Indian Journal of Radio & Space Physics
Vol. 50, March 2021, pp. 19-24

Accurate prognosis of Covid-19 using CT scan images with deep learning model and machine learning classifiers

Siddharth Gupta[a*], Palak Aggarwal[a], Nisha Chaubey[a], & Avnish Panwar[b]

[a]Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248 001,India
[b]Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun 248 001, India

Received: 12 February 2021; Accepted: 5 March 2021

The Covid-19 disease is caused by coronavirus or SARS-CoV-2 has wrecked havoc globally. This epidemic severely impacted the economy of most of the countries across the world and has taken away many lives. To control the pandemic situation many researchers, organizations, and institutes have come up with the pathogenesis and developing vaccines to decimate this disease. Out of the several techniques, one of the techniques use image patterns on Computed Tomography (CT) to detect whether a patient is Covid-19 positive or not. In this work, the SARS-COV-2 dataset has been used for the detection of Covid-19 images and normal images. These dataset images have been fed to various deep learning models for extracting the features and finally passed to various ML classifiers which classify the images as Covid-19 or normal images. The results have established that the VGG19 model along with Logistic Regression (LR) classifier gives the maximum AUC and accuracy of 98.5% and 94.6%.

Keywords: Machine Learning, Deep Learning, Coronavirus, Logistic regression (LR), Convolution neural network (CNN).

1 Introduction

In December 2019 many cases of pneumonia were registered in Wuhan, China. Later on the diagnosis and sampling analysis revealed the novel coronavirus as causing agent which is nowadays commonly called as COVID-19[1]. This virus belongs to beta-cov group 2 B and generally caused by SARS-COV-2 infection. The outbreak of coronavirus in healthcare workers revealed its tendency to spread via human-to-human transmission[2]. With the passing of time, the Covid-19 virus has spread across the globe and has taken away many lives across the world. As per the reports till 22nd October 2020, the total number of cases registered
across the world for Covid-19 are 41,351,673 and 1,133,415 people lost their lives in 217 countries[3]. Table 1 shows the total cases, active cases, and total death cases due to Covid-19 in the top 10 countries till October 2020[3]. By observing the tremendous increase in the number of patients for Covid-19, World Health Organization (WHO) declared it as "Worldwide Pandemic"[4]. The common symptoms for Covid-19 include cold, cough, fever, respiratory problems and shortness of breath (Dyspnea)[5]. During the mid of 2020, the condition across many countries of the world becomes worst and the number of people reported with positive Covid-19 drastically increases. With such a big number of cases, even the countries with the world's best medical facilities come to an end[6]. Many hospitals lack the shortage of beds and the whole system collapsed. During the peak time techniques such as Computed Tomography (CT)[7] and radiography techniques were opted to look at the chest images to figure out Covid-19 presence in real time as initially swab based tests were taking too much time.

Despite several attempts the drastic increase in Covid-19 patients and lack of radiologists resulted in very slow and few testing. Figure 1 shows the chest x-ray images of SARS-CoV-2 infection and without infection.

In order to increase the diagnosis of Covid-19 computer aided lung CT diagnosis techniques were adopted. Machine Learning (ML)[8] and Deep Learning (DL)[9] plays a major role in detection and diagnosis of coronavirus chest x-ray images. According to the research carried by Tao[10] used Reverse Transcription Polymerase Chain Reaction (RT-PCR) technique that uses the swab samples to diagnose the Covid-19 positive patient diagnosis requires a lot of time but by using the chest CT scan detection of Covid-19 viral infection becomes easy. Convolution Neural Network

*Corresponding author
(E-mail: sidgupta307@gmail.com)

Table 1 — COVID-19 Data for 10 countries across the world

Country Name	Covid-19 cases across World		
	Total Cases	*Active Cases*	*Death Cases*
USA	8,549,673	2,765,828	226,719
India	7,705,158	716,607	116,653
Brazil	5,276,942	400,443	154,906
Russia	1,447,335	N/A	24,952
Spain	1,046,641	162,252	34,366
Argentina	1,018,999	68,136	27,100
Colombia	974,139	815,721	29,272
France	957,421	51,749	34,048
Peru	874,118	146,237	33,875
Mexico	860,714	N/A	86,893

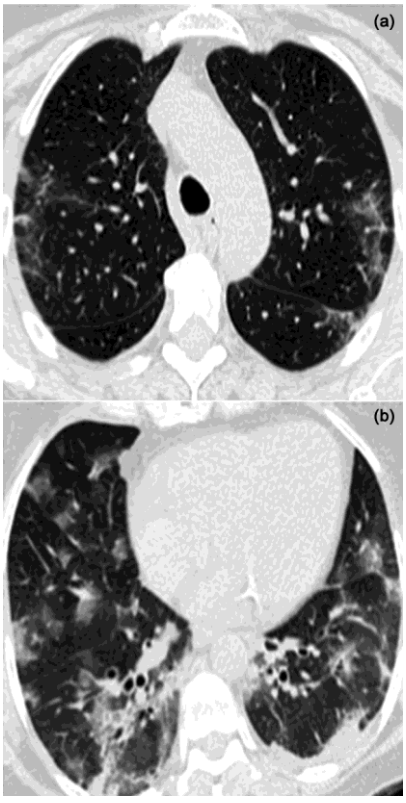

Fig. 1 — X-ray images: (a) with SARS-CoV-2 infection, and (b) without SARS-CoV-2 infection.

(CNN) proves to be a much powerful technique with the high rate of precision for the classification of Covid-19 using the chest x-ray or CT scan images. In this work, we used different CNN DL based model and ML based classifiers to classify the chest x-ray images as that of Covid-19 positive case or normal image.

The following works was carried out by several researcher: Gifani *et al.*[11] proposed a new technique established on an ensemble of deep learning for automatic diagnosis of Covid-19 infection. Several CNN architectures such as Efficient Nets (B0-B5), Inception V3, ResNet-50, Xception, DenseNet-121 were used. However, the results show an accuracy of 85% was obtained using five different deep transfer learning architectures. Karar *et al.*[12] proposed a new framework for automated computer aid diagnosis of Covid-19 disease using chest x-ray scans. The authors have used three different deep learning models such as ResNet-50v2, DenseNet169, and VGG16 for extracting the features and classifying the images. The results proved that all three models were accurately classified in the bacterial and viral infection. Wu *et al.*[13] evolved a deep learning method to assist the radiologist in the early diagnosis of Covid-19 using CT scans. The dataset used consisted of 495 CT scan images that were collected from three different hospitals located in China. ResNet50 architecture was used as the base model for the feature extraction. The obtained result shows an AUC of 73.2%, accuracy of 70%, the sensitivity of 73%, and specificity of 61.5% was obtained. Zebin *et al.*[14] used two publically available datasets of chest x-ray images for detection of Covid-19, pneumonia from the normal scan. The authors used a transfer learning pipeline for the diagnosis of Covid-19. Also, GAN networks were used to augment the Covid-19 class. An accuracy of 90% was achieved. Ouyang *et al.*[15] collected 2186 CT scans of 1588 patients who are suffering from Covid-19. These images were provided as an input to deep learning-based models. The results show an AUC of 94.4%, accuracy of 87.5%, sensitivity of 86.9%, specificity of 90.1%, and f1 score of 82%.

2 Materials and Methods

2.1 Dataset collection

SARS-COV-2 dataset[16] comprised of 2482 CT scan images out of which 1252 images were of patients infected by SARS-COV-2 infection and the remaining 1230 images were of the patients diagnosed with negative SARS-COV-2 infection but having other pulmonary diseases. The dataset images were gathered from Sao Paulo hospital situated in Brazil. Figure 2 shows the sample images of CT scan of patients suffering from SARS-COV-2 infection and those of non SARS-COV-2 images.

2.2 Pre-processing the dataset

For the ease of classification technique on the image dataset, all the images should be in the same resolution with equal width to height ratio. The images presented in

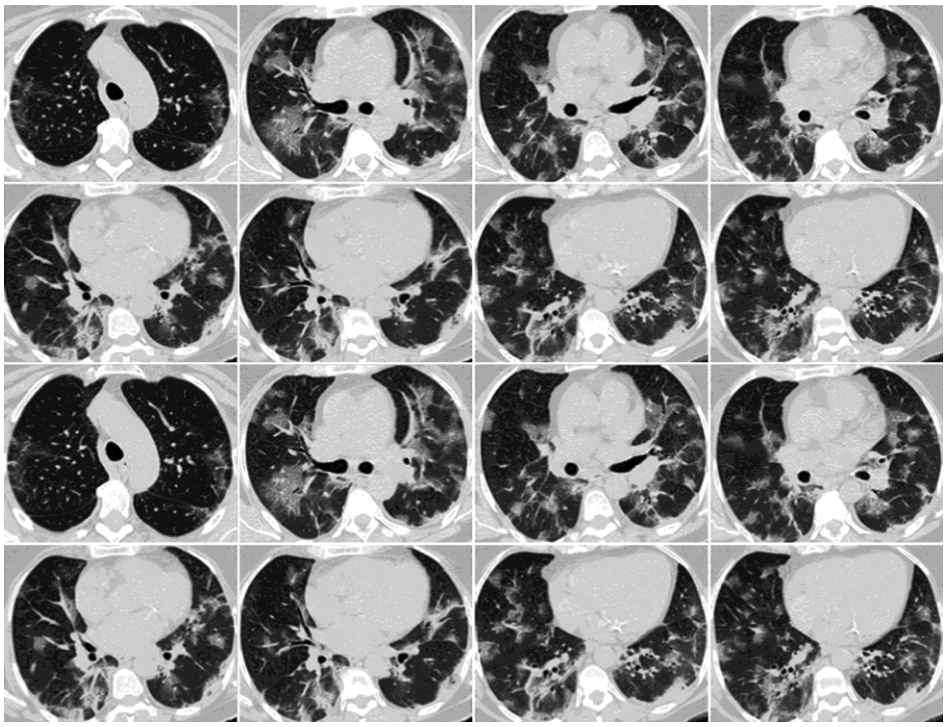

Fig. 2 — X-ray images for patient suffering and not suffering from SARS-CoV-2 infection.

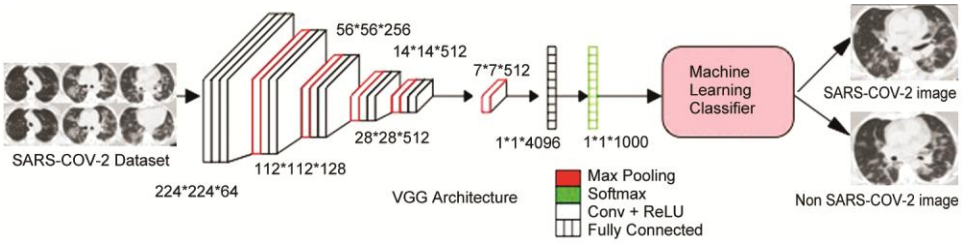

Fig. 3 — Overall architecture.

the SARS-COV-2 dataset varied in shape and size due to which classification were not possible. To overcome this problem, the image pre-processing techniques are applied to the dataset. After applying image cropping and image resizing technique the obtained images were of the same resolution.

2.3 Architecture

For the purpose of accurate prediction of Covid-19 positive patients, the chest x-ray or CT scans were used to feed on CNN[17] based deep learning models such as VGG16[18] and VGG19[18] and several ML classifiers such as KNN[19], Random Forest (RF)[20], Naïve Bayes[21], Logistic Regression (LR)[22] and AdaBoost[23] were applied to classify the images as Covid-19 or normal images. Figure 3 predicts the detailed framework for the classification of Covid-19 images. The detailed description about the architecture is explained as:

- The SARS-COV-2 dataset images of Covid-19 patient and normal persons were pre-processed and converted to the same shape, size, and resolution.
- After pre-processing, the dataset images were fed to various CNN models for feature extraction. VGG16 and VGG19 models were trained for the classification of SARS-COV-2 infection images and normal images. VGG network introduced by Simonyan and Zisserman. In this architecture several layers were present, convolutional layers were stacked over each other. Max pooling layer tackles the reduction in volume size. Finally, the two fully connected layers that constitute 4096 nodes were followed by the Softmax layer. The "16" and "19" referred to the number of weight layers in the network.

Table 2 — VGG16 Model along with several classifiers

Table	VGG16 Model				
	AUC	*Accuracy*	*F1 score*	*Precision*	*Recall*
KNN	0.982	0.942	0.942	0.942	0.942
RF	0.928	0.847	0.847	0.847	0.847
Naive	0.839	0.788	0.788	0.792	0.788
LR	0.986	0.942	0.942	0.942	0.942
Ada	0.779	0.779	0.779	0.779	0.779

- After pre-processing the images and extracting the features from the images, these were passed to several ML classifiers. These classifiers classified the images as of Covid-19 positive or normal images. For this purpose, we integrated KNN, RF, Naïve, LR, and AdaBoost classifiers.

2.4 Evaluation parameters

To evaluate the efficiency of the several classifiers used for classification of SARS-COV-2 dataset images, different parameters such as Area Under the Curve (AUC)[24], accuracy[24], precision[24], recall[24] and F1 score[24] were used. The accuracy might be as the ability to classify the image correctly as Covid-19 out of the total images. Similarly, precision may be defined as accurate predictions with respect to the overall predictions. These parameters can be calculated by using the equations:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \ldots(1)$$

$$Precision = \frac{TP}{TP+FP} \quad \ldots(2)$$

$$Recall = \frac{TP}{TP+FP} \quad \ldots(3)$$

$$F1\ Score = 2 * \frac{Precision\ +Recall}{Precision\ +Recall} \quad \ldots(4)$$

where, TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, FN stands for False Negative.

3 Results and Discussion

The SARS-COV-2 dataset chest x-ray images were used as input to the CNN based deep learning model such as VGG16 and VGG19 for feature extraction. Once the features were extracted these images are passed to several ML-based classifiers such as KNN, RF, Naïve, LR, Ada Boost for the purpose of classification of Covid-19 images, and normal images. The results obtained by using the VGG16 and VGG19 model can be seen in Table 2 and Table 3. It can be observed from Table 2. that by using the VGG16 model along with the KNN and LR classifier the accuracy obtained was 94.2% for the classification of SARS-COV-2 dataset images as Covid-19 infection and other any other infection. Table 3 shows by using the VGG19 model along with LR classifier gave 94.6% accuracy. Although the difference between the accuracies obtained in the case of VGG16 and VGG19 model were very less but still, the VGG19 model comprises 19 layers and the parameters used for training the model are 19 which was more than the 16 parameters used in the case of the VGG16 model. Hence, VGG19 was an efficient model for the classification of images as Covid-19 positive or any other infection in the case SARS-COV-2 dataset.

Table 3 — VGG19 Model along with several classifiers

Table	VGG19 Model				
	AUC	*Accuracy*	*F1 score*	*Precision*	*Recall*
KNN	0.981	0.939	0.940	0.939	0.939
RF	0.922	0.844	0.844	0.845	0.844
Naive	0.841	0.786	0.786	0.787	0.786
LR	0.985	0.946	0.946	0.946	0.946
Ada	0.777	0.777	0.779	0.777	0.777

Table 2 gives the values for VGG16 model along with the several classifiers. It can be seen that using KNN classifier the accuracy obtained was 94.2%, using RF classifier the accuracy is 84.7%, using Naïve Bayes classifier the accuracy was 78.8%, using LR classifier the accuracy was 94.2% and using AdaBoost classifier the classification accuracy was 77.9%. KNN and LR both the classifier gave the highest accuracy as compared with other classifier.

Table 3 gives the values for VGG19 model along with the several classifiers. It can be seen that using KNN classifier the accuracy obtained was 93.9%, using RF classifier the accuracy was 84.4%, using Naïve Bayes classifier the accuracy was 78.6%, using LR classifier the accuracy was 94.6% and using AdaBoost classifier the classification accuracy was 77.7%. LR classifier gives the highest classification accuracy of 94.6% as compared to the other classifiers.

2482 images were used to feed as an input to the VGG19 network for the feature extraction. Out of these 1252 CT scan images belonged to patients infected with SARS-COV-2 infection and 1230 images were of patients having other diseases. Once the features were extracted these images were passed to several ML

classifiers with which the accuracy of 94.6% was extracted. It has become very difficult for radiologists to differentiate the images with SARS-COV-2 infection from other respiratory diseases. The overall manual procedure for the classification of images was very much tedious. After analyzing the complexity of the problem, several researchers used the technique of deep learning to overcome this classification problem. By using the DL techniques the various models extracted the features from the images and these features were passed to several ML classifiers for the classification of disease in the right category. Various DL models were used for the testing purpose, and we found out of those models the VGG19 model performed exceptionally well with the accuracy of 94.6% as stated above. However, there will be scope for improving the accuracy of the classification of SARS-COV-2 images from several other diseases.

The other method to evaluate the efficiency of classifiers used was by creating the confusion matrix that gave a detailed understanding of the classification procedure. Also, ROC-AUC[25] curve was plotted to check the performance of classifiers. ROC curve was plotted between sensitivity and (100-specificity). ROC curve is used to measure the dynamics of clinical sensitivity with respect to specificity for a range of all probable cut offs in a test dataset. The ROC curve for VGG19 models with respective classifiers can be seen in Fig. 4(a) and similarly, Fig. 4(b) shows the ROC curve plotted for the VGG16 model and several classifiers[25].

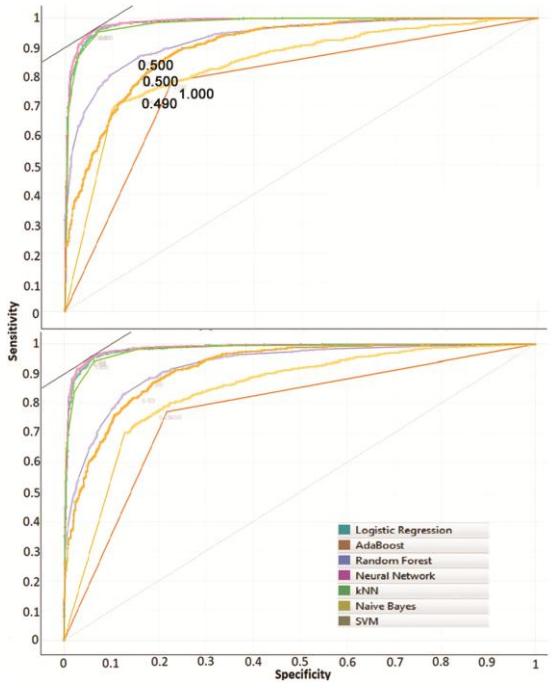

Fig. 4 — ROC curve for (a) VGG16, and (b) VGG19 models.

4 Conclusion

Adequate diagnosis of Covid-19 positive patient has been of utmost importance to provide them in time treatment and also save the others from being infected with the coronavirus. For diagnosing the Covid-19 patient the deep learning approach in this work has been used and the results have shown that using the aforementioned approach radiologist can easily observe and classify the patient as Covid-19 positive or not. In this work, the SARS-COV-2 dataset has been used on VGG16 and VGG19 model for feature extraction, and later on ML classifier has been used to classify the CT image as Covid-19 or any other disease. One of the big advantages of using a DL based approach is the less time taken for diagnosis of disease and the process of classification is simple and easy to apply. VGG19 model along with the LR classifier has given an accuracy of 94.6%. Therefore, using this approach of classification has proven to be an efficient, less time-consuming, and very much cost-effective technique for the classification of CT scan images. In the future, these accuracies can be improved by either enhancing the number of CT scan images in the dataset or apply the same dataset on several other deep learning models.

References

1 Hu S, Gao Y, Niu Z, Jiang Y, Li L, Xiao X, Wang M, Fang EF, Menpes-Smith W, Xia J, Ye H, & Yang G, *IEEE Access*, 8 (2020) 118869.
2 Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, Xing F, Liu J, Yip CC Y, Poon RW S, Tsoi HW, Lo S K F, & Chan K H, *The lancet*, 395 (2020) 514.
3 Worldmeters: Covid19 Coronavirus Pandemic. Accessed on: October 22, 2020 https://www.worldometers.info/coronavirus/
4 Santis E D, Martino A, & Rizzi A, *IEEE Access*, 8 (2020) 132527.
5 Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, & Tan W, *Jama*, 323 (2020) 1843.
6 Perry T S, *IEEE Spectrum*, 57 (2020) 4.
7 Fan D P, Zhou T, Ji G P, Zhou Y, Chen G, Fu H, Shen J, & Shao L, *IEEE Trans Med Imaging*, 39 (2020) 2626.
8 Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, & Cabitza F, *J Med Syst*, 44 (2020) 1.
9 Jamshidi M, Lalbakhsh A, Talla J, Peroutka Z, Hadjilooei F, Lalbakhsh P, Jamshidi M, Spada L L, Mirmozafari M, Dehghani M, Sabet A, Roshani S, Roshani S, Bayat M N, Mohamadzade B, Malek Z, Jamshidi A, Kiani S, Hashemi-Dezaki, & Mohyuddin A W, *IEEE Access*, 8 (2020) 109581.
10 Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, & Xia L, *Radiology*, 296 (2020) 32.

11 Shalbaf A, & Vafaeezadeh M, *Int J Comput Assist Radiol Surg*, 16 (2021) 115.
12 Karar M E, Hemdan E E, & Shouman M A, *Complex Intell Syst*, 7 (2021) 235.
13 Wu X, Hui H, Niu M, Li L, Wang L, He B, Yang X, Li L, Li H, Tian J, & Zha Y, *Eur J Radiol*, 128 (2020) 109041.
14 Zebin T, & Rezvy S, *Appl Intell*, 51 (2021) 1010.
15 Ouyang X, Huo J, Xia L, Shan F, Liu J, Mo Z, Yan F, Ding Z, Yang Q, Song B, & Shi F, *IEEE Trans Med Imaging*, 39 (2020) 2595.
16 https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset/
17 Abbas A, Abdelsamea M M, & Gaber M M, *arXiv preprint*, 51 (2021) 854.
18 Horry M J, Chakraborty S, Paul M, Ulhaq A, Pradhan B, Saha M, & Shukla N, *IEEE Access*, 8 (2020) 149808.
19 Altman N S, *The American Statistician*, 46 (1992) 175.
20 Breiman L, *Machine learning*, 45 (2001) 5.
21 David D Lewis, Naive (Bayes) at forty: The independence assumption in information retrieval, European conference on machine learning, Springer, Berlin Heidelberg, (1998) 4.
22 Hosmer D W, *Applied logistic regression*, 15 (2000) 143.
23 Hastie T, Rosset S, Zhu J, & Zou H, *Stat Interface*, 2 (2009) 349.
24 Siddharth Gupta, Avnish Panwar, Silky Goel, Ankush Mittal, Rahul Nijhawan, & Amit Kumar Singh, International Conference on Information Technology (ICIT), Bhubaneswar, India, (2019) 342.
25 Chowdhury ME, Rahman T, Khandakar A, Mazhar R, Kadir M A, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Emadi N A, & Reaz M B, *IEEE Access*, 8 (2020) 132665.