Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients

Sachin Sharma¹

Received: 4 April 2020 / Accepted: 14 July 2020 / Published online: 22 July 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

As the whole world is witnessing what novel coronavirus (COVID-19) can do to the mankind, it presents several unique features also. In the 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, Moscow and India? For training and testing the proposed system, custom vision software of Microsoft azure based on machine learning techniques is used. An overall accuracy of almost 91% is achieved for COVID-19 classification using the proposed methodology.

Keywords Coronavirus · COVID-19 · Machine learning · Computed tomography (CT) scan · Pneumonia · Polymerase chain reaction (PCR)

Introduction and literature review

The death toll from the new coronavirus surpassed 6000 in Europe, while the worldwide deaths surged past 12,000 according to data collected by the Johns Hopkins University in the USA up to the date when these data were 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 show 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. Figures 1 and 2 show 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 are 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 and 219 images of 110 patients infected with coronavirus. An overall accuracy of 87.6% was achieved using 3D-deep learning model. Shan et al. (2020) developed a system based on deep learning mechanism for segmenting and quantifying 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.

Responsible editor: Philippe Garrigues

✉ Sachin Sharma
sachin.sharma@iar.ac.in

¹ Department of Engineering and Computing, Institute of Advanced Research, Gandhinagar, India

Fig. 1 X-ray image of a patient with severe COVID-19 pneumonia (Source: RNSA)

The rest of the paper is arranged as follows. In the “Distinguishing COVID-19 pneumonia from other viral pneumonia” section, how to distinguish COVID-19 pneumonia from other viral pneumonia such as respiratory syncytial virus (RSV) and Influenza A (H1N1, H5N1) is mentioned. In the “Selection of database” section, selection of database is discussed. Proposed research methodology is mentioned in the “Research methodology” section. The “Procedure for training and testing the system” section discusses the procedure for training and testing the system. Experimental work and result analysis are mentioned in the “Experiment and result analysis” section. Final conclusion and future work are mentioned in the “Conclusion and future work” section.

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 Fig. 3, the image shows a CT scan from a person 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

Fig. 2 X-ray and CT scan image of a coronavirus victim showing white patches in the lower corners of the lungs (Source: RNSA)

Fig. 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)

▲ 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

regions) of CT scan of a patient with COVID-19 are shown in Fig. 4.

As per the information shared by various radiologists and respiratory physicians, there are predominantly three features/symptoms, namely ground-glass opacities (GGO), consolidation and pleural effusion seen in the chest CT scan image of a COVID-19 patient. As per the definition, ground-glass opacities refer to the hazy appearance of the lungs on imaging studies, almost as if sections are obscured by ground-glass. It may be due to the filling of pulmonary airspaces with fluid, the collapse of the airspaces or both. It is a pattern that can be seen when the lungs are sick. Normal lung CT scans appear black; an abnormal chest CT with GGOs will show lighter-coloured or gray patches. Consolidation refers to the filling of pulmonary airspaces with fluid or other products of inflammation. Pleural effusion refers to abnormal fluid, which develops in the spaces around the lungs. Figure 5 shows a sample CT scan image of a COVID-19 patient with ground-glass opacities (GGO), consolidation and pleural effusion marked/filled with different colours for better understanding. Though it is really a challenge to distinguish COVID-19 pneumonia

Fig. 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

from other viral pneumonia but after discussing with various radiologists and respiratory physicians, the focus of infections located close to the pleura apart from having other features/symptoms mentioned earlier in the chest CT scan is more likely to be recognized as COVID-19.

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 that had COVID-19 symptoms (CT scan showing typical patches on the outer edges of the lungs), other viral pneumonia symptoms such as respiratory syncytial virus (RSV), Influenza A (H1N1, H5N1) and normal healthy lungs.

Fig. 5 Ground-glass opacities in blue, consolidation in yellow and pleural effusion in green (Source: <http://medicalsegmentation.com/covid19/>)

Figure 6 shows sample lung CT scan image of normal healthy person, COVID-19 patient and other viral pneumonia patient.

Research methodology

As shown in Fig. 7, first the chest CT scan images of COVID-19 patient, other viral pneumonia patient and normal healthy person are taken and stored in the computer. Then we are doing some image pre-processing steps, i.e. image cropping (ROI) and image resizing to extract effective pulmonary regions before using the dataset. Ground-glass opacities (GGO), consolidation and pleural effusion are the features that are used as they are the predominant features seen in the CT scan image of a COVID-19 patient. Then we are making separate folders for all the three different categories. For training the system, custom vision software based on machine learning techniques, i.e. residual neural network (ResNet) architecture of Microsoft azure, is used. ResNet are used to build a deeper network compared with other plain networks and simultaneously find an optimised number of layers to negate the vanishing gradient problem. Once the system gets trained, we test the system on the unseen images for all the three cases. Proper selection of ROI and Grad-Cam/heatmap, which are basically used to understand where the network is “looking” in the input image, which series of neurons activated in the forward-pass during inference/prediction, and how the network arrived at its final output (to make it easy for doctors to test the reliability of the model) are used for improving the performance. After getting the results of testing, we check it with the actual/ground truth condition (COVID-19 or other viral pneumonia or normal healthy case) of the patient for validating the accuracy of the trained model. Once the desired accuracy is achieved, next step is to deploy the model. Figure 8 shows the training process using ResNet architecture (CNN-based) for COVID-19 classification. Figure 9 shows the sample Grad-Cam for COVID-19 patient.

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

Fig. 6 Sample lung CT scan images of normal healthy person, COVID-19 patient and other viral pneumonia patient

Fig. 7 Block diagram of the system

of China, Italy, Moscow and India (mentioned below). Around 2200 images were collected consisting of the following: COVID-19 patient CT scan images 800, other viral pneumonia patient CT scan images 600 and normal healthy person CT scan image 800.

Database Sources:

- Dataset of more than 100 axial CT images from > 80 patients with COVID-19 provided by Italian Society of Medical and Interventional Radiology (more details: <https://www.sirm.org/en/>).
- Dataset of 349 CT images containing clinical findings of COVID-19 from 216 patients and more than 350 CT images of normal healthy person. The utility of this dataset was confirmed by a senior radiologist in Tongji Hospital,

Fig. 8 Training process using ResNet architecture (CNN-based) for COVID-19 classification

Wuhan, China. (More details: <https://github.com/UCSD-AI4H/COVID-CT>)

- Dataset of more than 1000 images containing anonymised human lung computed tomography (CT) scans with COVID-19-related findings, as well as without such findings. CT scans were obtained between 1 March 2020 and 25 April 2020 and provided by medical hospitals in Moscow, Russia (More details: <https://mosmed.ai/en/>).
- Dataset of more than 100 CT images from > 50 patients with COVID-19 case and other pneumonia case provided by SAL Hospital, Ahmedabad, India (more details: <http://www.salhospital.com/>).

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

On cloud training was used. The average time it took to train the system was 21 h. Figure 10 shows the environment setup in custom vision software of Microsoft azure. Figure 11 a, b and c shows the sample CT scan images of COVID-19 patient, normal healthy person and other viral pneumonia patient.

Experiment and result analysis

We built our database of almost 2200 images consisting of the following:

CT scan COVID-19 patient case images 800 (Fig. 11a), normal healthy person CT scan image 800 (Fig. 11b) and other viral pneumonia patient case images 600 (Fig. 11c). 600 images (200 COVID-19 case, 250 normal healthy case, 150 other viral pneumonia case) collected from various

Fig. 9 Sample Grad-Cam for COVID-19 patient

Fig. 10 Environment setup for collecting, labelling, training and testing the images

hospitals of Italy, China, Moscow and India (database sources mentioned earlier) were kept for testing. The trained model had never seen these images before. We already had all the information like COVID-19/normal healthy/other viral pneumonia 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 COVID-19 with any place and also to develop a model that 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 (Lalkhen and McCluskey 2008), which are given as

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

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

Fig. 11 Sample CT scan images of COVID-19 patient, normal healthy person and other viral pneumonia patient

Fig. 12 True positive and true negative cases detected correctly by the trained model

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

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 91%, sensitivity equal to 92.1% and specificity equal to 90.29% with TP = 200, TN = 400, FP = 43 and FN = 17.

Fig. 13 False positive and false negative cases detected wrongly by the trained model

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 12 a and b shows the true positive and true negative case as detected correctly by our trained model. Figure 13 shows the false positive and false negative cases that were detected wrongly by our trained model. Figure 14 shows some cases where model was little confused in classifying COVID-19 with two categories. i.e. other viral pneumonia and normal healthy case (shown in percentage of classification).

Some of the important outcomes of the experiment are as follows:

- 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

Fig. 14 Some confusing cases

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.
- Our proposed trained model (based on ResNet architecture and Grad-Cam) achieved a higher accuracy, i.e. 91% compared with the work done in Xu et al. (2020), who reported an overall accuracy of 87.6% in classifying chest CT scan images into three classes as healthy cases, other viral pneumonia case (respiratory syncytial virus (RSV), Influenza A) and COVID-19 case.

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, Moscow and India.

Training and testing were done using custom vision software based on machine learning techniques of Microsoft azure. An accuracy of almost 91% 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 for early detection of

coronavirus. Training and testing on more quality image datasets may further improve accuracy of the model.

Acknowledgments Author would like to acknowledge Dr. Raj Rawal (Critical Care Specialist, SAL Hospital, Ahmedabad-India) for helping with the medical terms and diagnosis of COVID-19 patients.

Funding: None

Declaration:

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 Moscow, Italy, China and India (Links are provided in the reference sections). These are repositories of anonymised images accessible locally for educational purposes and no identifiable information is stored or available.

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.

References

- Lalkhen A, McCluskey A (2008) Clinical tests: sensitivity and specificity. *Continuing education in anaesthesia critical care & pain* 8(6): 221–223. <https://doi.org/10.1093/bjaceaccp/mkn041>
- <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://github.com/UCSD-AI4H/COVID-CT> (Last accessed on 20th May 2020)
- <https://mosmed.ai/en/> (Last accessed on 5th June 2020)
- <https://radiopaedia.org/articles/COVID-19-3> (Last accessed on 15th March 2020)
- <https://www.sirm.org/en/> (Last accessed on 28th May 2020)
- <http://www.salhospital.com/> (Last accessed on 2nd June 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)
- Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, et al (2020) “Lung infection quantification of COVID-19 in CT images with deep learning.” *arXiv preprint arXiv:200304655*
- Xu X, Jiang X, Ma C, Du P, Li X, Lv S, et al. (2020) “Deep learning system to screen coronavirus disease 2019 pneumonia”. *arXiv preprint arXiv:200209334*

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.