Abstract

SARS-CoV-2 virus caused COVID-19 pandemic worldwide in December 2019. The early COVID-19 cases were reported in Wuhan, Hubei Province, China in December 2019.
The World Health Organization (WHO) declared COVID-19 a pandemic on March 11, 2020. In the face of this epidemic, it was critical to reduce fatalities and prevent the epidemic from spreading. People were running out of testing kits and many other resources. Those living in remote and rural areas, in particular, faced numerous challenges due to a lack of immediate access and awareness. Even though most of the people have been immunized by now, new mutations and rising cases are always a concern. There is a clear need for the development of novel computer-assisted diagnosis tools to provide rapid and cost-effective screening in locations where massive traditional testing is not feasible.
As a result, it will be extremely beneficial if we can have parallel diagnostic and testing procedures that use artificial intelligence and machine learning while also taking advantage of historical data. It can solve the greater purpose in detecting disease at an early stage, reducing mortality risks and allowing faster medical assistance. In this age of automation, machine learning and data science play an important role in the healthcare industry. As these technologies are so interconnected, medical professionals can manage their roles and patient care with ease. Analysis of radiography images can be used to diagnose the disease.
In this paper, we leveraged the use of different AI and ML models to work on diagnosis of COVID-19 infections. Our goal was to train these models with good accuracy so that they can help in analyzing the situation at the earliest, implementing correct strategies and ensuring right diagnosis. CNN architectures VGG16 and ResNet50 were used to build the models for detecting COVID from X-Ray and CT images datasets which were taken from Kaggle. CNNs are extremely efficient at image processing. Extracted features can become hierarchically and progressively more complex as one layer feeds its output into the next layer. It has a number of building blocks, including convolution layers, pooling layers, and fully connected layers. They were trained on 100 and 150 epochs respectively. Image paths were given and the output was given in the form of predicted condition and chances of it being covid or normal.

Introduction

Coronavirus (COVID-19) public health crisis which is caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), has produced a devastating toll both in terms of human life loss and economic disruption. COVID-19 started in December 2019 in Wuhan province of China. Thereafter it spread to many countries and caused global pandemic in 2021. As of 3rd August, 2022, over 584 million cases (**584,358,167** cases) have been reported around the world (https://www.worldometers.info/coronavirus/) and over 44 million cases (44,067,144 cases) have been reported in India with deaths reported to be 526,477. To prevent the infection from spreading, most countries restricted social interaction through precautionary measures such as isolation, quarantine and social distancing. Many countries-imposed lockdowns for months to prevent the spread of the virus. Hospitals used test kits to diagnose COVID-19 disease for the treatment. But still many infected patients did not benefit from the treatment due to late diagnosis and the novel and unknown nature of the virus. Vaccines have been developed and each country has a greater number of people vaccinated. But the strains of COVID-19 are getting mutated which have put the whole world in concern.

It will be tremendously valuable if we can have concurrent diagnostic / test operations utilizing artificial intelligence and machine learning while also taking advantage of past data to train the models. It can also aid in the selection of individuals who will be examined first. On radiographic pictures, the majority of COVID-19 patients had comparable features, such as early-stage frost-glass opacities and advanced-stage pulmonary consolidation. A rounded shape and a peripheral pulmonary distribution have also been reported. This means that a high number of patients will frequently need to be examined in short time intervals by a small number of professionals with limited resources. In general, one of these three tests is used to diagnose COVID-19.

Though there are many tests available, the problems and limitations with these tests could be simplified and made more effective using computer-based models to quickly diagnose the disease and predict its future number of cases. These models use some sort of algorithms to diagnose the disease by training past data to the model used.

Despite recent progress in the COVID-19 vaccine approval and distribution, the pandemic continues to pose a huge burden to our healthcare system and to medical professionals. Global resources to manage this crisis continue to be in short supply. It remains critical to quickly and efficiently identify, screen and monitor individuals with the highest risks for COVID-19 so that distribution of medicines and treatments can be based on the level of individual risks. In this era of automation, Machine Learning (ML) and data science have an important role in the healthcare industry. These technologies are well-connected so that medical professionals can manage their roles and patient care with ease. Using these technologies, many diseases such as cancer, diabetes, HIV, and other diseases are being diagnosed. These technologies use Machine Learning algorithms to diagnose particulate diseases and also predict their future spread. Similarly, we can use Machine Learning algorithms to diagnose COVID-19 disease at an early stage and also predict its future cases of infection for better hospital treatments and measures to be taken by the government. The datasets of COVID-19 patients can be integrated and analysed by ML algorithms to improve diagnostic speed, better accuracy and potentially identify the most susceptible people based on personalized clinical and laboratory characteristics. These datasets of COVID-19 patients are different for diagnosis and for predicting its cases.

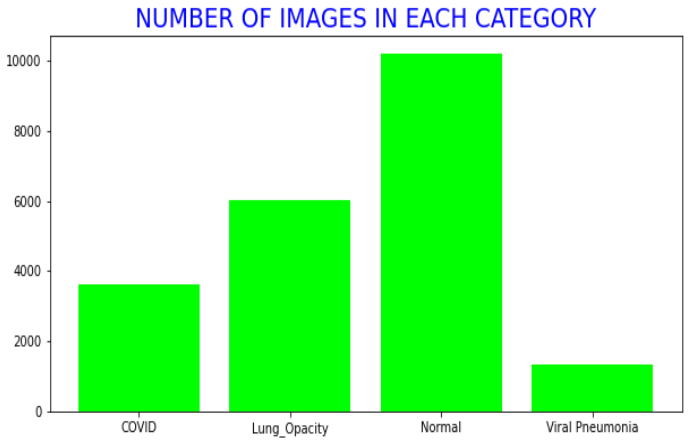
So, we decided to come up with diagnosis model for the COVID-19 virus. The proposed model would detect COVID-19 infected patients from their CT-scan images and X-Ray images using Machine Learning algorithm Convolutional Neural Network (CNN) with high accuracy and quicker diagnosis and prediction thus improving hospital capacity planning and timely treatment.

Convolutional Neural Network is nothing but Deep Learning algorithm which takes images as input, differentiate these images and gives output based on input. It is an artificial network where the connection between its neurons is inspired by the human visual cortex. It uses three layers viz., Convolution layers, Max pooling layers and fully connected layers. Two CNN models are used, VGG16 and ResNet50. VGG16 is one of convolutional neural networks and the 16 number indicates that it is 16 layers deep. ResNet50 is also one of convolutional neural networks and the 50 number indicates it is 50 layers deep.

Methodology

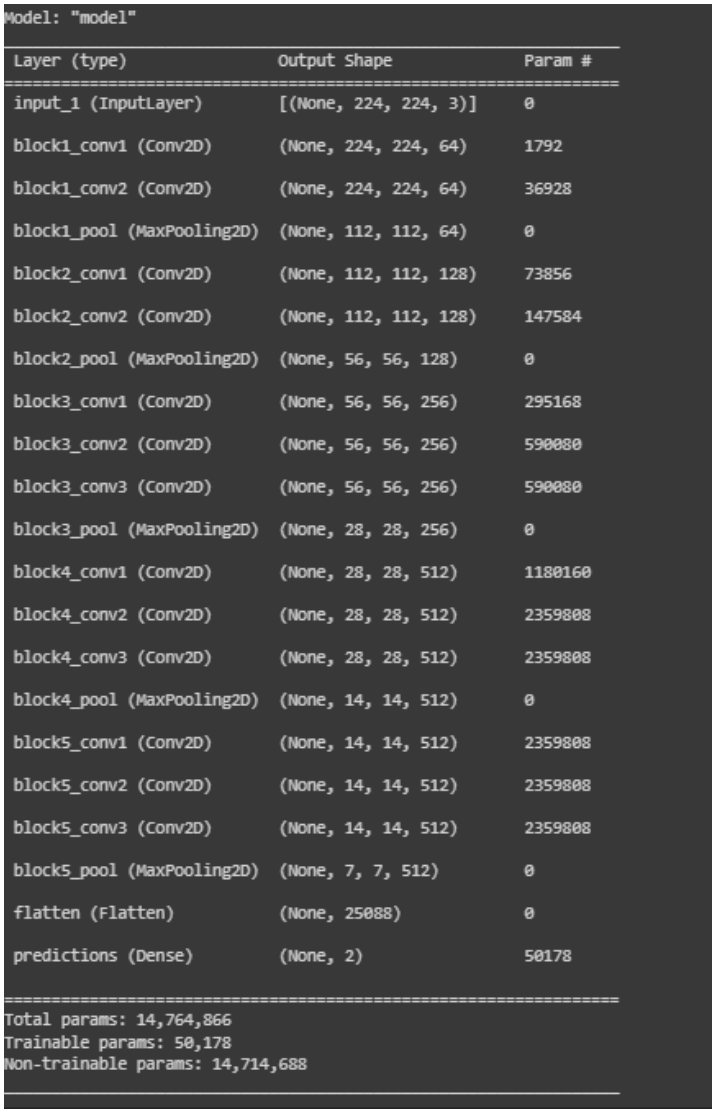
Covid Detection from X-Ray

- **Dataset Collection:** The X-Ray images dataset was collected from Kaggle, one of the most famous sites for the data science community. The dataset available at https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database was downloaded. This database of chest X-ray pictures of viral pneumonia, lung opacity, and COVID-19 positive and normal cases was established by a research team from Qatar, Doha, Bangladesh, and Dhaka universities in collaboration with their collaborators and medical doctors. The dataset contains four.csv files with the following number of photos in each category: 3616 for COVID-19, 6012 for Lung Opacity, 10192 for Normal, and 1345 for Viral Pneumonia.
- **Reading metadata and visualizing number of images:** All the required dependencies were imported. .csv files of each category were read by giving the respective file paths. The columns contain the 'filename', 'image format', 'image size' and 'URL of the image source'. The number of images in each category were visualized using *for* loop and count parameter with an initialization of 0 (Figure 1). As each category of images had a separate folder, all the images were transferred to a new single directory to split the images without any bias using train_test_split ().



Bar graph showing number of images in each category

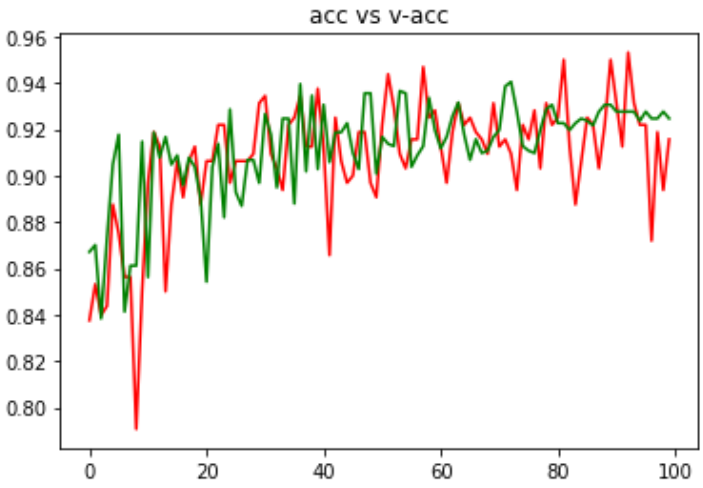
- **Data Augmentation and Model Training:** Input images were read and for data augmentation, the Keras ImageDataGenerator class was utilized, with input parameters such as zoom range, horizontal flip, shear range, and rescale. It employs several techniques such as standardization, rotation, shifts, flips, and brightness changes, among others. The ImageDataGenerator class ensures that the model receives new versions of the photos at each epoch, with the added benefit of using less memory. The VGG16 model was then defined (Figure 2) and the algorithm was run at 100 epochs with learning rate as "1e-4" (Figure 3). This code was written in google colab interface.



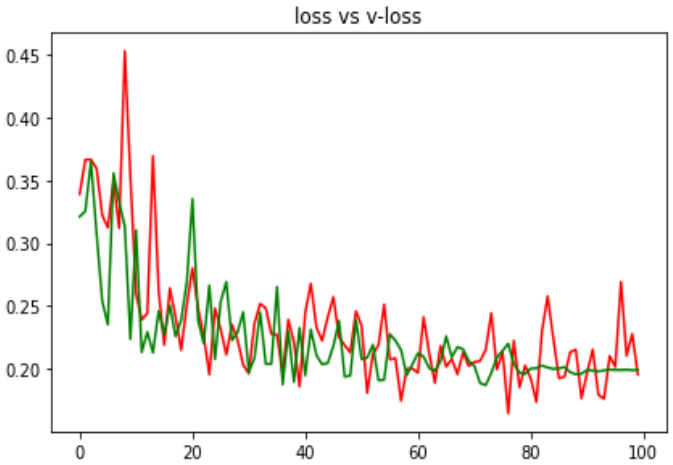
```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 224, 224, 3)]     0
block1_conv1 (Conv2D)        (None, 224, 224, 64)      1792
block1_conv2 (Conv2D)        (None, 224, 224, 64)      36928
block1_pool (MaxPooling2D)   (None, 112, 112, 64)      0
block2_conv1 (Conv2D)        (None, 112, 112, 128)     73856
block2_conv2 (Conv2D)        (None, 112, 112, 128)     147584
block2_pool (MaxPooling2D)   (None, 56, 56, 128)       0
block3_conv1 (Conv2D)        (None, 56, 56, 256)       295168
block3_conv2 (Conv2D)        (None, 56, 56, 256)       590080
block3_conv3 (Conv2D)        (None, 56, 56, 256)       590080
block3_pool (MaxPooling2D)   (None, 28, 28, 256)       0
block4_conv1 (Conv2D)        (None, 28, 28, 512)       1180160
block4_conv2 (Conv2D)        (None, 28, 28, 512)       2359808
block4_conv3 (Conv2D)        (None, 28, 28, 512)       2359808
block4_pool (MaxPooling2D)   (None, 14, 14, 512)       0
block5_conv1 (Conv2D)        (None, 14, 14, 512)       2359808
block5_conv2 (Conv2D)        (None, 14, 14, 512)       2359808
block5_conv3 (Conv2D)        (None, 14, 14, 512)       2359808
block5_pool (MaxPooling2D)   (None, 7, 7, 512)         0
flatten (Flatten)            (None, 25088)             0
predictions (Dense)          (None, 2)                 50178
=================================================================
Total params: 14,764,866
Trainable params: 50,178
Non-trainable params: 14,714,688
_________________________________________________________________
```

Model Summary

- **Predication and GRAD-CAM visualization:** The graphs of accuracy (Figure 6) and loss (Figure 4) were plotted. get img array () is a function that takes an image path and returns a pre-processed image. The function gradcam map () is then used to create a heat map for an image. After defining and initializing the respective functions, image paths were given. Type of x-ray image, chances of being covid, chances of being normal, original image and grad Cam image visualization were predicted.



Accuracy curve of the model

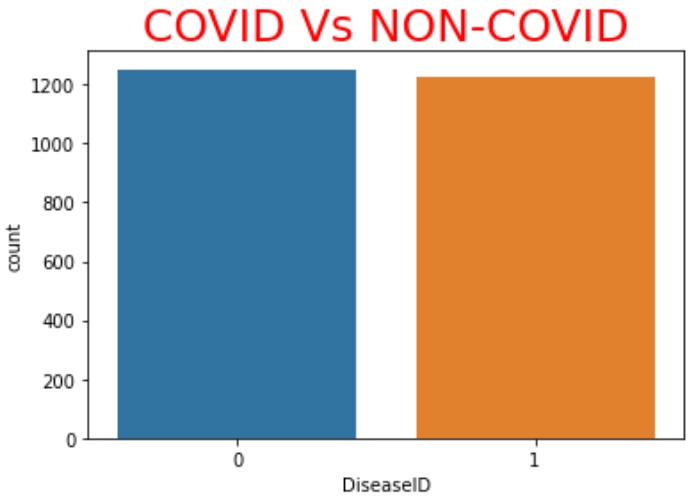


Loss curve of the model

Covid Detection from CT images

- **Dataset Collection:** The CT images dataset was collected from Kaggle, one of the most famous sites for the data science community. The dataset available at https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset was downloaded containing 1252 CT scans that are positive for SARS-CoV-2 infection (COVID-19) and 1230 CT scans for patients non-infected by SARS-CoV-2, 2482 CT scans in total. These data have been collected from real patients in hospitals from Brazil.
- **Reading and Visualization of Images:** All the required dependencies were imported. The dataset does not contain .csv files and contains only images in two categories: covid and non-covid. The images were transferred into a single directory and the entire data was converted into a data frame with 'File', 'DiseaseID' and 'Disease Type' as columns. The disease IDs were allotted as follows- covid: 0, non-covid:1. The number

of images in each category were using seaborn's count plot and train['DiseaseID'] as input parameters. First 100 images in two categories were also visualized using plt.subplots by matplotlib. Input images were then visualized.



Count plot showing number of images in each category

Subplots visualizing first 100 images

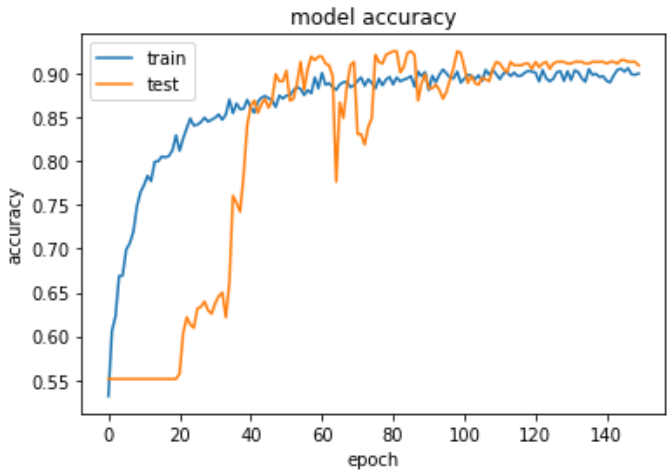
Subplots visualizing first 100 non-covid images

- **Data Normalization and splitting for training:** To organize the data, normalization was used during the data preparation to change the values of numeric columns in the dataset to use a common scale. Data normalization is the organization of data to appear similar across all records and fields. It increases the cohesion of entry types leading to cleansing, lead generation, segmentation, and higher quality data. The images were then processed for. train_test_split() with a test_size=0.2.
- **Data augmentation and model training:** Keras ImageDataGenerator class was used for augmentation of data and input parameters like zoom_range, horizontal_flip, shear_range and rescale were given. Resnet50 model was then defined and the algorithm was run at 150 epochs with learning rate as "1e-4". This code was written in anaconda's jupyter notebook interface.

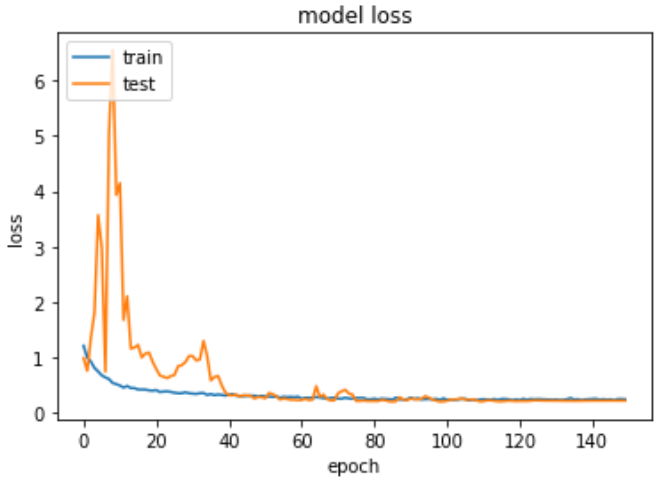
```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         [(None, 64, 64, 3)]       0
_________________________________________________________________
conv2d (Conv2D)              (None, 64, 64, 3)         84
_________________________________________________________________
resnet50 (Functional)        (None, None, None, 2048)  23587712
_________________________________________________________________
global_average_pooling2d (Gl (None, 2048)              0
_________________________________________________________________
batch_normalization (BatchNo (None, 2048)              8192
_________________________________________________________________
dropout (Dropout)            (None, 2048)              0
_________________________________________________________________
dense (Dense)                (None, 256)               524544
_________________________________________________________________
batch_normalization_1 (Batch (None, 256)               1024
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
root (Dense)                 (None, 2)                 514
=================================================================
Total params: 24,122,070
Trainable params: 24,064,342
Non-trainable params: 57,728
```

Model Summary

- **Prediction and visualization:** Accuracy and loss graphs were plotted. get_img_array () was defined which takes the image path as input and gives out a pre-processed image. After defining and initializing the respective functions, image paths were given. Type of CT scan image, chances of being covid and chances of being normal were predicted. Truth labels vs Model Predictions were shown using a confusion matrix. ROC Curve was also plotted to see the exchange between the true positive rate and false positive rate using different probability thresholds.



Model Accuracy



Model Loss

Results

COVID-19 detection model:

Tensorflow was used to create CNN models, which were wrapped in the Python framework Keras. Google Colab and Jupyter Notebook were used to conduct the research. The CNN model was trained using the Adam optimization algorithm for hyperparameter optimization and cross-entropy as the loss function.

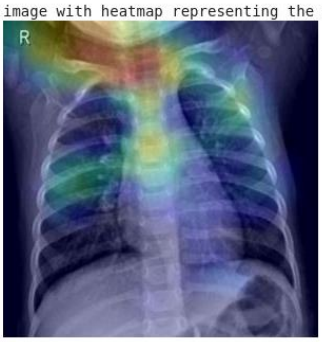
COVID-19 detection from X-rays

The VGG16 model used in this case achieved the best test accuracy of 0.9237499833106995. The actual image disease condition and the predicted disease condition were found to be the same thus fulfilling the objective of detection. Image paths were given and changes of being covid or non-covid were visualized.

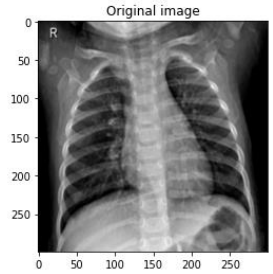
The given X-Ray image is of type = Normal

The chances of image being Covid is : 8.306345343589783 %
The chances of image being Normal is : 90.17686247825623 %

image with heatmap representing the covid spot



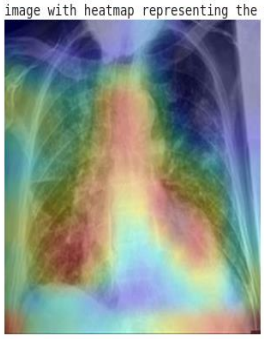
the original input image



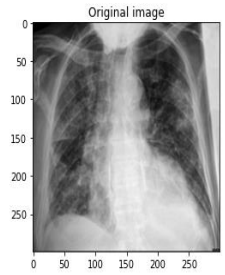
The given X-Ray image is of type = Covid

The chances of image being Covid is : 87.96427249908447 %
The chances of image being Normal is : 9.105725586414337 %

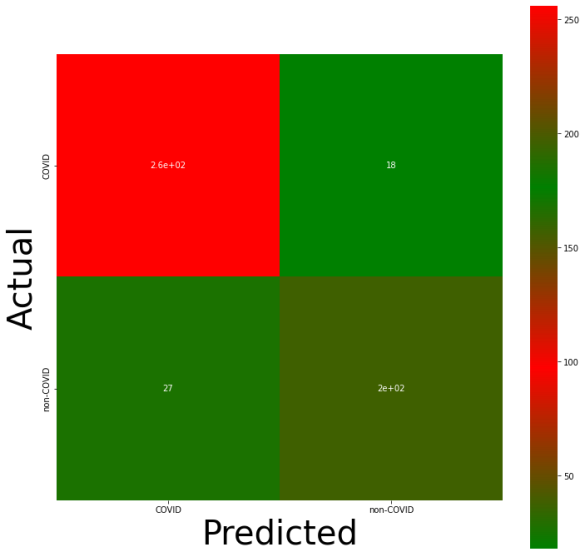
image with heatmap representing the covid spot

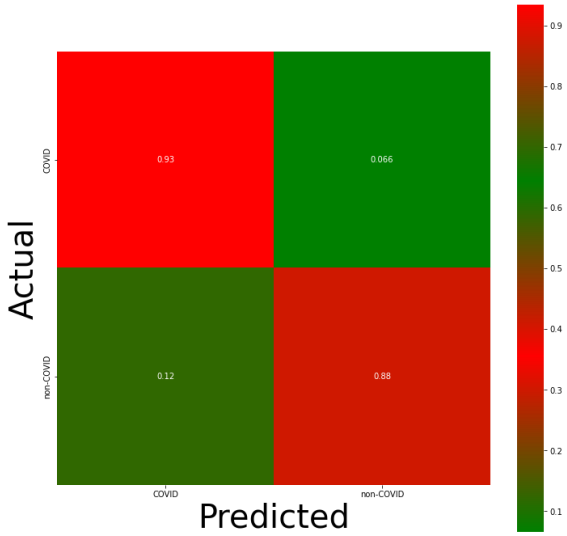


the original input image



Prediction of Infection





Confusion Matrix before and after normalization

COVID-19 detection from CT

Confusion Matrix is used as a performance measurement for this COVID-19 classification. Before and after normalization confusion matrices were plotted with TP=0.93, FP=0.066, FN=0.12, TN=0.88. Classification report was generated with precision, recall, f1-report and support scores of both covid and non-covid.

	precision	recall	f1-score	support
0	0.90	0.93	0.92	274
1	0.92	0.88	0.90	223
accuracy			0.91	497
macro avg	0.91	0.91	0.91	497
weighted avg	0.91	0.91	0.91	497

Classification Report

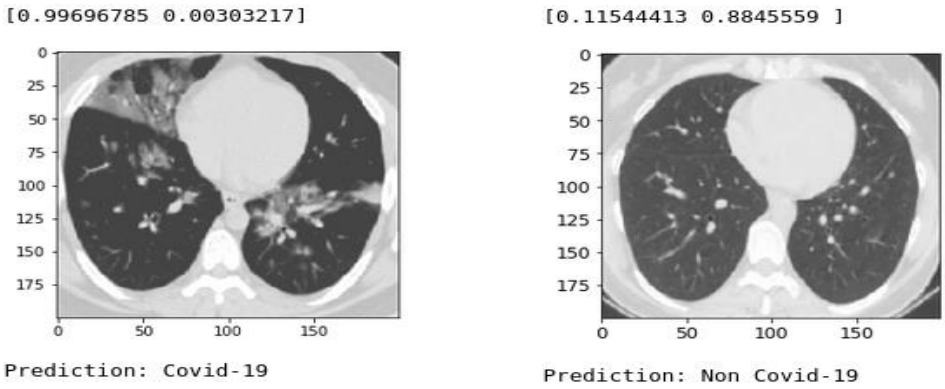
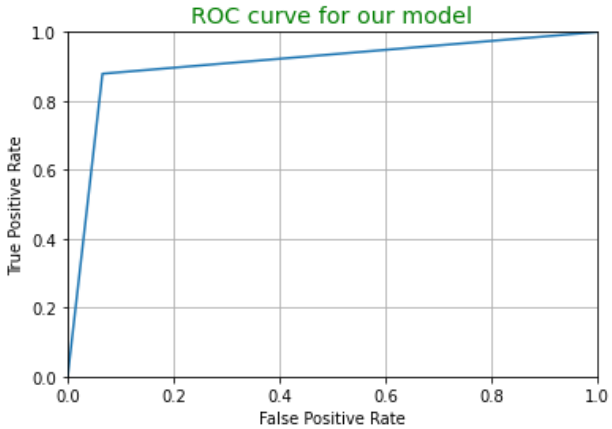
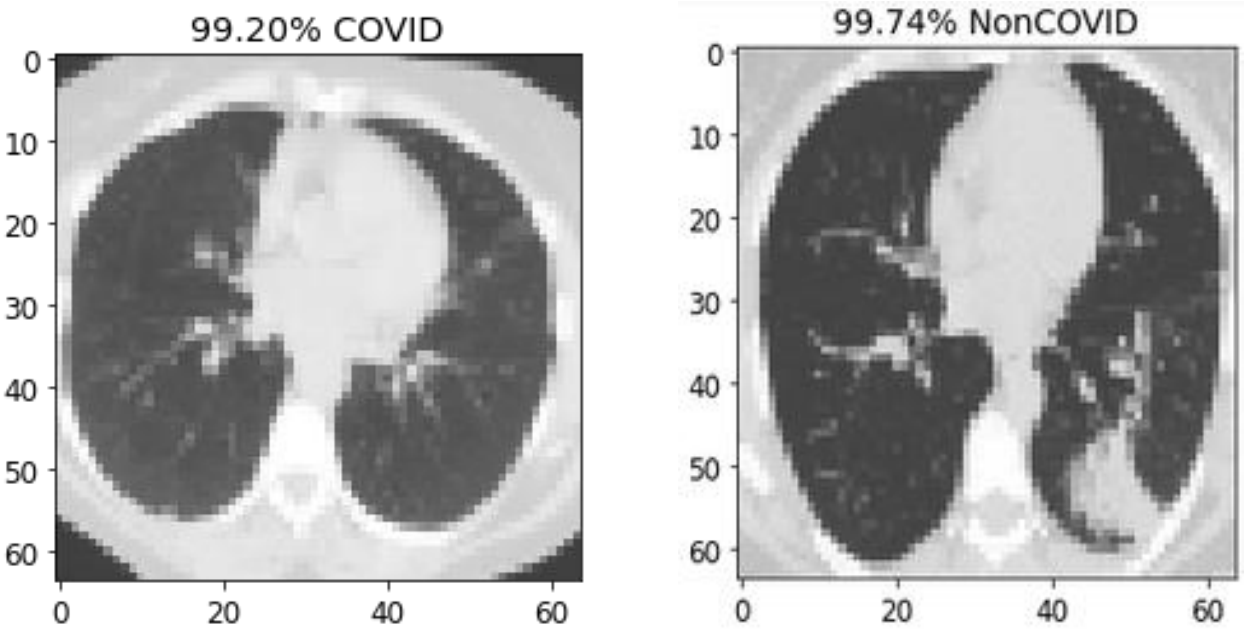


Figure 25: Prediction of



ROC curve of the model



A snippet of two images of prediction of probability of being covid or non-covid for first 50 images.

Conclusion

After analyzing how COVID-19 affected the entire healthcare system and how a lack of appropriate diagnosis resulted in high mortality rates, we realized that deploying machine learning models might aid in the faster detection and treatment of one of these kinds of diseases. The main objectives were to use CNN to detect COVID-19 from radiography pictures and predict future curves from COVID-19 cases.

Our objective was to detect the COVID-19 infection from radiography image dataset. Both the X-Ray and CT CNN models achieved good accuracy in detecting the infection. Deep Learning algorithms to classify two classes COVID-19 and normal using Transfer Learning concept and using the pre-trained architectures VGG16 and ResNet50 were able to determine the characteristics of infection correctly. Rolling out these kinds of ML models can really be very helpful in rural areas and at the time of crises.

References

1. İ. Mertyüz, T. Mertyüz, B. Taşar and O. Yakut, "Covid-19 Disease Diagnosis From Radiology Data With Deep Learning Algorithms," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2020, pp. 1-4, doi: 10.1109/ISMSIT50672.2020.9255380.
2. Alballa N, Al-Turaiki I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Inform Med Unlocked. 2021;24:100564. doi: 10.1016/j.imu.2021.100564. Epub 2021 Apr 3. PMID: 33842685; PMCID: PMC8018906.
3. Qiblawey Y, Tahir A, Chowdhury MEH, Khandakar A, Kiranyaz S, Rahman T, Ibtehaz N, Mahmud S, Maadeed SA, Musharavati F, Ayari MA. Detection and Severity Classification of COVID-19 in CT Images Using Deep Learning. *Diagnostics*. 2021; 11(5):893. https://doi.org/10.3390/diagnostics11050893
4. Aktar S, Ahamad M, Rashed-Al-Mahfuz M, Azad A, Uddin S, Kamal A, Alyami S, Lin P, Islam S, Quinn J, Eapen V, Moni M. Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data: Statistical Analysis and Model Development. JMIR Med Inform 2021;9(4): e25884. URL: https://medinform.jmir.org/2021/4/e25884. DOI: 10.2196/25884
5. Quiroz-Juárez MA, Torres-Gómez A, Hoyo-Ulloa I, León-Montiel RdJ, U'Ren AB (2021) Identification of high-risk COVID-19 patients using machine learning. PLOS ONE 16(9): e0257234. https://doi.org/10.1371/journal.pone.0257234

6. Abdullha and S. Abujar, "COVID-19: Data Analysis and the situation Prediction Using Machine Learning Based on Bangladesh perspective," 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2020, pp. 1-8, doi: 10.1109/iSAI-NLP51646.2020.9376812.

7. A. Abdullha and S. Abujar, "COVID-19: Data Analysis and the situation Prediction Using Machine Learning Based on Bangladesh perspective," 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2020, pp. 1-8, doi: 10.1109/iSAI-NLP51646.2020.9376812.
8. Mondal MRH, Bharati S, Podder P. Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A Review. Curr Med Imaging. 2021;17(12):1403-1418. doi: 10.2174/1573405617666210713113439. PMID: 34259149.
9. V. Chamola, V. Hassija, V. Gupta and M. Guizani, "A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact," in IEEE Access, vol. 8, pp. 90225-90265, 2020, doi: 10.1109/ACCESS.2020.2992341.
10. Sumayh S. Aljameel, Irfan Ullah Khan, Nida Aslam, Malak Aljabri, Eman S. Alsulmi, "Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients", Scientific Programming, vol. 2021, Article ID 5587188, 10 pages, 2021. https://doi.org/10.1155/2021/5587188
11. Sarki R, Ahmed K, Wang H, Zhang Y, Wang K (2022) Automated detection of COVID-19 through convolutional neural network using chest x-ray images. PLoS ONE 17(1): e0262052. https://doi.org/10.1371/journal.pone.0262052
12. Haque KF, Abdelgawad A. A Deep Learning Approach to Detect COVID-19 Patients from Chest X-ray Images. *AI*. 2020; 1(3):418-435. https://doi.org/10.3390/ai1030027
13. Nasser H. Sweilam, A.A. Tharwat, N.K. Abdel Moniem, Support vector machine for diagnosis cancer disease: A comparative study, Egyptian Informatics Journal, Volume 11, Issue 2, 2010, Pages 81-92, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2010.10.005
14. Rahman, Khaleelur & Sathik, Mohamed & Kaliyaperumal, Senthamarai. (2012). Multiple Linear Regression Models in Outlier Detection. International Journal of Research in Computer Science. 2. 10.7815/ijorcs.22.2012.018.
15. Mustafa Ghaderzadeh, Farkhondeh Asadi, "Deep Learning in the Detection and Diagnosis of COVID-19 Using Radiology Modalities: A Systematic Review", Journal of Healthcare Engineering, vol. 2021, Article ID 6677314, 10 pages, 2021. https://doi.org/10.1155/2021/6677314

16. Qjidaa, M. & Mechbal, Yassine & Ben-fares, A. & Amakdouf, Hicham & Maaroufi, Maher & Alami, Badreeddine & Qjidaa, H.. (2020). Early detection of COVID19 by deep learning transfer Model for populations in isolated rural areas. 1-5. 10.1109/ISCV49265.2020.9204099.