform 5-fold cross-validation and apply various machine learning classifiers with the grid search algorithm.

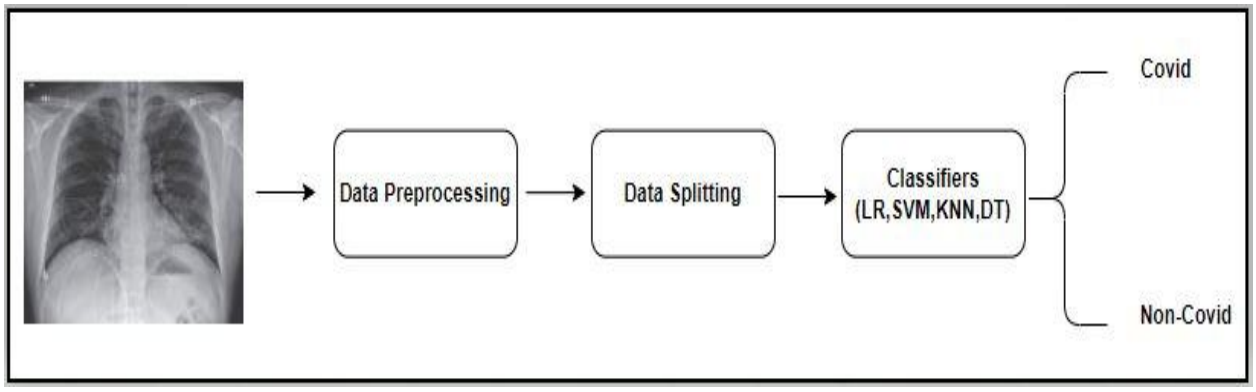


Figure 1: Framework of Proposed Approach.

(1). Dataset Description ----------

We have worked on the covid prediction dataset that we have acquired from github [15]. The dataset consists of 5184 total images in which we have taken 2084 images for training and 3100 images for Testing purposes. In training we have used 84 images for Covid and 2000 images for Non-Covid. In testing we have taken 100 images for Covid and 3000 images for Non-Covid.

Classes	Training Data	Testing Data
Covid	84	100
Non-Covid	2000	3000
Total	2184	3100

Table 1: Showing the total number of images in the train and test data.

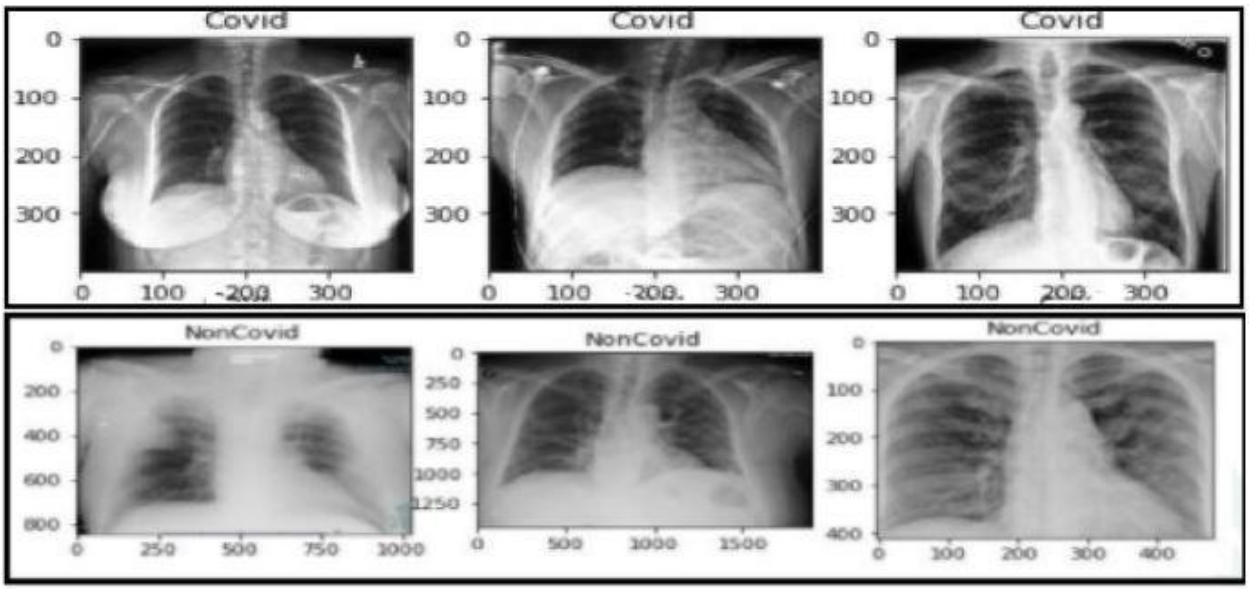


Figure 2: Covid and Non-Covid images from the dataset.

(2). Cross Validation ----------

It is a mechanism for evaluating how well our machine learning models perform on data that has never been seen before. The steps of Cross-Validation are as follows:-

1. We will keep some sample data set aside.
2. Now, we will use the remaining data to train the model.
3. Use the data-reserve set's component to test the model.

We have used 5 - fold cross-validation on the dataset.In this approach, the dataset is divided into five parts in which four components are used in the model's training process, and the remaining part is used for the testing process.

(3). Machine Learning Classifiers ----------

A machine learning classifier can be defined as it is a machine learning algorithm that assigns a classification label to data. We apply various machine learning Classifiers on the dataset like KNN, Logistic Regression, SVM, DT, and RF[17].

Support Vector Machine (SVM) - Boser, Guyon, and Vapnik proposed support Vector Machine in 1992. This classifier is highly effective on smaller datasets. This technique tries to build a decision boundary that can divide n-dimensional space into distinct classes so that fresh data points may be rapidly assigned to the appropriate category. We forecast the incidence of stroke across a predefined period in our stroke prediction issue, making it a binary classification problem that fits within the SVM framework[18].

K-Nearest Neighbour (KNN) - KNN can be defined as it Classifies unknown objects based on similarity/distance of the annotated object. It Searches similar things in a database of known objects. KNN algorithms classify a new example by comparing it to previous models. For forecasting the categorization of the present example, the classification of the k most comparable previous standards applied.

Logistic Regression (LR) - A supervised machine learning technique used to address classification issues is logistic regression. True or false, Spam or not Spam, and other discrete values are predicted using classification issues. The sigmoid function is used to model the data in logistic regression.

Decision Tree (DT) - The decision tree is a supervised machine learning approach for segmenting data based on specific features. It divides the dataset into smaller and smaller subgroups as it develops the tree. There are two sorts of nodes in the tree: decision nodes and leaf nodes. In a decision tree, decision nodes indicate outcomes, while leaf nodes are where the data is divided.

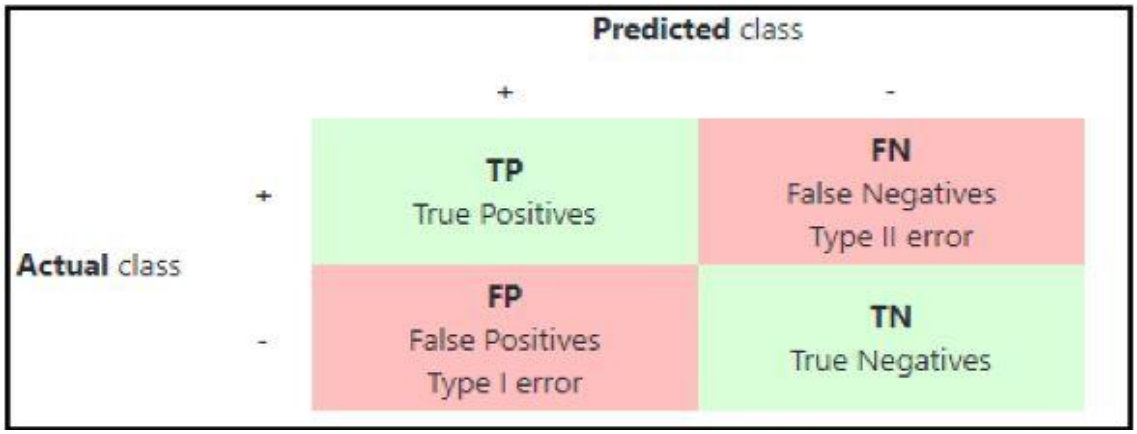
Random Forest (RF) - A supervised machine learning method is used to create the random forest. Random forest is a versatile and user-friendly algorithm. Random forest is a method that, in most cases, produces good results without the use of hyper-parameter tweaking. It creates a forest out of a collection of decision trees. It is mainly used for classification problems.

Evaluation Parameters ----------

The confusion matrix, accuracy, (AUC-ROC), precision, Recall, f1-score, Matthews correlation coefficient (MCC) score, and Cohen kappa score were all used to make the assessment. The confusion matrix is basically a NXN matrix which generally divides in four parts.

The first is a true positive (TP), in which items are designated as true and this is also true. False positive (FP) is the second category, in which the values are specified as false but are really defined as true. The third category is false negative (FN), which denotes a true value that was incorrectly labelled as negative. The fourth digit in the value, true negative (TN), was negative and was accurately detected as such.

The below images[16] show the confusion matrix and evaluation parameters formula.



Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall performance of model
Precision	$\frac{TP}{TP + FP}$	How accurate the positive predictions are
Recall Sensitivity	$\frac{TP}{TP + FN}$	Coverage of actual positive sample
Specificity	$\frac{TN}{TN + FP}$	Coverage of actual negative sample
F1 score	$\frac{2TP}{2TP + FP + FN}$	Hybrid metric useful for unbalanced classes

Result and Analysis ----------

For the sake of our research, we divided our dataset 80:20 for training and testing. Using the K-Nearest Neighbor approach, we can observe that our work gets the highest accuracy of 99.86%. We employed a variety of machine learning methods in our research, including LR, KNN, SVM, DT, RF, and neural networks. We achieved an accuracy of 97.01% using Logistic Regression. We acquired an accuracy of 99.86% when we used the K-Nearest Neighbor technique. The accuracy using the SVM classifier is 98.78 %. The accuracy using the Decision tree classifier is 99.32 %.

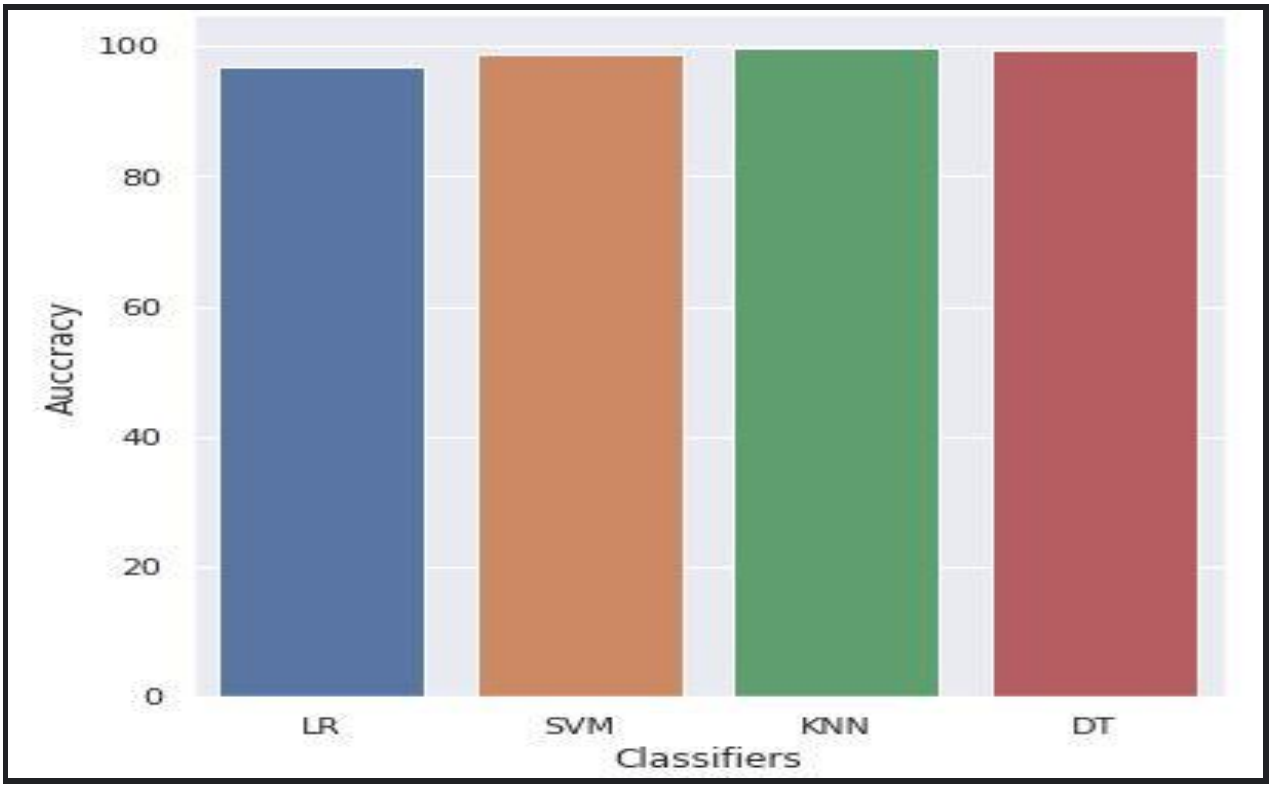
Table 2 and Table 3 showing the training and testing results on a complete dataset with cross validation.

Classifier	ACC	ROC AUC	Specificity	Sensitivity	F1-Score	MCC	Kappa
SVM	98.91	98.91	98.91	97.83	98.90	0.98	0.98
KNN	99.97	99.97	99.97	99.93	99.97	0.97	0.97
LR	97.32	97.32	97.24	94.63	97.24	0.95	0.95
DT	99.42	99.42	99.93	98.91	99.42	0.99	0.99

Table 2: Showing the training result on a complete dataset with cross validation.

Classifier	ACC	ROC AUC	Specificity	Sensitivity	F1-Score	MCC	Kappa
SVM	98.78	98.78	98.76	97.55	98.76	0.98	0.98
KNN	99.86	99.86	99.86	99.73	99.96	0.98	0.98
LR	97.01	97.01	96.92	94.02	96.92	0.94	0.94
DT	99.32	99.32	99.73	98.91	99.32	0.99	0.99

Table 3: Showing the testing result on a complete dataset with cross validation.



Conclusion and Future Scope ----------

For this Research, we applied the concept of machine learning to build a system for the prediction of coronavirus. We attained an accuracy of 99.86 percent and an auc roc score of 99.86 percent using a K-Nearest Neighbour classifier. In the future we could implement this system on the smartphone in which people can upload a picture of the chest x-ray and send it to the server. The server will automatically identify and classify the type of diseases (covid, pneumonia etc.) and send results along with prescribed medicine back to the smartphone.

Acknowledgement ----------

We, Mritunjay Ashish(MT20102), Naveen(MT20078), and Shubham Sharma(20315), are thankful to Mr. Sumeet Patiyal and Dr. G.P.S.Raghava for their guidance and support throughout the project timeline.

References ----------

[1]. Tyrrell DA, Bynoe ML. Cultivation of viruses from a high proportion of patients with colds. *Lancet* 1966: 1: 76–77.

[2]. GISAID Global Initiative on Sharing All Influenza Data . Phylogeny of SARS‐like beta coronaviruses including novel coronavirus (nCoV). (Available from: https://nextstrain.org/groups/blab/sars-like-cov).

[3]. Zhou P, Yang XL, Wang XG et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020. 10.1038/s41586-020-2012-7

[4]. Guan W, Ni Z, Yu H, et al.Clinical characteristics of 2019 novel coronavirus infection in China. medRxiv preprint posted online on Feb. 9, 2020; 10.1101/2020.02.06.20020974.

[5]. Li Q, Guan X, Wu P et al. Early transmission dynamics in Wuhan, China, of novel coronavirus‐infected pneumonia. *N Engl J Med* 2020. 10.1056/NEJMoa2001316

[6]. Richardson P, Griffin I, Tucker C et al. Baricitinib as potential treatment for 2019‐nCoV acute respiratory disease. *Lancet* 2020. S0140-6736(20)30304-4. 10.1016/S0140-6736(20)30304-4

[7].Availablefrom:https://clinicaltrials.gov/ct2/results?cond=2019nCoV&term=&cntry=&state=&city=&dist

[8]. S. Tabik et al., "COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images," in IEEE Journal of Biomedical and Health Informatics,vol.24, no. 12, pp. 3595-3605, Dec. 2020, doi:10.1109/JBHI.2020.3037127.

[9]. A. Mohammed et al., "Weakly-Supervised Network for Detection of COVID-19 in Chest CT Scans," in IEEE Access, vol. 8, pp. 155987-156000, 2020, doi: 10.1109/ACCESS.2020.3018498.

[10]. L. Sun et al., "Adaptive Feature Selection Guided Deep Forest for COVID-19 Classification With Chest CT," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2798-2805, Oct. 2020, doi:

10.1109/JBHI.2020.3019505.

[11]. W. Xie, C. Jacobs, J. -P. Charbonnier and B. van Ginneken, "Relational Modeling for Robust and Efficient Pulmonary Lobe Segmentation in CT Scans," in IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2664-2675, Aug. 2020, doi:10.1109/TMI.2020.2995108.

[12]. E. -S. M. El-Kenawy, A. Ibrahim, S. Mirjalili, M. M. Eid and S. E. Hussein, "Novel Feature Selection and Voting Classifier Algorithms for COVID-19 Classification in CT Images," in IEEE Access, vol. 8, pp. 179317-179335, 2020, doi: 10.1109/ACCESS.2020.3028012.

[13]. Z. Han et al., "Accurate Screening of COVID-19 Using Attention-Based Deep 3D Multiple Instance Learning," in IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2584-2594, Aug. 2020, doi: 10.1109/TMI.2020.2996256.

[14]. S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio and S. K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays," in IEEE Access, vol. 8, pp. 115041-115050, 2020, doi: 10.1109/ACCESS.2020.3003810.

[15]. https://github.com/shervinmin/DeepCovid/tree/master/data

[16]. https://learnanalyticshere.wordpress.com/

[17]. S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[18] R. S. Jeena and S. Kumar, "Stroke prediction using SVM," 2016 InternationalConference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kumaracoil, India, 2016, pp. 600-602, doi:10.1109/ICCICCT.2016.7988020.