Drawing Insights from COVID-19 Infected Patients With no Past Medical History Using CT Scan Images and Machine Learning Techniques: A Study on 200 Patients

Dr. Sachin Sharma (sachin.sharma@iar.ac.in)
Institute of Advanced Research, Gandhinagar https://orcid.org/0000-0002-3072-8698

Method Article

Keywords: Coronavirus, COVID-19, Machine Learning, Computed Tomography (CT) Scan, Pneumonia, Polymerase Chain Reaction (PCR)

Posted Date: April 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-23863/v1

Version of Record: A version of this preprint was published on July 22nd, 2020. See the published version at https://doi.org/10.1007/s11356-020-10133-3.

Abstract

As the whole world is witnessing what novel Coronavirus (COVID-19) can do to the mankind, it presents several unique features also. In absence of specific vaccine for COVID-19, it is essential to detect the disease at an early stage and isolate an infected patient. Till today there is a global shortage of testing labs and testing kits for COVID-19. This paper discusses about the role of machine learning techniques for getting important insights like whether lung Computed Tomography (CT) scan should be the first screening /alternative test for real-time reverse transcriptase-polymerase chain reaction (RT-PCR), Is COVID-19 pneumonia different from other viral pneumonia and if yes how to distinguish it using lung CT scan images from the carefully selected data of lung CT scan COVID-19 infected patients from the hospitals of Italy, China and India having no past medical illnesses i.e. high blood pressure, diabetes and heart disease. The reason for selecting images of the patients having no past illnesses history is that people with medical problems are already known to develop serious illness because of COVID-19 but we wanted to check and analyse condition of those patients (lung condition) which don't have past medical history using CT scan images.

I. Introduction And Literature Review

The death toll from the new coronavirus surpassed 6,000 in Europe while the worldwide deaths surged past 12,000 according to data collected by the Johns Hopkins University in the United States up to the date when this data was shared to this literature. More than 299,000 people have been infected, while some 91,500 have recovered. As per the definition and information shared by World Health Organization (WHO), coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. People having medical problems like heart disease, diabetes and high blood pressure are more likely to develop serious illness. Currently, there are no specific vaccines or treatments for COVID-19.

As per the information shared by Radiological Society of North America (RSNA), X-ray images of a Chinese person who was killed by COVID-19 shows what the virus does to sufferers' lungs. Taking a deep look at images, it shows white patches in the lower corners of the lungs which indicate what radiologists say ground glass opacity - the partial filling of air spaces. Similar symptoms were seen in case of 54-year-old woman caught with Covid. Figure 1 and 2 shows the images of it. So, if some distinctive patterns are there, machine learning techniques can be used for early detection of it.

Studies related to understanding and early detection of coronavirus using x-ray images and by other means is still going on. Work done in [Xu et al. 2020] classified CT scan images of COVID-19 patients into three classes as healthy cases, Influenza viral pneumonia and COVID-19. A total of 618 images were taken for the database, which included 175 images of 175 healthy people, 224 images of 224 patients with Influenza-A pneumonia, 219 images of 110 patients infected with coronavirus. An overall accuracy of 87.6% was achieved using 3D-dimensional deep learning model. Authors in [Shan et al. 2020] developed a system based on deep learning mechanism for segmenting and quantification the infected regions and the entire lung using chest CT images. In their study, 249 COVID-19 patients and 300 new

COVID-19 patients for validation were used. They used Dice similarity 2 coefficient concept and got around it 91.6%. It is mentioned in their study that system reduced the delineation time to almost four minutes.

Ii. Distinguishing Covid-19 Pneumonia From Other Viral Pneumonia

As per the information shared by a respiratory physician in The Guardian (a leading British daily newspaper), COVID-19 pneumonia is different from the most common cases that people are admitted to hospitals for. As per him, cases of coronavirus pneumonia tend to affect all the lungs, instead of just small parts. As shown in figure 3, the image shows a CT scan from a man with Covid-19. Pneumonia caused by the coronavirus shows a typical hazy patch on the outer edges of the lungs, indicated by arrows. Some other labelled images (infected regions) of CT scan of a patient with Covid-19 are shown in figure 4.

Iii. Selection Of Database

As the accuracy of any machine learning algorithm depends on the type and quality of data that is provided to it, the database for our experiment was very carefully selected keeping in mind the goals that we wanted to achieve. Only those patients(images) were selected who were having no past medical history such as high blood pressure, diabetes and heart disease and also whose lung CT scan images showed typical patches on the outer edges of the lungs.

Iv. Research Methodolgy

As shown in figure 5, first the CT scan images of positive covid and normal healthy person are taken and stored in the computer. Then we are doing some pre-processing steps to enhance the image. We are making separate folders for positive cases and normal cases. For feature extraction and learning of the system, we are using custom vision software based on machine learning algorithm of Microsoft azure. Once the system gets trained, we test the system on the unseen images of positive and normal cases. After getting the results of testing, we check it with the actual condition (positive/normal) of the patient for getting accuracy of the trained model. Next step is to deploy the model.

V. Procedure For Training And Testing The System

We have collected and added all chest CT scan images in the database for training the system. All the images are collected from the official database of different hospitals of Italy, China and India. We built our database of almost 1200 images consisting of:

CT scan positive case images 400, normal healthy person CT scan image 800.

Following is the proposed procedure for training and testing of the data for COVID-19 detection:

- Collect all positive and normal images in the data folder
- Image annotation/labelling
- Training the model based on machine learning algorithm
- Testing
- Retrain if needed and finally deploy or export the model for offline use

The average time it took to train the system was 19 hours. Figure 6 shows the environment setup in custom vision software of Microsoft azure. Figure 7(a) and 7(b) shows the sample positive case images and normal case images.

Vi. Experiment And Result Analysis

We built our database of almost 1200 images consisting of:

CT scan positive case images 400 (fig 7(a)), normal healthy person CT scan image 800 (fig 7(b)). 450 images (200 positive cases of COVID, 250 normal healthy cases) were kept for testing. The trained model had never seen these images before. We already had all the information like positive/negative covid category, health data (collected from the hospitals) that the trained model was to test. The reason for collecting the images from different countries was that we wanted to check if there is any influence or bias of C0VID-19 with any place and also to develop a model which is robust and gives the same accuracy irrespective of the location or people.

Parameters important for validating the performance of the classifier are Sensitivity (True Positive Rate), Specificity (True Negative Rate) and Accuracy [Gupta et al. 2013] which are given as

$$Sensitivity = \frac{TP}{TP+FN} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Here in above equations, TN stands for true negative; TP stands for true positive; FN stands for false negative, and FP stands for false positive. True positive (TP) and true negative (TN) are the most relevant and correct parameters of classification.

After training and testing the model, we got accuracy close to 93%, sensitivity close to 92% and specificity close to 95% with TP= 200, TN=250, FP=15 and FN=19. We performed extensive experiments and spent so many hours testing the system. Training and testing on more quality image datasets may improve accuracy of the model.

Figure 8(a) and 8(b) shows the true positive and true negative case as detected correctly by our trained model. Fig 9 shows the false positive and false negative cases that were detected wrongly by our trained

model. Figure 10 shows some cases where model was little confused in classifying positive and normal cases (shown in percentage of classification).

Some of the important outcomes of the experiment are:

- Of the images (patients) that were wrongly classified by our trained model based on CT scan, PCR was carried out by the hospital and they were confirmed positive which means that PCR is necessary for the final diagnosis but as our model based on CT scan showed good results in terms of accuracy and also as it takes less time (no blood sample collection, shipping issues), we can say that CT scan diagnosis can be the first screening test for the patients.

- As per the information shared by Radiopaedia which is a wiki-based international collaborative radiology educational web resource containing reference articles, radiology images, and patient cases, though the definitive test for COVID-19 is the real-time reverse transcriptase-polymerase chain reaction(RT-PCR) test and is believed to be highly specific, but there are cases reported with sensitivity as low as 60-70% and as high as 95-97% depending on the country. Thus, false negatives are a real clinical problem and several negative tests might be required in a single case to be confident about excluding the disease.

- Pneumonia caused by COVID-19 is particularly severe. Cases of coronavirus pneumonia tend to affect all of the lungs, instead of just small parts. Pneumonia caused by coronavirus shows a typical patch on the outer edges of the lungs.

Vii. Conclusion And Future Work

Coronavirus is a global problem and it not only has a huge impact on health of citizens but also on the global economy. In this paper we discussed about the role of machine learning techniques for getting important insights like whether lung Computed Tomography (CT) scan be first screening /alternative test for real-time reverse transcriptase-polymerase chain reaction (RT-PCR), is COVID-19 pneumonia different from other viral pneumonia and if yes how to distinguish it using lung CT scan images from the carefully selected data of lung CT scan COVID-19 infected patients from the hospitals of Italy, China and India having no previous medical history (high blood pressure, diabetes and heart disease).

Training and testing were done using custom vision software based on machine learning techniques of Microsoft azure. An accuracy of almost 93% was achieved. Though there were some false indications also. As the model based on CT scan showed good results in terms of accuracy and as it takes less time (no blood sample collection, shipping issues), we can conclude that CT scan diagnosis can be the first screening/alternative test for real-time reverse transcriptase-polymerase chain reaction (RT-PCR) test for the patients. Pneumonia caused by coronavirus shows a typical hazy patch on the outer edges of the lungs which suggests a pattern and so machine learning techniques can be used early detection of corona virus. Training and testing on more quality image datasets may further improve accuracy of the model.

References

N. Gupta, A. Rawal, V. L. Narasimhan, S. Shivani, "Accuracy sensitivity and specificity measurement of various classification techniques on healthcare data", *IOSR J. Comput. Eng.*, vol. 11, no. 5, pp. 70-73, May/Jun. 2013

Shan. F, Gao. Y, Wang. J, Shi. W, Shi. N, Han. M, et al. “Lung Infection Quantification of COVID-19 in CT Images with Deep Learning”. arXiv preprint arXiv:200304655. 2020

Xu X, Jiang X, Ma C, Du P, Li X, Lv S, et al. “Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia”. arXiv preprint arXiv:200209334. 2020

https://www.aljazeera.com/news/2020/03/coronavirus (Last accessed on 19th March 2020)

https://www.dailymail.co.uk/news/article-8101383 (Last accessed on 20th March 2020)

https://radiopaedia.org/articles/COVID-19-3 (Last accessed on 15th March 2020)

https://www.theguardian.com/world/2020/mar/24/coronavirus (Last accessed on 11th March 2020)

https://www.who.int/health-topics/coronavirus (Last accessed on 18th March 2020)

Declarations

Ethics approval and *consent* to participate:
Acquisition of all clinical images was granted by subject verbal *consent.* The images in this paper are *obtained* from an open database of hospitals in Italy (It is an open source database.
Link: https://www.sirm.org/en/). These are repositories of anonymised images accessible locally for educational purposes and no identifiable information is stored or available. *Ethical approval* for their use in publication was therefore deemed unnecessary.
Consent for publication
Consent for publication from the patient was *obtained*
Availability of data and materials
Extra data is available by emailing to sachin.sharma@iar.ac.in on reasonable request.
Competing interests
The authors declare that they have no competing interests.
Funding
None

Figures

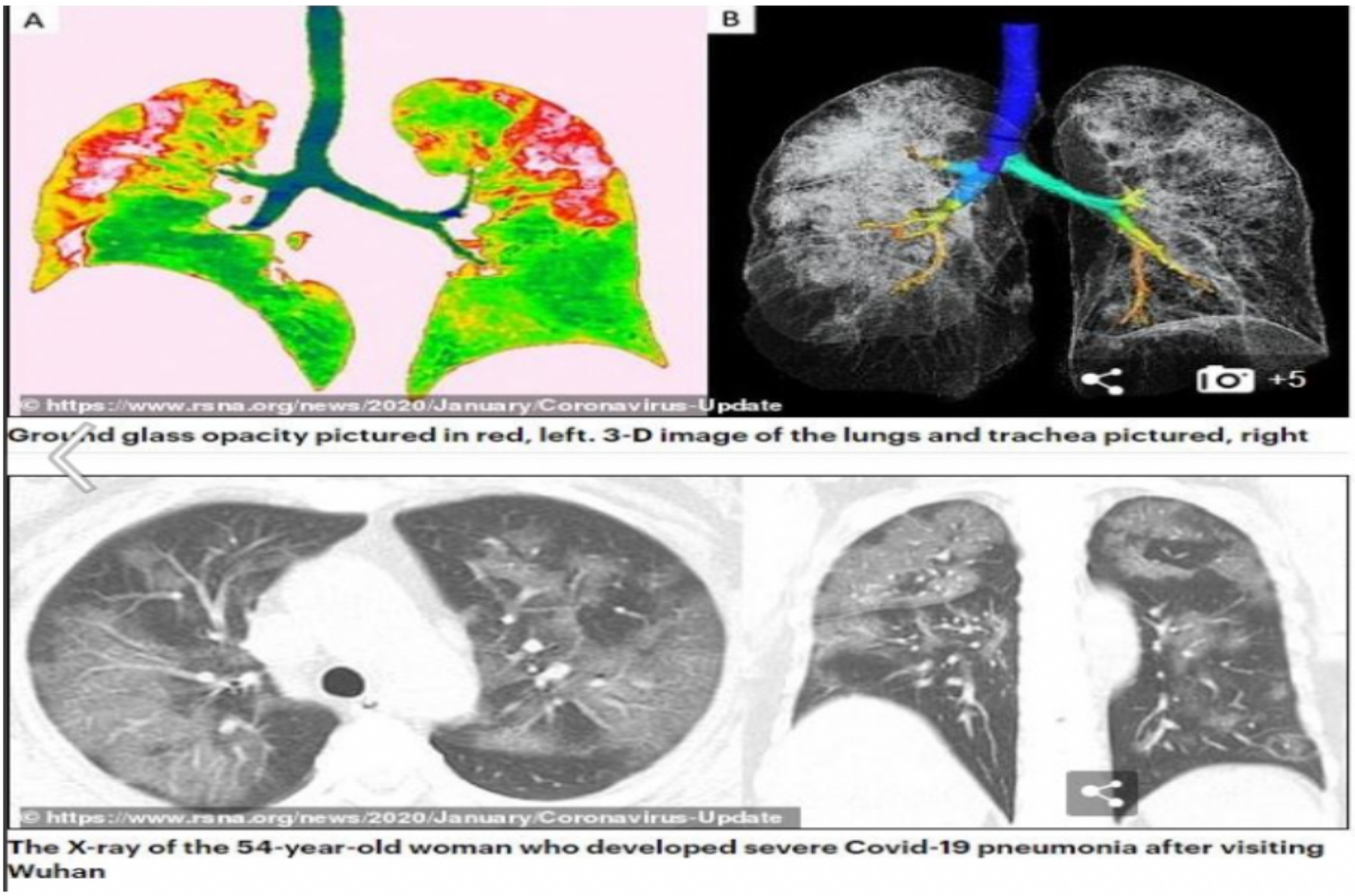


Figure 1

X-ray image of a patient with severe COVID-19 pneumonia (Source: RNSA)

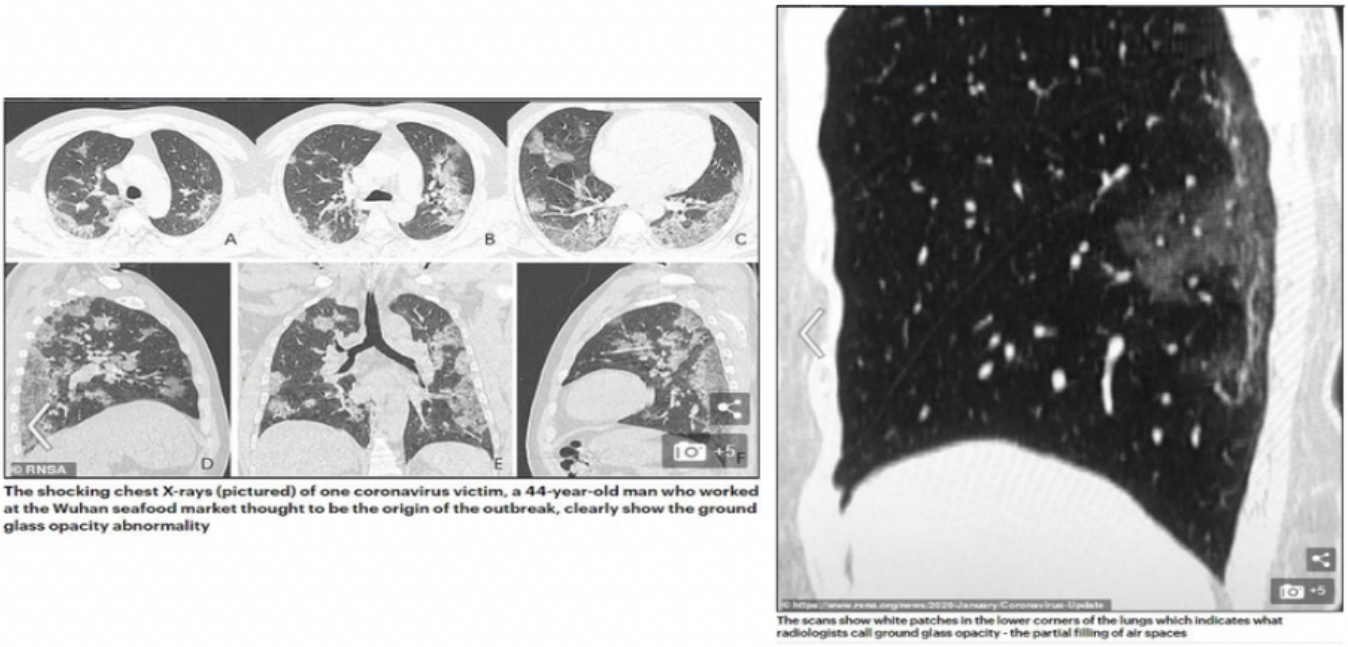
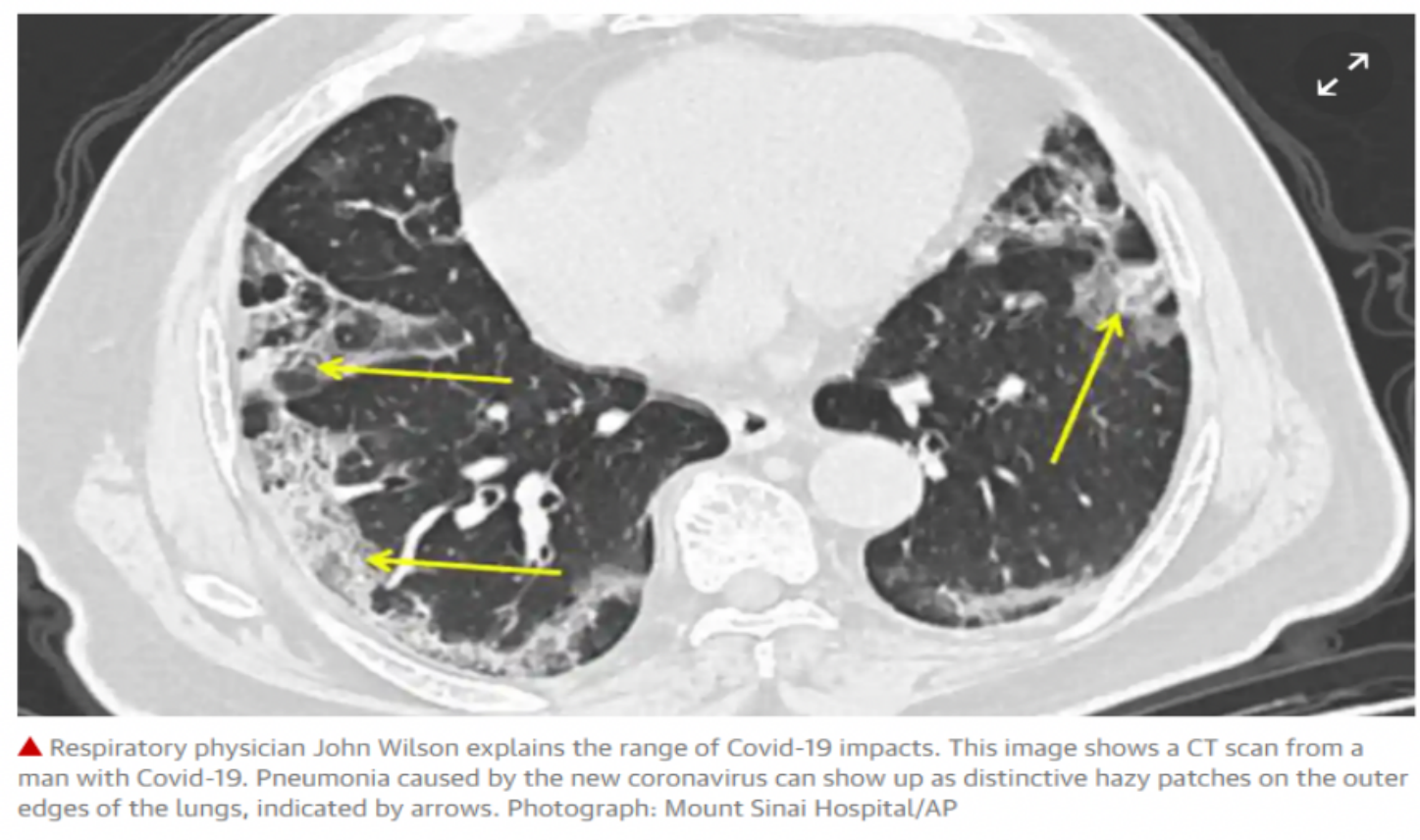


Figure 2

X-ray and CT scan image of a coronavirus victim showing white patches in the lower corners of the lungs (Source: RNSA)



▲ Respiratory physician John Wilson explains the range of Covid-19 impacts. This image shows a CT scan from a man with Covid-19. Pneumonia caused by the new coronavirus can show up as distinctive hazy patches on the outer edges of the lungs, indicated by arrows. Photograph: Mount Sinai Hospital/AP

Figure 3

CT scan image of a patient with severe COVID-19 pneumonia showing a distinguishing hazy patch on the outer edges of the lungs, indicated by arrows (Source: Mount Sinai Hospital/AP)

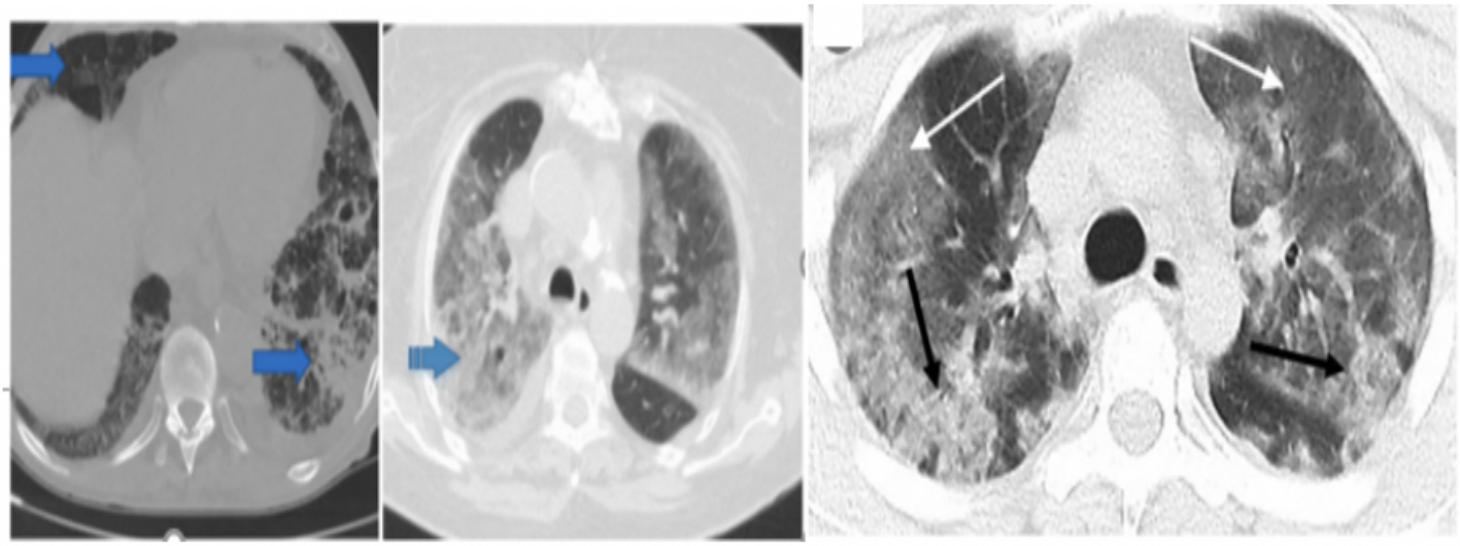


Figure 4

CT scan image of patients with severe COVID-19 pneumonia showing distinguishing patches on the outer edges of the lungs, indicated by different coloured arrows

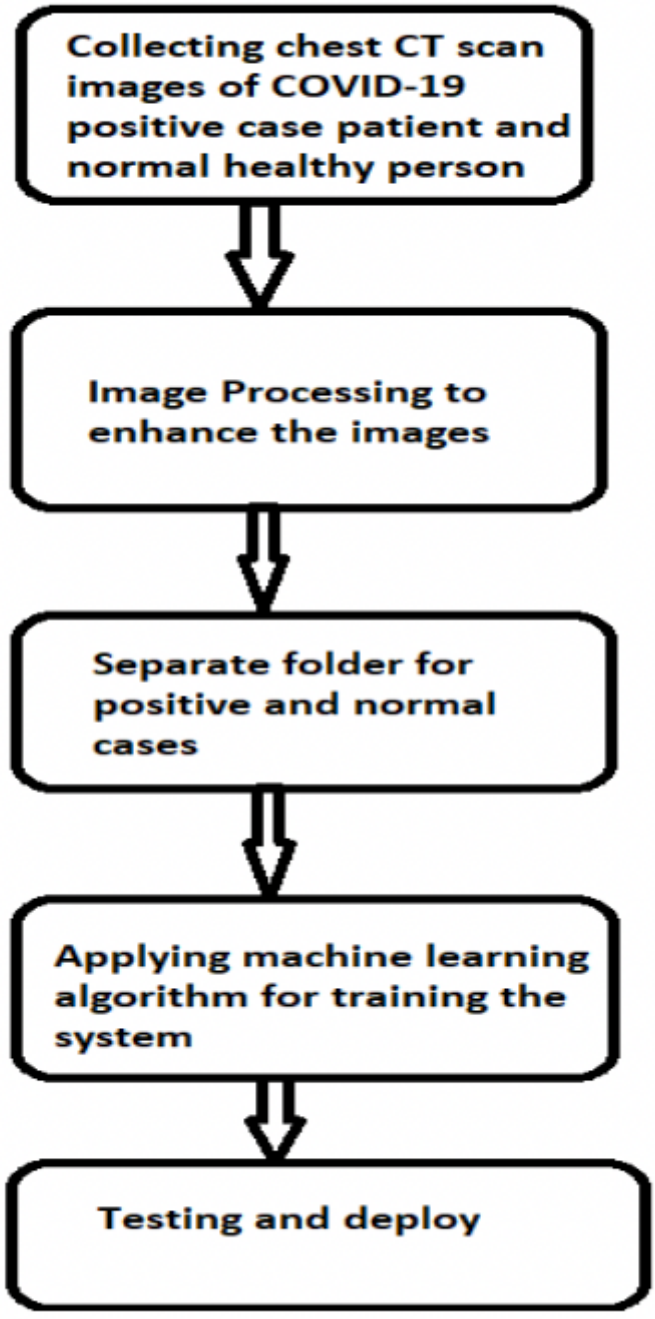


Figure 5

Block diagram of the system

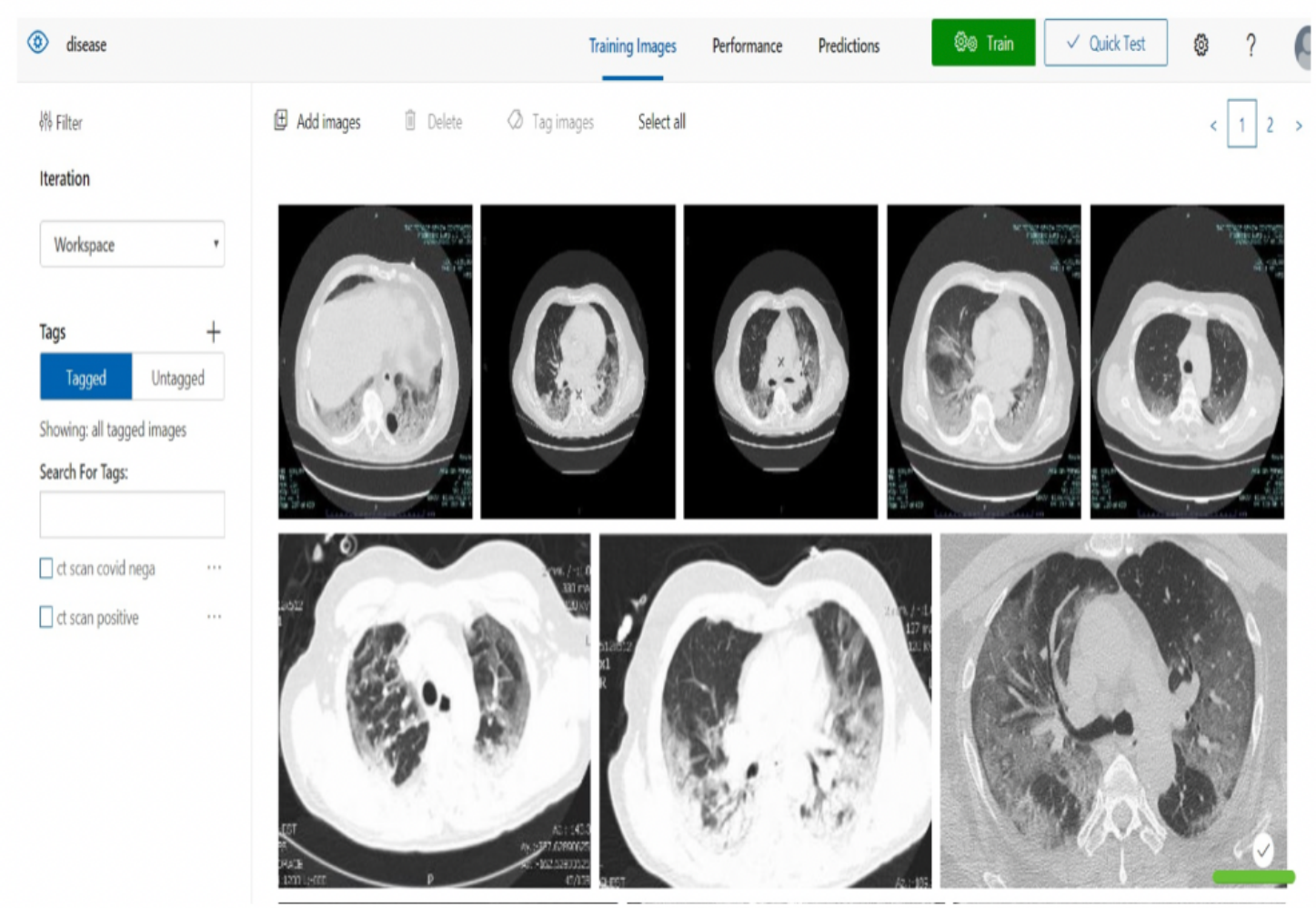


Figure 6

Environment setup for collecting, labelling, training and testing the images

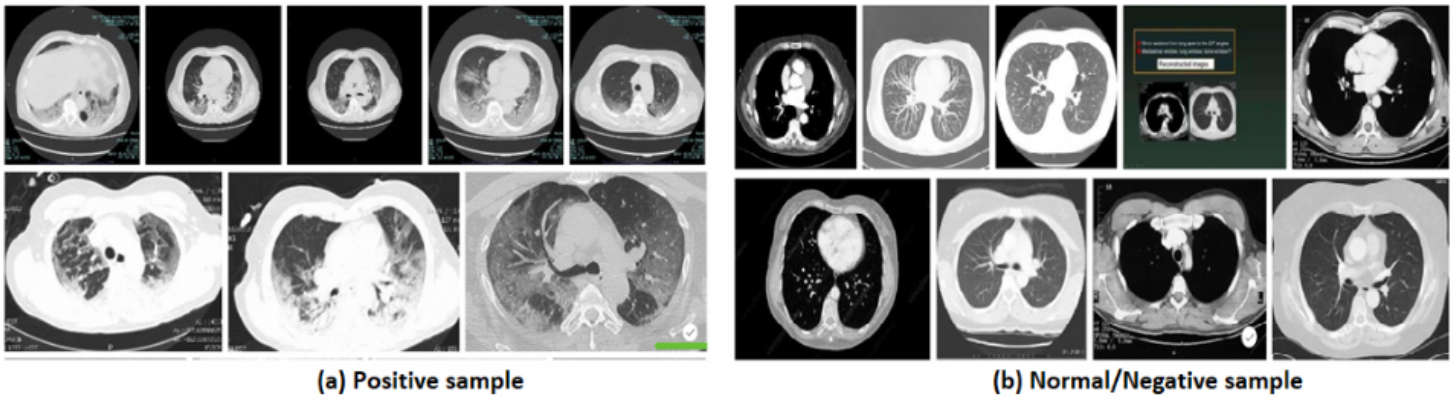


Figure 7

Sample CT scan images of positive COVID-19 and normal healthy person

Figure 8

True Positive and True Negative cases detected correctly by the trained model

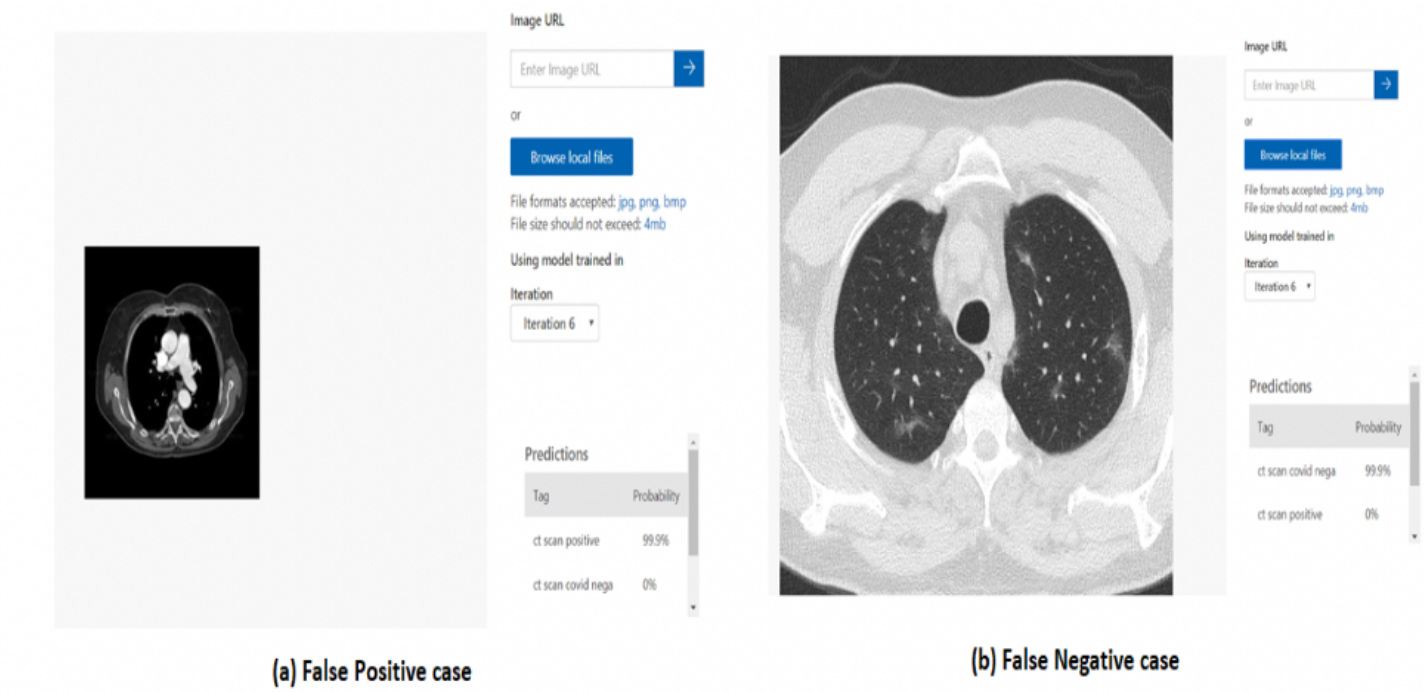


Figure 9

False Positive and False Negative cases detected wrongly by the trained model

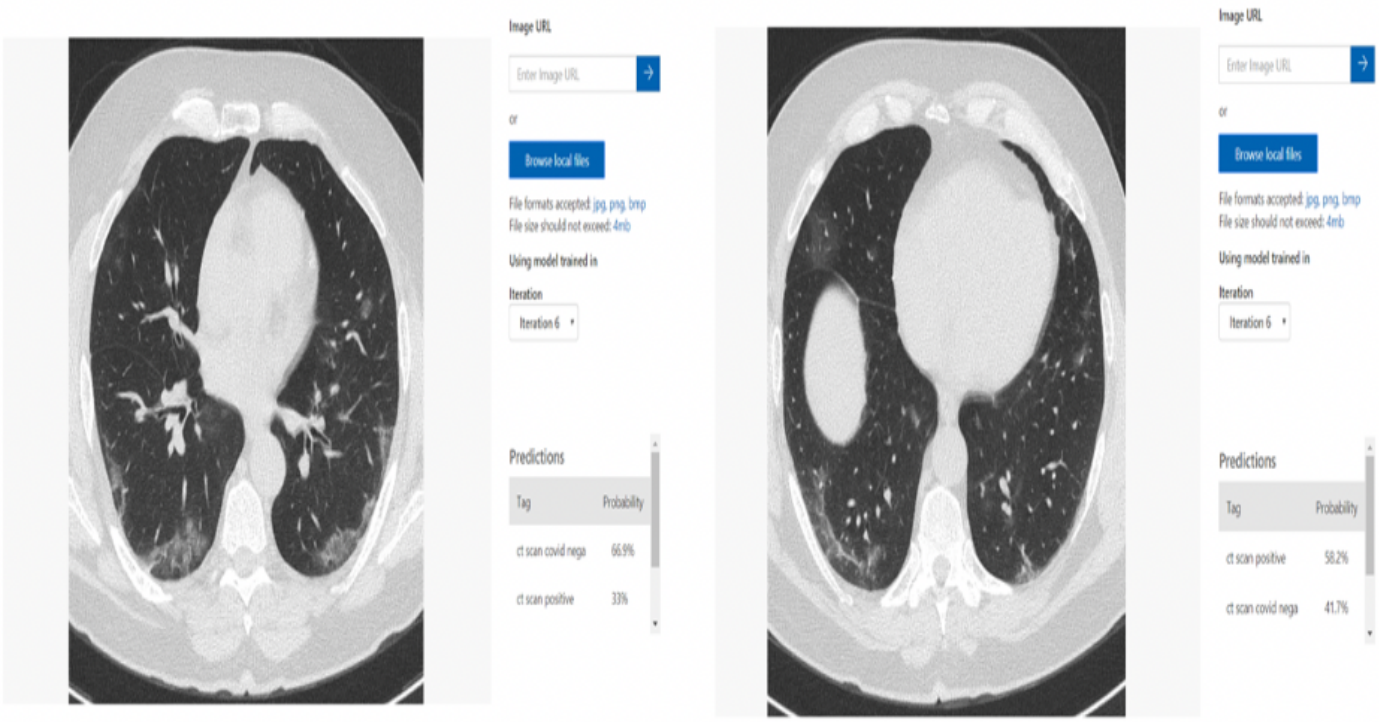


Figure 10

Some confusing cases

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- COVID19EARLYDETECTIONUSINGCTSCANIMAGES.mp4