Deep Learning Approach to Recognition of Novel COVID-19 Using CT Scans and Digital Image Processing

H.M.K.K.M.B. Herath (kasunherathlive@gmail.com)
The Open University of Sri Lanka https://orcid.org/0000-0002-1873-768X
G.M.K.B. Karunasena
The Open University of Sri Lanka
S.V.A.S.H. Ariyathunge
The Open University of Sri Lanka
H.D.N.S. Priyankara
The Open University of Sri Lanka
B.G.D.A. Madhusanka
The Open University of Sri Lanka
H.M.W.T. Herath
University of Moratuwa Sri Lanka
U.D.C. Nimanthi
National Eye Hospital of Sri Lanka

Method Article

Keywords: Artificial Intelligence (AI), COVID-19 Pneumonia, CT-Scan, Deep Learning, Ground-Glass Opacity (GGO), Machine Vision, SARS-CoV-2

Posted Date: June 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-646890/v1

Abstract

COVID-19 was announced as a global pandemic by the World Health Organization (WHO) in March 2020. With more than 31.3 million confirmed cases and over 965 thousand deaths recorded as of September 2020, it has inflicted catastrophic damage worldwide. The aim of this study is to develop an algorithm based on artificial intelligence (AI) and image processing techniques to identify COVID-19 patients with the aid of CT chest scan images. This study used a CT scan image dataset that is publically available for the researchers at Kaggle. We randomly extracted 27% of positive CT (pCT) images and 11% of negative CT (nCT) images from the original dataset. In the testing process, 120 of the test subjects in both nCT and pCT were used to validate the algorithm. Based on the experimental findings, the proposed COVID-19 detection algorithm shows promising results for the identification of COVID-19 patients with 90.83% accuracy at an average precision of 0.905.

1 Introduction

Coronavirus is a large family of viruses that can cause a human being to develop a serious illness. The first reported major epidemic was Severe Acute Respiratory Syndrome (SARS) [1] in 2003, while the second severe outbreak of Middle East Respiratory Syndrome (MERS) [2, 15] in Saudi Arabia began in 2012. The latest outbreak of coronavirus disease was announced in late December 2019. This new virus is very infectious and has spread globally rapidly. On January 30, 2020, as it had spread to 18 countries, the World Health Organization (WHO) declared this outbreak a Public Health Emergency of International Concern (PHEIC) [3]. This virus was named 'COVID-19' by the World Health Organization on February 11, 2020 [4]. As of September 2020, the WHO reported that 31.3 million confirmed cases and over 965 thousand deaths have been registered in 213 countries.

Figure 1 shows confirmed cases of global COVID-19 as of September 2020. The disease has spread rapidly around the globe since it was first identified and has become an international concern. An analysis performed by Jiang et al. [5] found that COVID-19's death rate is 4.5% worldwide. In the age group of 70–79 years, the death rate for patients is 8.0%, while 14.8% for patients over 80 years. Patients over 50 years of age with chronic diseases are at the highest risk and it is critically important to find a way to detect illness before getting into serious conditions.

As the COVID-19 epidemic has become a global pandemic, real-time analysis of epidemiological data is required to prepare society for better disease response plans. COVID-19 belongs to the SARS-CoV and MERS-CoV families, where symptoms of the common cold to severe respiratory diseases, causing trouble breathing, exhaustion, fever, and dry cough, start at the initial level. Real-time Reverse Transcription-Polymerase Chain Reaction or also known as RT-PCR is the latest approach used to make a definitive diagnosis of SARS-CoV-2 infection [6]. PCR testing was found to have a high specificity (Sp) but rather low sensitivity (Sn) with a reported positive rate of only 38%~57%. In addition to etiological laboratory confirmation, Clinical Features (CFs) and chest Computed Tomography (CT) imaging include other key diagnostic elements that could facilitate the identification of COVID-19 pneumonia.

Early identification of patients with COVID-19 pneumonia for timely treatment is crucial to contain the spread, particularly in epidemic regions. According to information shared by the Radiological Society of North America (RSNA), X-ray and CT images of a Chinese person dead by COVID-19 showed the damages done to the human lungs. A research team led by Lucas [7] at The University of São Paulo demonstrated the chest imaging finding of COVID-19 on different modalities such as Chest Radiography (CXR), Computed Tomography (CT), and Ultrasonography. According to them, chest CT is the main imaging method used in the assessment of COVID-19 pneumonia. A structured chest CT report standardizes imaging results and optimizes contact with the prescribing physician, making it a valuable tool in the pandemic scenario. In addition, the CT imaging properties of infected lungs include Ground-Glass Opacity (GGO) and severity-correlated consolidation. In Hubei Province, China, CT scans have been used widely and on presentation in an attempt to rapidly detect, isolate, and control the spread of the epidemic.

Many studies have documented a high degree of chest CT sensitivity in the diagnosis of COVID-19 pneumonia. Previous studies have shown that the most common CT characteristic of COVID-19 pneumonia is the presence of multifocal Ground-Glass Opacity (GGOs). Figure 2 displays the CT scan image of a COVID-19 patient. Arrowheads reveal the recognizable hazy area on the outer edges of the lungs. As per the description, Ground-Glass Opacities refer to the distorted presence of the lungs in imaging experiments, almost as though parts were obscured by Ground-Glass. This may be due to the fluid filling of pulmonary airspace, the collapsing of airspace, or both. This is a trend that can be seen while the lungs are sick. Regular lung’s CT scans look black; rare chest CTs with GGOs reveal lighter colored or gray spots. Consolidation refers to the saturation of fluids or other inflammatory products in pulmonary airspace. Pleural effusion refers to abnormal fluids that form in the spaces surrounding the lungs.

PCR tests are taking time to diagnose COVID-19 patients and the test results appear to be of low accuracy compared to CT scan tests. However, CT scans can be used as a simple and quick way of categorizing patients into "probably positive" and "probably negative" cohorts. As the hospital admission rate of COVID-19 patients increases, the PCR test is not appropriate. Nowadays, tools for the identification of COVID-19 patients with high efficacy and accuracy are essential. Due to the poor contrast of infection regions of CT images and the large differences in both the shape and location of the lesions in different patients, the delineation of the infection regions in CT scans in the chest is very difficult for the physician. Image processing techniques may open new pathways to describe the state of the lungs using CT scans. The objective of this study is to develop a deep learning algorithm to detect COVID-19 patients using CT chest scan images and validate results for both COVID-19 positive and healthy test subjects.

2 Related Works

Scientific, technological knowledge and resources have been widely used to prevent COVID-19 globally. The number of studies related to the novel COVID-19 is increasing daily. Researchers have recently used imaging patterns on chest CT to detect COVID-19 infection. Research contribution in the field of machine vision and artificial intelligence to prevention novel COVID-19 is described in this section.

Previous studies led by Xu and his research team categorized CT scan images of COVID-19 patients in three groups as healthy cases, Influenza viral pneumonia, and COVID-19 [8]. This study used 175 images of healthy people, 224 images of patients with Influenza-A Pneumonia, and 219 images of patients infected with the coronavirus. The overall accuracy of 86.7% was observed using the 3D-deep learning model.

Shan et al. [9] have developed a method focused on a deep learning mechanism for the segmentation and quantification of contaminated regions and the entire lung using CT images in the chest. A total of 249 COVID-19 patients and 300 new COVID-19 patients have been used for validation in their study. They used the Dice similarity 2 coefficient concepts and achieved 91.6% accuracy.

Sachin Sharma [10] from the Institute of Advanced Research of India has engaged with a study about the role of machine learning techniques in obtaining important insights, such as whether a lung CT scan is a first screening/alternative test for RT-PCR. Training and testing have been carried out using custom vision software based on Microsoft Azure machine learning techniques. The accuracy of nearly 91% has reached, although some false indicators were found in their analysis.

Harmon [11] and her research team from the USA have shown that a number of deep learning algorithms have been trained in a multi-national cohort of 1,280 patients to locate parietal pleura/lung parenchyma followed by a COVID-19 pneumonia classification. They achieved 90.8% accuracy, with 84% sensitivity and 93% specificity.

Xavier [12] has engaged in a study to evaluate the performance of Artificial Intelligence methods to detect COVID-19 using chest X-Rays and CT scan images. A total of 363 patients have been used by combining two different data sources. 191 patients have COVID-19 positive and the rest of them were healthy subjects. The accuracy of the proposed system has reached 90.9% for the 121 testing samples.

Table 1
Advantages and disadvantages of related works

Ref	Advantages	Disadvantages
[8]	High accuracy achieved (86.7%), The promising results of an additional diagnostic tool for frontline clinical physicians.	The CT manifestation of COVID-19 contrasted only with that of IAVP. In this analysis, A limited number of model samples were used. A limited range of testing and training samples utilized.
[9]	High accuracy (91.6%) achieved a large dataset for training and testing has used.	Validation data of CT datasets have been collected in one location, which may not be indicative of all COVID-19 patients in other geographical regions.
[10]	CT scan images obtained from various geological sites (Italy, China, Moscow, and India), As the model based on the CT chest images showed strong results in terms of precision and time-consuming.	A high false detection rate has been observed in some experiments. A limited range of test samples has been used for the study in each geological location.
[11]	High accuracy showed (90.8%), Larger dataset has been used for training and testing. The multinational dataset has been used to cover the different geological locations of the world.	Model training has been limited to patients with positive RT-PCR testing and COVID-19 related pneumonia on chest CT.
[12]	The results have been achieved with high accuracy (91.8%), Larger dataset has been used for the study.	They also state that it is unknown if the tested procedures may be used to diagnose asymptomatic patients.

Table 1 depicts the advantages and disadvantages of previous related studies referred to. From a systematic study, it has been found that CT images of the chest can be used for the early classification of COVID-19 infected patients. Therefore, the Convolutional Neural Networks (CNN) model was used in this study to distinguish COVID-19 patient's identification using the CT scan images of the chest.

3 Methodology

Ground-Glass Opacities (GGO) [13], consolidation, and pleural effusion are the characteristics that are seen as the primary features used in the CT scan picture of a COVID-19 patient. This study mainly focused on detecting the GGO features based on the presence of the COVID-19 infection of the human lungs. The proposed methodology of the system consists of two separate sections, such as the selected CT scan image dataset (Sect. 3.1) and the structure of the COVID-19 artificial intelligence (AI) algorithm (Sect. 3.2). The MATLAB development environment was used to develop the proposed algorithm.

3.1 CT Image Dataset

This study used the CT scan image dataset from the Kaggle [14] which is publically available for the researchers. The dataset consists of three types of CT images obtained from Union Hospital (HUST-UH) and Liyuan Hospital (HUST-LH). The dataset consists of non-informative CT (NiCT) images, positive CT

(pCT) images, and negative CT (nCT) images. We were randomly extracted 27% of pCT and 11% of nCT data for this study. In the testing process, 120 of the extracted data in nCT and pCT were used to validate the method. Table 2 depicts the dataset description used in the study. All the images in the dataset were originally sized to 512×512 pixels.

Table 2
Description of the COVID-19 positive and negative CT scan image dataset

CT Image Type	Data used	Description
Positive CT – pCT	27% of data from original dataset	Imaging features are associated with the COVID-19 Pneumonia
Negative CT – nCT	11% of data from original dataset	Imaging features in both lungs were irrelevant to the COVID-19 Pneumonia

In terms of lung changes, the presence of various types of lungs was observed in COVID-19 positive patients. Figure 3 illustrates the CT chest scan images of the COVID 19 positive and healthy test subject's lungs.

3.2 Proposed CNN Architecture

As the initial phase of this study, the chest CT scan images of COVID-19 subjects and normal healthy subjects are taken and stored in the computer. Then we have performed some image pre-processing steps, such as image cropping and image resizing to extract effective pulmonary regions before using the dataset.

Convolutional Neural Networks (CNN) is a versatile method that is commonly used for image classification. The hierarchical structure and the powerful functionality of image extraction render CNN a complex model for image classification. The proposed CNN architecture is composed of two stages: a feature learning stage and a classification stage as shown in Fig. 4.

The developed feature learning step consists of two convolutional layers and two pooling layers. The first convolution layer includes a 3×3 convolutional filter for initial feature extraction. Then resultant features passed into the first pooling layer which consists of 2×2 max-pooling filters. Then, the extracted features from the first convolution and pooling passed to the second convolution and pooling layers. Furthermore, the second convolution layer consists of 3×3 convolutional filters and the pooling layer consists of 2×2 max-pooling filters. In the classification stage, the feature score matrix passed into a fully connected layer which consists of fully connected three neural layers. Each layer includes 500, 100, and 2 artificial neurons. Finally, the softmaxLayer was used to obtain probability and classify input samples to indicate whether the test subjects are COVID positive or negative. In this model, we have used a 200×200 size input layer.

The Model Hyperparameters are properties that control the whole training process. These include the variables that determine the structure of the network and the variables that determine how the network is trained. The Stochastic Gradient Descent with Momentum (SGDM) optimizer was used as the solver of the training network. The ReLU activation function was used to activate the nodes. The initial learning rate of 0.1 and 0.01 learn rate drop factor was observed at the 20 maximum epochs. We used a mini-batch with 20 observations in each iteration.

4 Results And Analysis

The proposed design was tested with 120 randomly selected nCT and pCT chest images from the extracted dataset. Figure 5 illustrates the COVID-19 positive test subjects whose lungs have filled with hazy areas. Hazy areas suggested that patients have COVID-19 infection in the body at what level. These subjects are identified as the COVID-19 positive patients by the clinical trials.

Figure 6 depicts the sample of healthy test subjects used in this study. According to the images, lungs are observed to be clear and detect less gray spots. Detection of less gray spots suggested that the test subject is negative from COVID-19. These healthy test subjects are identified as COVID-19 negative by the clinical trials.

For the training, a classification model was developed. A total of 20 epochs and 2000 iterations were undertaken during the training process in order to achieve optimal model parameters. The accuracy curve of the training process is shown in Figure 7. Based on the training results, the average classification accuracy for each individual mini-batch was 94.25% and the classification accuracy for each individual mini-batch was reached to the maximum at epoch 8.

According to the mini-batch loss curve shown in Figure 8, the mini-batch loss for multi-class classification decreased from 0.695 to 0.0977 at the end of the 20 epochs Based on the test results, positive subjects for COVID-19 were classified with a range of 0.65-1.00 probabilities and healthy subjects (COVID-19 negative subjects) ranged from 0.10-0.40 probabilities. Figure 9 illustrates the experiment results for 120 test subjects of COVID-19 positive and healthy subjects (COVID-19 negative).

A confusion matrix or also known as an error matrix is a representation of the performance of an algorithm. The confusion matrix is commonly used in the area of machine learning, typically supervised learning. The entries in the confusion metrics were calculated from the coincidence matrix by using the following hypothesis,

True Negative (TN) is the number of correct predictions that an instance is negative.

True Positive (TP) is the number of correct predictions that an instance is positive.

False Positive (FP) is the number of incorrect predictions that an instance is positive.

False Negative (FN) is the number of incorrect predictions that an instance is negative.

Figure 10 illustrates the confusion matrix of the results. For the development of the confusion matrix, the following parameters were identified by considering actual and observed values.

True Positives (TP) : 56
True Negatives (TN) : 53
False Positives (FP) : 04
False Negatives (FN) : 07

Equations 1 and 2 are the accuracy and precision of the test results based on the data extracted from the confusion matrix.

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

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

Fig. 10. Confusion matrix of the test results.

Table 3. Accuracy, precision and recall values of the results

Class	n (truth)	n (classified)	Accuracy	Precision	Recall	F1 Score
1	63	60	90.83 %	0.93	0.89	0.91
2	57	60	90.83 %	0.88	0.93	0.91

According to Table 3, the accuracy of the proposed design was 90.83% with 0.91 of average precision. Precision for both classes (class 1 and class 2) were 0.93 and 0.88 observed.

$$MAE = \frac{\sum_{t=1}^{N} |Y_t - F_t|}{N} \quad (3)$$

Mean Absolute Error (MAE) is a calculation of errors between paired measurements that express the same phenomena. Equation 3 represents the relationship between real data and the prediction data. The best Mean Absolute Error of a system is considered to be less than 0.200 and the MAE of the proposed system was 0.095. Therefore, the Lower MAE validates the accuracy of the proposed model for the identification of COVID-19 using chest CT scan images. Root Mean Square Error (RMSE) of the system was 0.149 calculated. Lower RMSE suggested the higher accuracy of the proposed algorithm.

5 Conclusion

Coronavirus disease outbreak 2019 (COVID-19) is a worldwide epidemic that has a significant effect not only on the health of peoples but also on the global economy. Pneumonia caused by coronavirus reveals a common hazy spot on the outer edges of the lungs, which indicates a trend such that machine learning methods can be used for early coronavirus identification.

In this paper, we addressed the role of artificial intelligence (AI) techniques in identifying the novel COVID-19 using CT chest scan images of corona patients. Training and testing were carried out using the dataset published by Ning and his research team at the Huazhong University of Science and Technology in China. The model based on the CT scan showed good results in terms of accuracy and as it took less time to identify the COVID-19. The results show that the detection of COVID-19 is possible using the CT images with deep neural network methodology. The proposed algorithm was 90.83% accurate with the 0.095 Mean Absolute Error (MAE). Moreover, patients with lung disorders are identified as positive cases of COVID-19. As a result, an error may be developed in the other lung condition that has hazy regions of the lungs. The probability of the CT image result indicates COVID-19 detection at what level. Higher probability means larger hazy spots detect in the lungs which means the subject has a higher level of COVID-19 virus.

This study found some drawbacks, such as validation data of the CT dataset collected from one geographical region, which may not be representative of all COVID-19 patients in other geographic areas. In our future work, we will extend the algorithm to quantify the severity of other pneumonia using transfer learning and to validate the results using data obtained from many geographical regions of the world.

References

1. Cheng, V.C.C., Lau, S.K.P., Woo, P.C.Y., Yuen, K.Y.: Severe Acute Respiratory Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection. Clinical Microbiology Reviews. 20, 660–694 (2007)
2. Ramadan, N., Shaib, H.: Middle East Respiratory Syndrome Coronavirus (MERS-CoV): A Review. Germs. 9, 35–42 (2019)
3. Kamradt-Scott, A.: A Public Health Emergency of International Concern? Response to a Proposal to Apply the International Health Regulations to Antimicrobial Resistance. PLoS Medicine. 8. 1–5 (2011)
4. Punn, N.S., Sonbhadra, S.K., Agarwal, S.: COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms. MedR. XIV. 1–10 (2020)
5. Jiang, F., Deng, L., Zhang, L., Cai, Y., Cheung, C.W., Xia, Z.: Review of the Clinical Characteristics of Coronavirus Disease 2019 (COVID-19). Journal of General Internal Medicine. 35. 1545–1549 (2020)
6. Jin, Y.H., Cai, L., Cheng, Z.S., Cheng, H., Deng, T., Fan, Y.P.: A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version). Military Medical Research. 7, 1–23 (2020)
7. Farias, L., Fonseca, E., Strabelli, D., Loureiro, B., Neves, Y., Rodrigues, T., Chate, R., Nomura, C., Sawamura, M.V.Y., Cerri, G.: Imaging findings in COVID-19 pneumonia. Clinics (Sao Paulo, Brazil). 75, 1–9 (2020)
8. Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Chen, Y., Su, J., Lang, G., Li, Y., Zhoa, H., Xu, K., Ruan, L., Wu, W.: Deep learning system to screen coronavirus disease 2019 pneumonia. arXiv. 2020, 1–29 (2020)

9. Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z. and Shi, Y.: Lung infection quantification of COVID-19 in CT images with deep learning. arXiv. 2020, 1–19 (2020)
10. Sachin, S.: Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients. Environmental science and pollution research international. 27, 1–29 (2020)
11. Harmon, S.A., Sanford, T.H., Xu, S.: Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun. 11, 1–7 (2020)
12. Xavier, B.: Computer-aided covid-19 patient screening using chest images (X-Ray and CT scans). medRxiv. 2020, 1–19 (2020)
13. Schmitt, W., Marchiori, E.: Covid-19: Round and oval areas of ground-glass opacity. Pulmonology. 26, 1–5 (2020)
14. Ning, W., Lei, S., Yang, J., Cao, Y., Jiang, P., Yang, Q., Zhang, J., Wang, X., Chen, F., Geng, Z., Xiong, L., Zhou, H., Guo, Y., Zeng, Y., Shi, H., Wang, L., Xue, Y., Wang, Z.. iCTCF: an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia. Europe PMC. 2020, 1–37 (2020)
15. Herath, H. M. K. K. M. B., Karunasena, G. M. K. B., & Herath, H. M. W. T. (2021). Development of an IoT Based Systems to Mitigate the Impact of COVID-19 Pandemic in Smart Cities. In Machine Intelligence and Data Analytics for Sustainable Future Smart Cities (pp. 287–309). Springer, Cham.

Competing Interests

The authors declare no competing interests.

Figures

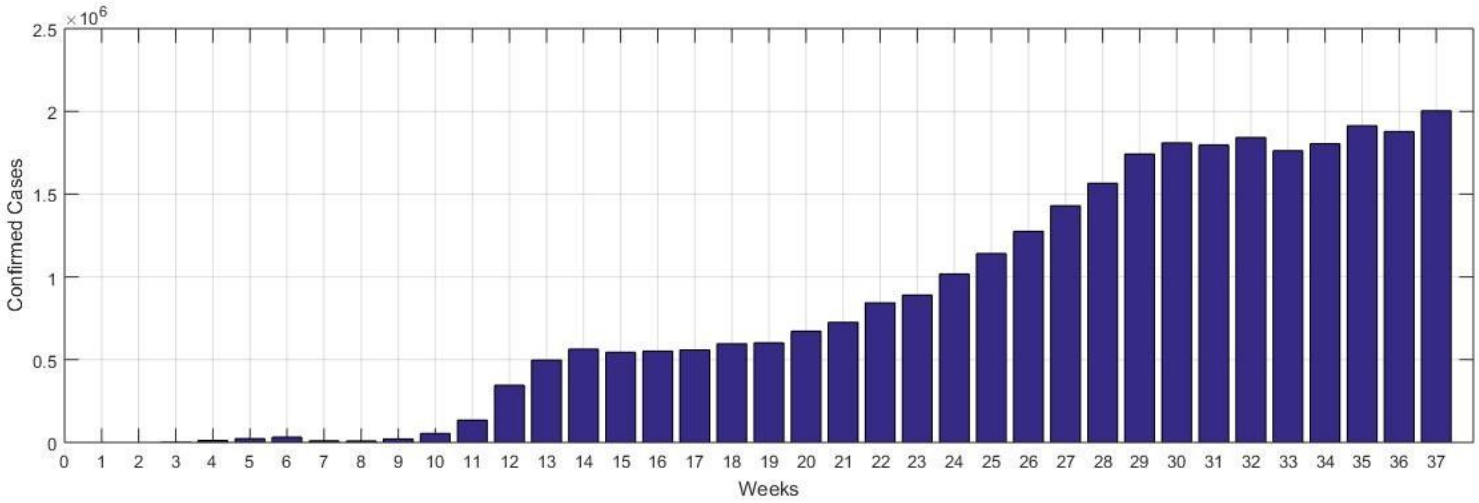

Figure 1

Global COVID-19 confirmed cases as of September 2020.

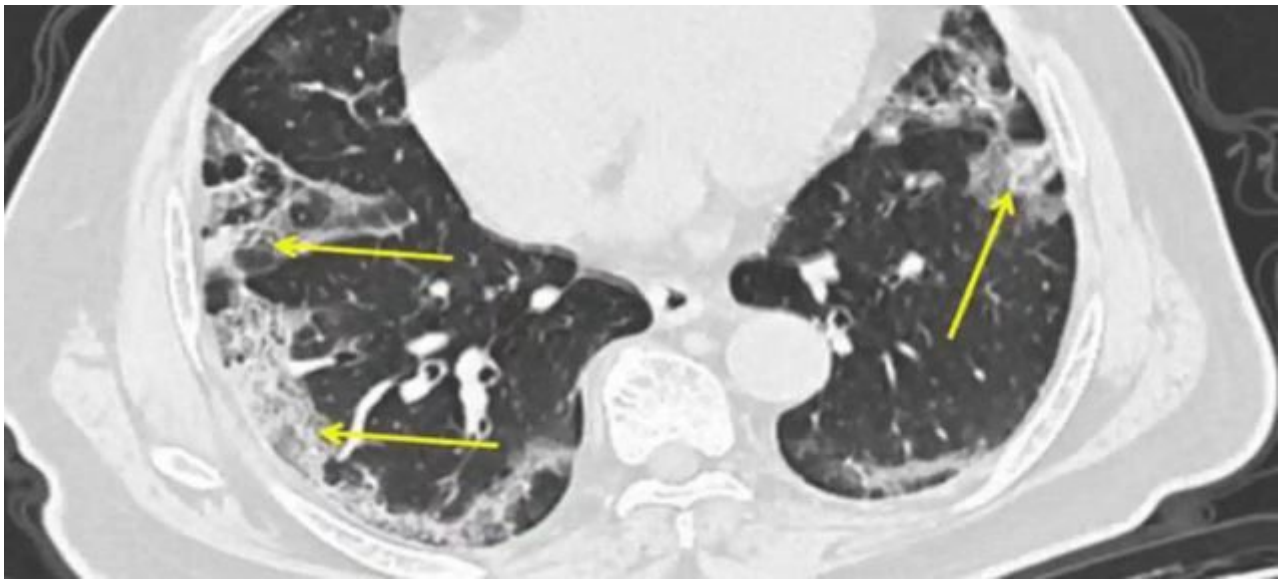

Figure 2

CT scan image of a patient with severe COVID-19, (Photograph: Mount Sinai Hospital)

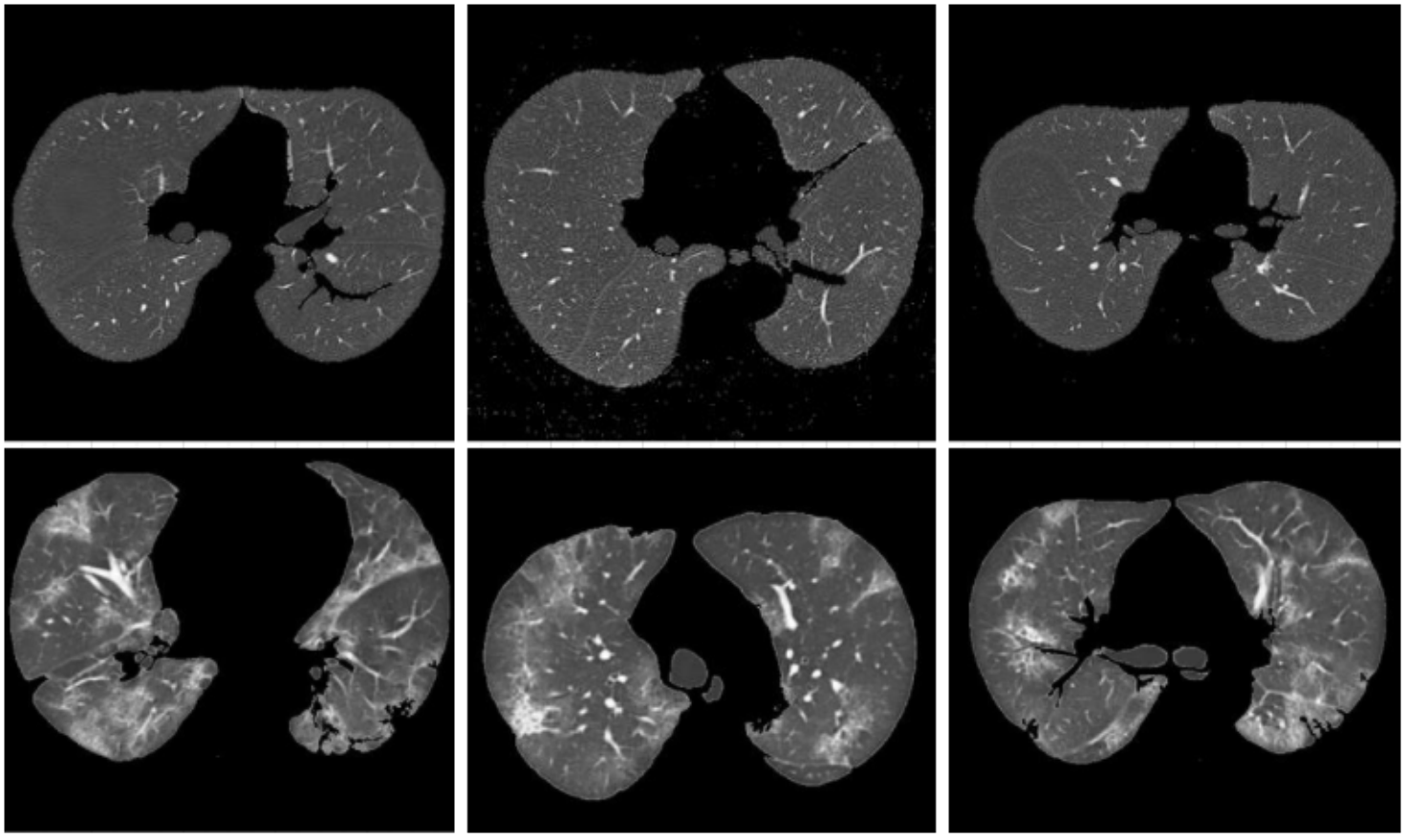

Figure 3

CT scan images of COVID-19 negative (top row) and positive (bottom row) patients

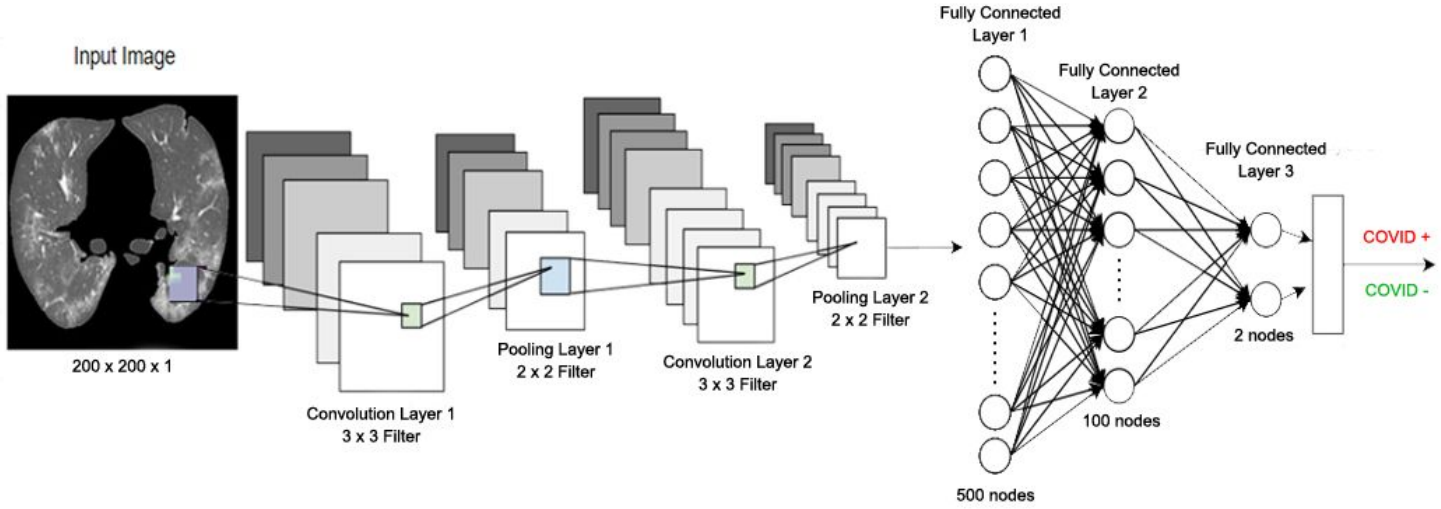

Figure 4

Proposed CNN architecture of the COVID-19 detection algorithm

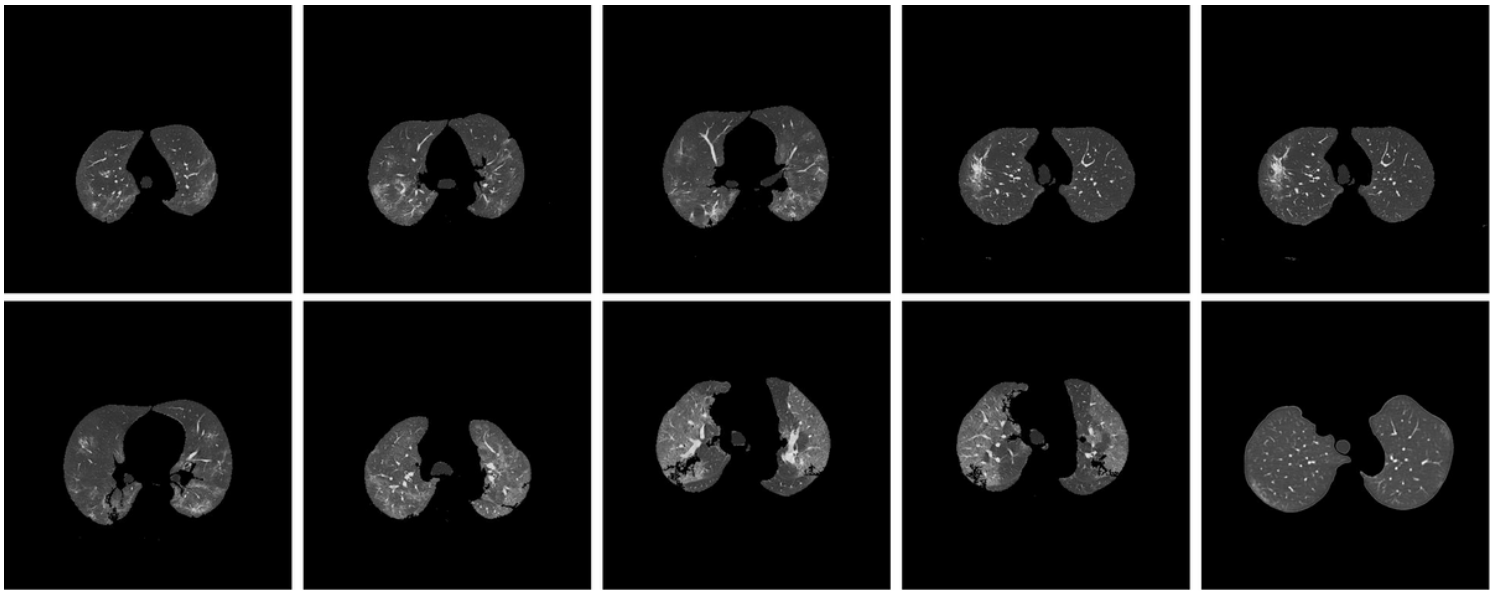

Figure 5

CT chest scan images for COVID-19 positive test subjects

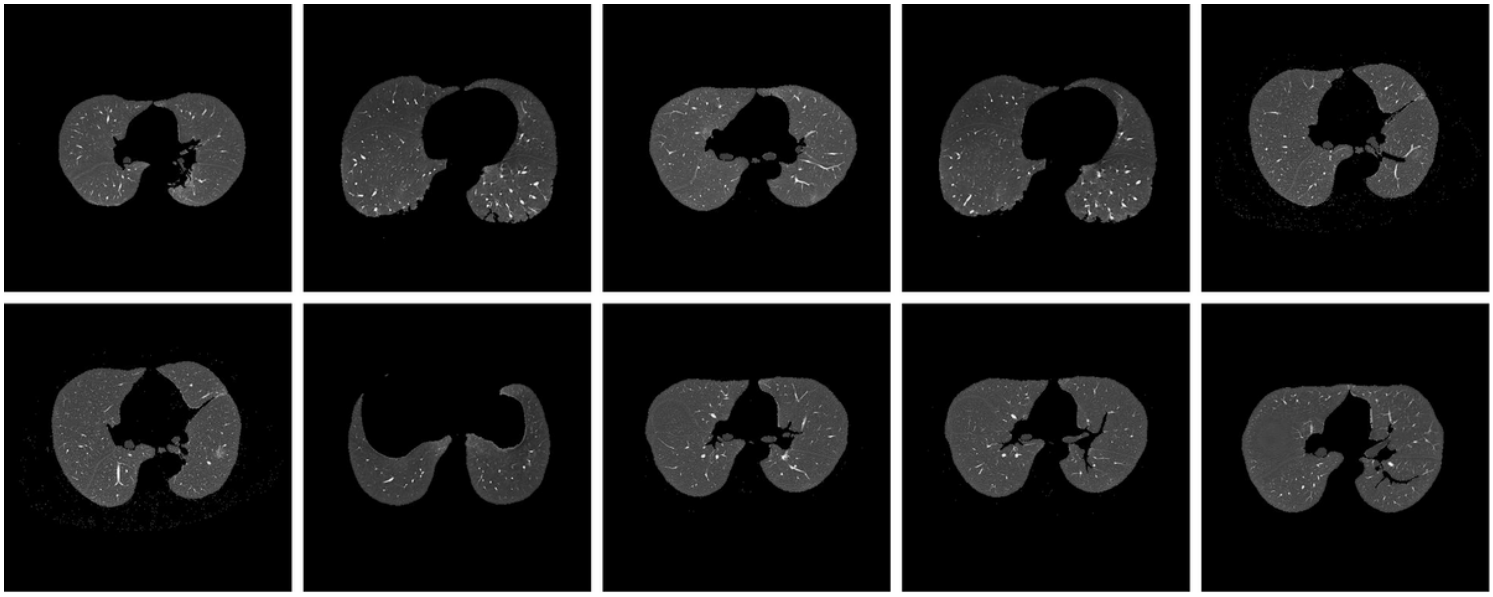

Figure 6

CT chest scan images for healthy test subjects (COVID-19 negative test subjects)

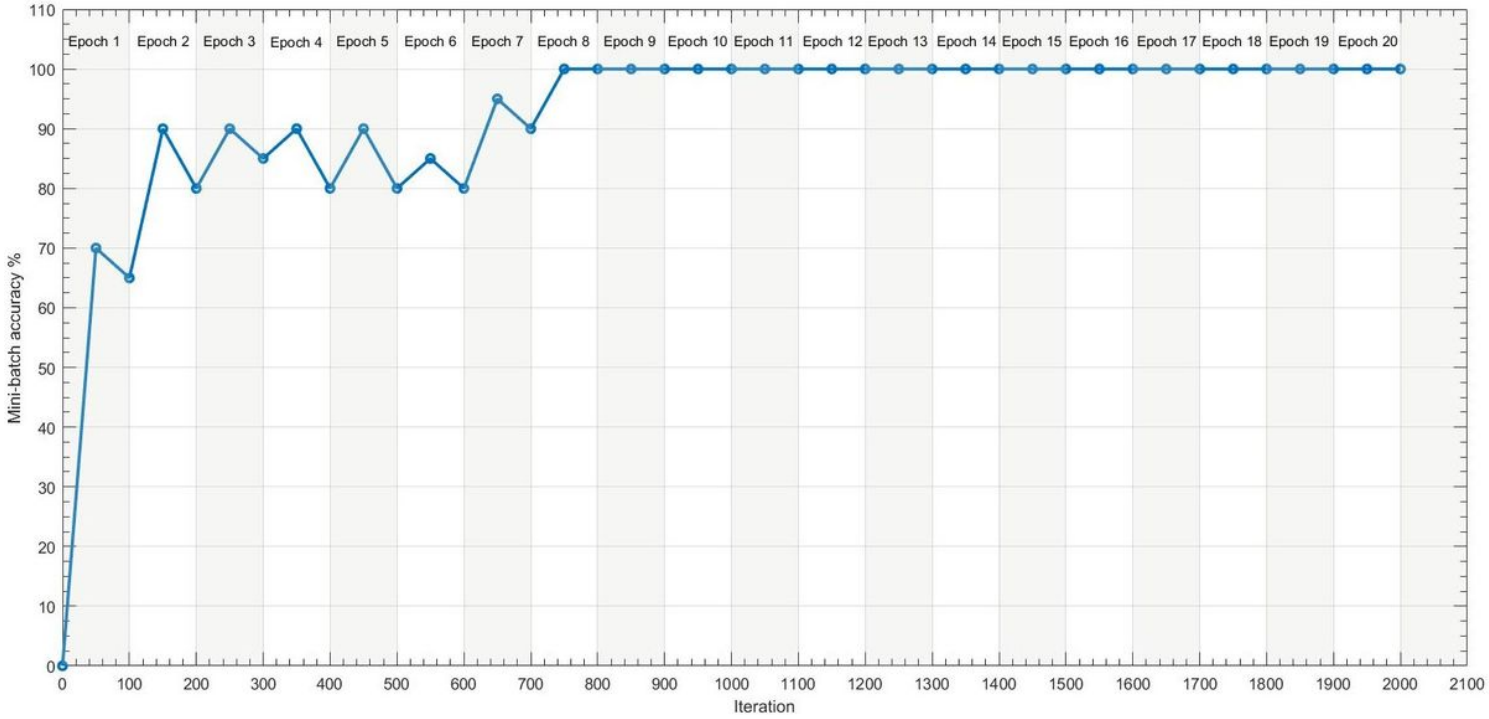

Figure 7

Mini-batch accuracy of the model training for 20 epochs

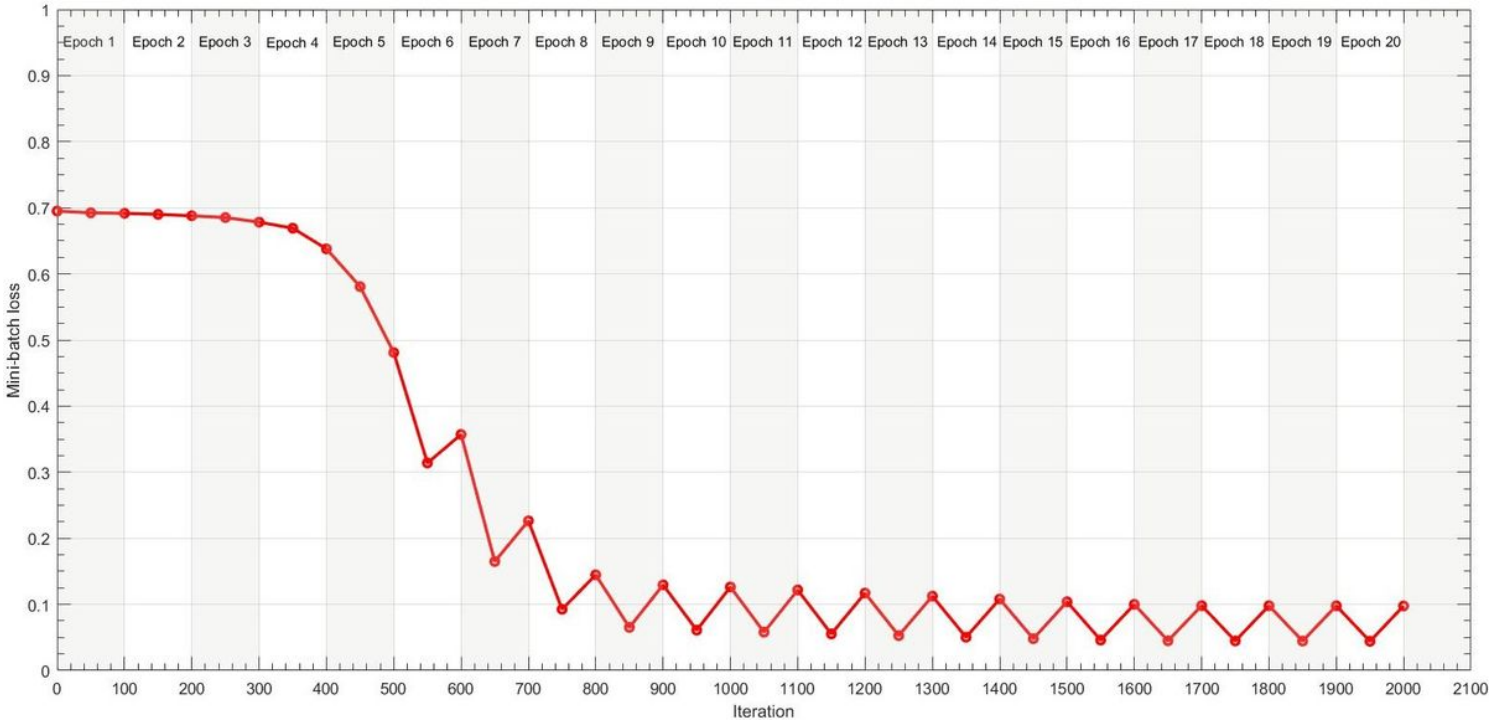

Figure 8

Mini-batch loss of the model training for 20 epochs

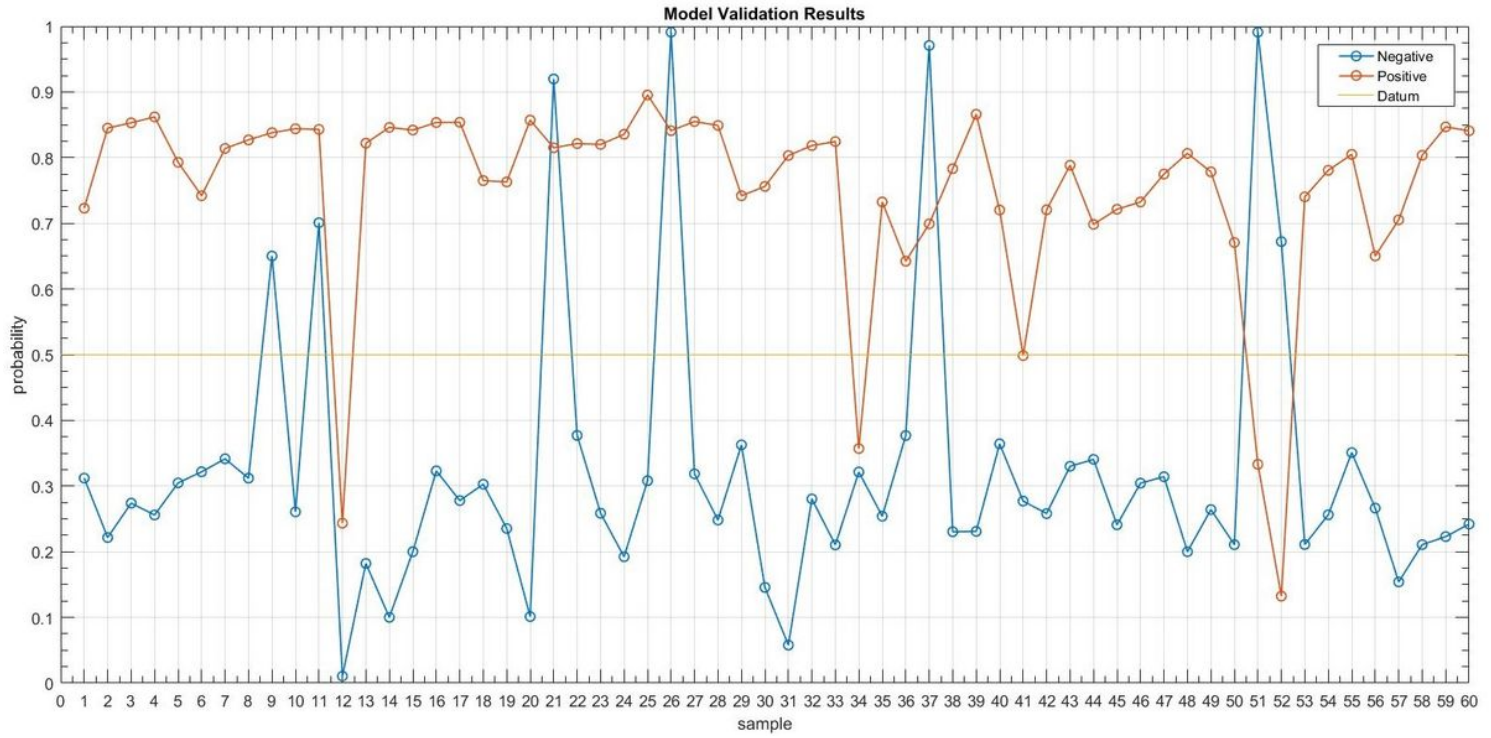

Figure 9

Model validation results of the COVID-19 positive and healthy test subjects

N=120	Actual: Yes	Actual: No	
Predicted: Yes	TP = 56	FP = 04	60
Predicted: No	FN = 07	TN = 53	60
	63	57	

Figure 10

Confusion matrix of the test results