# Network for subclinical prognostication of COVID 19 Patients from data of thoracic roentgenogram: A feasible alternative screening technology

**Authors:** Akash Bararia[1], Abhirup Ghosh[2], Chiranjit Bose[3], Debarati Bhar[4,$]

**Affiliations:**

[1]Department of Haematology, Tata Medical Center, Kolkata

[2]My Jio-Jio.com, Jio Platforms Ltd, Navi Mumbai

[3]Dept. of Endocrinology and Metabolism, IPGME&R and SSKM Hospital, Kolkata

[$4]Department of Medicine, R. G. Kar Medical College and Hospital, Kolkata.

**[$]Corresponding Author: Dr. Debarati Bhar,** Department of Medicine, R. G. Kar Medical College and Hospital, Kolkata. Email: debaratibhar1234@gmail.com

**Credit authorship contribution statement:**

**Akash Bararia:** Conceptualization, Data sorting and mining, Investigation**,** Writing - original draft.

**Abhirup Ghosh:** Data curation and augmentation, Computational analysis, Investigation**,** Writing - original draft.

**Chiranjit Bose: Statistical analysis,** Writing - original draft, Editing, Dataset curation.

**Debarati Bhar:** Supervision, Clinical Correlation and analysis, Draft proofreading.

**Conflict of interest**: None

**Ethical Approval**: None

**Funding:** None

**Competing interest:** No benefits in any form have been received or will be received from a

commercial party related directly or indirectly to the subject of this article.

**Acknowledgement:** We would like to thank all the patient and normal individuals whose data has been incorporated into the datasets. We would also like to thank Prof. Dr. Satinath Mukhopadhyay, Professor at the Institute of Post Graduate Medical Education and Research (IPGME&R) in Kolkata for his motivational support. We would like to thank The Department of Science & Technology for financially supporting Mr. Chiranjit Bose via DST-INSPIRE FELLOWSHIP (Inspire fellow registration number: IF180004). Alongside We would also like to thank our respective employer's. We would also like to thank Joseph Paul Cohen, Director of the Institute for Reproducible Research, for GitHub dataset and Paul Mooney, Developer Advocate at Kaggle for the Kaggle dataset development. Last but not least we would also like to thank all doctors and health support teams working across globe to fight against coronavirus pandemic.

**Abstract: Background and Study Aim:** COVID 19 is the terminology driving people's life in the year 2020 without a supportive globally high mortality rate. Coronavirus lead pandemic is a new found disease with no gold standard diagnostic and therapeutic guideline across the globe. Amidst this scenario our aim is to develop a prediction model that makes mass screening easy on par with reducing strain on hospitals diagnostic facility and doctors alike. For this prediction model, a neural network based on Chest X-ray images has been developed. Alongside the aim is also to generate a case record form that would include prediction model result along with few other subclinical factors for generating disease identification. Once found positive then only it will proceed to RT-PCR for final validation. The objective was to provide a cheap alternative to RT-PCR for mass screening and to reduced burden on diagnostic facility by keeping RT-PCR only for final confirmation.

**Methods:** Datasets of chest X-ray images gathered from across the globe has been used to test and train the network after proper dataset curing and augmentation. **Results:** The final neural network-based prediction model showed an accuracy of 81% with sensitivity of 82% and specificity of 90%. The AUC score obtained is 93.7%. **Discussion and Conclusion:** The above results based on the existing datasets showcase our model capability to successfully distinguish patients based on Chest X-ray (a non-invasive tool) and along with the designed case record form it can significantly contribute in increasing hospitals monitoring and health care capability.

**Keywords:** Covid-19 prediction model, neural network, chest X-ray based Covid-19 diagnosis, machine learning prediction model, alternative to RT-PCR.

## I. Introduction:

Towards the completion of a semi-annual tenure beginning from the first reported case of COVID-19 on 29th January, 2020 in India, there has been a substantial rise in the number of new cases and associated death rates across the country. Amongst 21 countries, a total of 9996 cases has been reported as on 30th January, 2020 [1]. COVID-19 is an extremely contagious disease capable of producing a global threat to health based on benchmark criteria such as prevention, diagnosis and treatment [2]. According to WHO, 16,341,920 confirmed cases has been reported as on 30th July, 2020 out of which deaths occurred in 650,805 cases. In India alone, 1,531,669 confirmed infected cases have been reported out of which 34,193 died as on 30th July, 2020.

The SARS-COV2 viral disease has similarities with influenza in the early stages, which makes it very difficult to diagnose. On the contrary, late diagnosis can lead to the death of the subject as well as spread of infection. During an approach to develop a mass screening

procedure, a lot was accessed about the quantum of errors that might be there in the current molecular as well as the serological procedures like RT-PCR and antibody tests. Since none of the above tests are accurate and in the scenario of the non-availability of a gold-standard procedure, the choice of test is based on its sensitivity and specificity. Country-wide RT-PCR based approach to detect viral RNA is currently being used [3, 4].

Owing to the rise in COVID-19 cases across the globe, systemic utilization of healthcare resources is of crucial importance. Machine learning technologies can be a significant strategy in this situation for proper futuristic prediction that can help the administrators to plan accordingly in terms of health care resources and its associated support mechanisms. As the onset of disease is primarily in the lung of the infected person, so images of the lung could give us a quicker and better predictive approach. But to make it quicker, we need to use artificial intelligence and machine learning as a front-line analytical tool of the images [5-7].

Our main goal in this work is to develop a neural network-based prediction model that is capable of distinguishing the COVID-19 viral infected patients from normal people as well as from bacterial pneumonia based on the images of thoracic roentgenogram. The model takes an input of the X-ray image and then analyses it to state the outcome. The benefit of this procedure is based on user flexibility and associated facts: X-ray test is cheap, very fast image development time, availability of the instrument in most clinics and hospitals across the country. The additional benefits include scanning of the x-ray plate and uploading it into the network doesn't always require skilled manpower or medical personnel, hence it is effective in mass screening and helps to reduce burden on skilled hospital labour and prevents overcrowding of hospitals. Associatively a Case record form (in Supplementary file) is developed that has certain essential input spaces along with the network predicted result which needs to be filled and can be done by anyone without prior training. Based on

collective outcome of the rest of inputs and network result, an initial screening is possible and then, if needed, the patient can be sent to doctors based on the government and hospital workflow guidelines. A schematic representation of the combined model is stated in figure 1.

## II. Materials and Methods:

### II.I Data source/Dataset Development and structuring.

During the course of this study, all data gathered is in the form of jpeg image files of thoracic roentgenograms. These files have been acquired from different sources and have different sizes. The main two main sources of Chest X-ray images were taken from GitHub (https://github.com/ieee8023/covid-chestxray-dataset) and Kaggle (https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia). These two datasets have been fused and their data used to create our final dataset comprising of equal proportions of COVID-19, bacterial, other viral and normal chest X-ray images [8, 9]. We have also assumed that COVID-19 cases and viral pneumonia have similar symptoms and so these images are merged into a single feature. After acquiring the images, we have created a pandas data-frame of the image path and its corresponding label. Thus, our final dataset is a pandas data frame containing 3840 images, which are equally distributed.

### II.II Data cleaning and Data Augmentation.:

Preliminary data cleaning steps involved validating the image's readability and conversion. Since most of the images obtained were already pre-cleaned and verified, this step ran relatively fast and only a small percentage, about 0.1% of the images were found to be corrupted and hence were rejected from the study. Since the dataset has multiple images with different image sizes, some amount of pre-processing had to be performed. The pre-processing step firstly converted these images to a single channel gray-scale image.

After this step, slight image enhancement techniques were done by passing it through a sharpening filter. The final step was to convert the image to a 200*200 matrix and normalising the values.

Also, in order to satisfy limited memory usage during the training phase of the model, a suitable data generator was created which could generate random images of batch size 10 and feed it to the network. This data generator was used for training as well as in the serving phase, and all pre-processing steps used during training were also applied during the production phase. The protocol structuring for the experiment was done in a systematic manner as indicated in figure 2.

II.III Computational setting.

The training of the model was computationally expensive, primarily because it involved convolutional neural networks, being trained on about 1000 images. This required high memory and processing capabilities and so conventional laptops or PC's could not satisfy this. Hence majority of the training was carried out on a python notebook hosted on Google Cloud platform. The underlying system configuration was 8 CPU's, 32 GB of RAM and 1 TB of hard disk space. The models were written in Keras using a tensor flow backend.

Owing to the fact that Google Cloud platform has out-of-the-box support for tensor flow, to ensure production level robustness and ease of deployment, it was chosen and all the training was done on this platform. This also gave us the capability to deploy the model in future in a scalable manner without the hassle of setting up servers and configurations.

II.IV Statistical analysis.:

Kolmogorov-Smirnov test was used to determine the distribution of the subjects. Then for the group of subjects having normal distribution, unpaired t-test used. If one or both groups didn't have normally distributed data, Mann-Whitney test was used for comparison. All statistical analysis was performed using GraphPad Prism version 8.0.2.

## III. Results:

### III.I FEATURE SELECTION

There were not many feature selection steps involved in the study and hence all of the feature selection steps have been omitted or carried out in the pre-processing steps.

### III.II NETWORK ARCHITECTURE AND OPTIMIZATION

The input layer of the network consists of a 2D convolutional network with 32 filters followed by a dropout layer. This is followed by another 2D convolutional layer with 64 filters and another dropout layer. This is again followed by 3 layers of a dense network, each with 64, 32 and 16 cells respectively. The layers are activated using a RELU activation function. The output layer consists of a soft-max layer with 3 output nodes. Categorical Cross Entropy is selected as the loss function and the layers are trained using an Adam optimizer. The final model consists of 157,373,475 trainable parameters. The output of the model is a probability distribution spread across 3 class labels.

### III.III TRAINING

The model was trained on a batch size of 10 images each of 200 * 200 in dimension. Furthermore, the model was trained for 20 epochs and the dataset was shuffled after the end of each epoch. The total number of samples trained in each epoch was 2688. The total

training time of the model was 1hour 15 minutes on the hardware. This time could be further reduced by training on a lower number of samples or increasing the computational power of the system. At the end of the training, the model was saved as a .pb file in the local file system which could be later used to recreate the model. After multiple iterations, the best performing model was chosen and was used for final evaluation study.

III.IV Prediction accuracy

The model gave an overall accuracy of 82% among the 3 different classes. This can be further expanded as the following scenarios: -

- The model provided 93.2 % accuracy when trying to detect whether a person is normal or not i.e. it is able to differentiate a normal person from an infected person with very high accuracy.
- The model is able to detect with 74.9% accuracy that a person has Bacterial pneumonia whereas in the rest cases it detects the person as having Viral pneumonia.
- The model is able to detect with 76.5% accuracy that the person has viral pneumonia whereas in the rest 23.4% cases it detects that the person has bacterial pneumonia.

| Actual/Predicted | Bacterial | Normal | Viral/COVID-19 |
|---|---|---|---|
| Bacterial | 251 | 9 | 75 |
| Normal | 5 | 302 | 17 |
| Viral/COVID-19 | 68 | 12 | 261 |

Furthermore, other performance parameters of the model are as follows: -

| | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| Bacterial | 0.77 | 0.75 | 0.76 | 335 |
| Normal | 0.93 | 0.93 | 0.93 | 324 |
| Virus/COVID-19 | 0.74 | 0.77 | 0.75 | 341 |
| accuracy | 0.81 | | | 1000 |
| macro average | 0.82 | 0.82 | 0.82 | 1000 |
| weighted average | 0.81 | 0.81 | 0.81 | 1000 |

The AUC score for the above data is 93.7%. Final accuracy of the model as tested on 1000 samples was found to be 81% and the sensitivity is 82% with a specificity of 90%. With more and better-quality data, the accuracy can be further increased. The performance metrics of the data was obtained using the confusion matrix with the help of Scikit Learn library.

III.V Statistical analysis:

COVID-19 is a respiratory pathogen caused by SAR-COV2 virus which has a probable zoonotic origin. It mainly enters the body via the upper aero-digestive tract, and produces a deadly attack on the lungs of the infected person with severe disease. Subjects predominantly have respiratory distress or even worse being 'happy hypoxia', i.e. a state of oxygen

desaturation without significant shortness of breath and hence leads to a crucial delay in seeking medical help, which might result in fatality. Mainstay of supportive therapy is oxygen support and/or ventilation. We assumed that oxygen saturation may play a crucial role in the survival of these patients. But from the datasets we saw that there was no significant difference in the oxygen saturation level of the subjects who died of SARS-COV2 viral illness vis-a-vis those who recovered from the disease as shown in Figure 3. Hence, we may say that although oxygen saturation is an important clinical marker of the disease, it probably doesn't play a big role in predicting mortality.

## IV. Discussion and Conclusion

The first documented case of corona virus appeared in 1960 followed by the death of about a thousand patients in 2003, and finally a pandemic led by a different strain of corona virus in 2020. Irrespective of such a long history of affliction of the human race, scientists have been unable to produce an officially available vaccine against this deadly air-borne virus [10].

Evidence of pathogenicity is directed by the fact that it can spread via air-borne droplets and via asymptomatic patients across all ages which makes every age group susceptible. Even patients can spread this virus after immediate recovery [11]. Hence, under the prevailing circumstances, there is an urgent need to keep health-care officials safe from nosocomial infection of COVID-19. This can be only done by early diagnosis followed by isolation, and decrease of the patient burden on hospitals and associated intensive care units (ICU). This scenario might be critically dealt with by implementation of machine learning based methodology like a dynamic neural network based on thoracic roentgenogram [12]. Nevertheless, the only drawback comes from the dataset development part as developing a dataset during a pandemic is troublesome but very crucial for strategic planning in healthcare organisations [5-7]. However, we predict that our model along with case record form (CRF)

can significantly contribute in quicker and cheaper way, for mass screening of patients (non-invasively) and also help in managing and streamlining health care centres and its associative monitoring

==========

**References:**


1 Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med 2020;382:929-936. PMID:32004427. doi: 10.1056/NEJMoa2001191.

2 Wang L, Wang Y, Ye D, Liu Q. Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. Int J Antimicrob Agents 2020;55:105948. PMID:32201353. doi: 10.1016/j.ijantimicag.2020.105948.

3 He JL, Luo L, Luo ZD, Lyu JX, Ng MY, Shen XP, et al. Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China. Respir Med 2020;168:105980. PMID:32364959. doi: 10.1016/j.rmed.2020.105980.

4 Shih HI, Wu CJ, Tu YF, Chi CY. Fighting COVID-19: A quick review of diagnoses, therapies, and vaccines. Biomed J 2020. PMID:32532623. doi: 10.1016/j.bj.2020.05.021.

5 Akhtar M, Kraemer MUG, Gardner LM. A dynamic neural network model for predicting risk of Zika in real time. BMC Med 2019;17:171. PMID:31474220. doi: 10.1186/s12916-019-1389-3.

6 Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. PLoS One 2019;14:e0212356. PMID:30779785. doi: 10.1371/journal.pone.0212356.

7 Uhlig S, Nichani K, Uhlig C, Simon K. Modeling projections for COVID-19 pandemic by combining epidemiological, statistical, and neural network approaches. 2020:2020.2004.2017.20059535. doi: 10.1101/2020.04.17.20059535 %J medRxiv.

8 Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018;172:1122-1131 e1129. PMID:29474911. doi: 10.1016/j.cell.2018.02.010.

9 Cohen JP, Dao L, Roth K, Morrison P, Bengio Y, Abbasi AF, et al. Predicting COVID-19 Pneumonia Severity on Chest X-ray With Deep Learning. Cureus 2020;12:e9448. PMID:32864270. doi: 10.7759/cureus.9448.

10 Dharmendra Kumar RM, Pramod Kumar Sharma. Corona Virus: A Review of COVID-19. EJMO 2020;4:8-25. doi: doi:10.14744/ejmo.2020.51418.

11 Singhal T. A Review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr 2020;87:281-286. PMID:32166607. doi: 10.1007/s12098-020-03263-6.

12 Phua J, Weng L, Ling L, Egi M, Lim CM, Divatia JV, et al. Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. Lancet Respir Med 2020;8:506-517. PMID:32272080. doi: 10.1016/S2213-2600(20)30161-2.

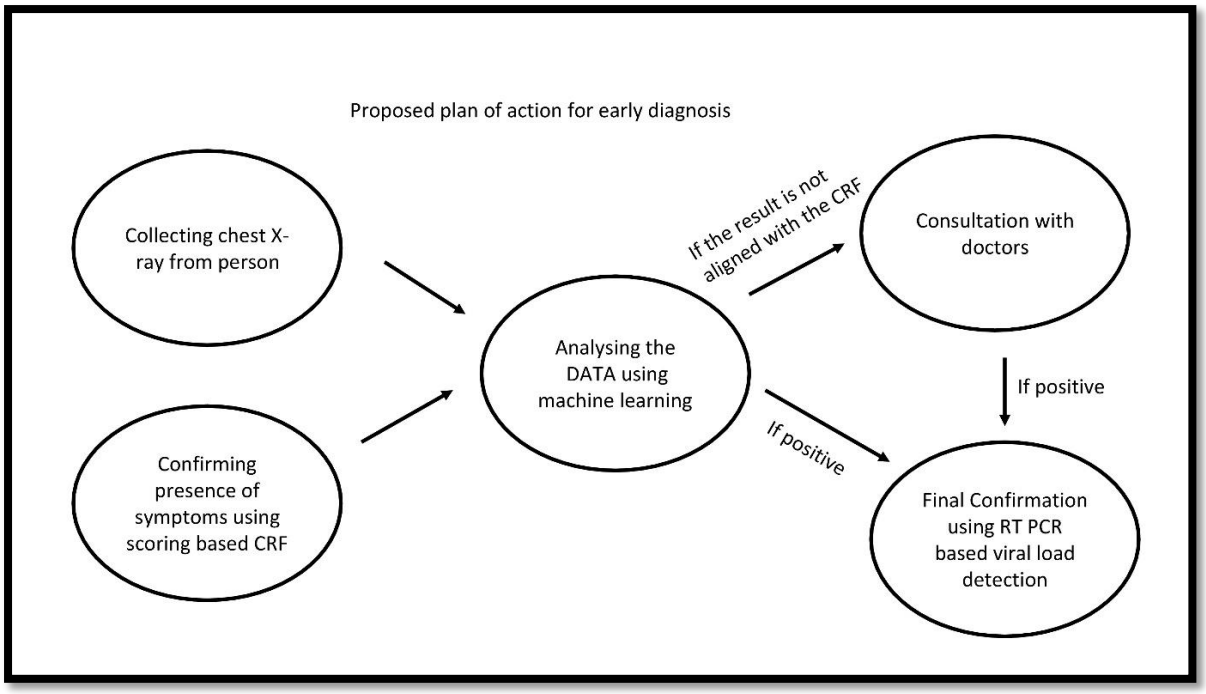


Figure 1: Prognostication model developed by the successive use of the neural network-based prediction model and the associative case record form developed.

.

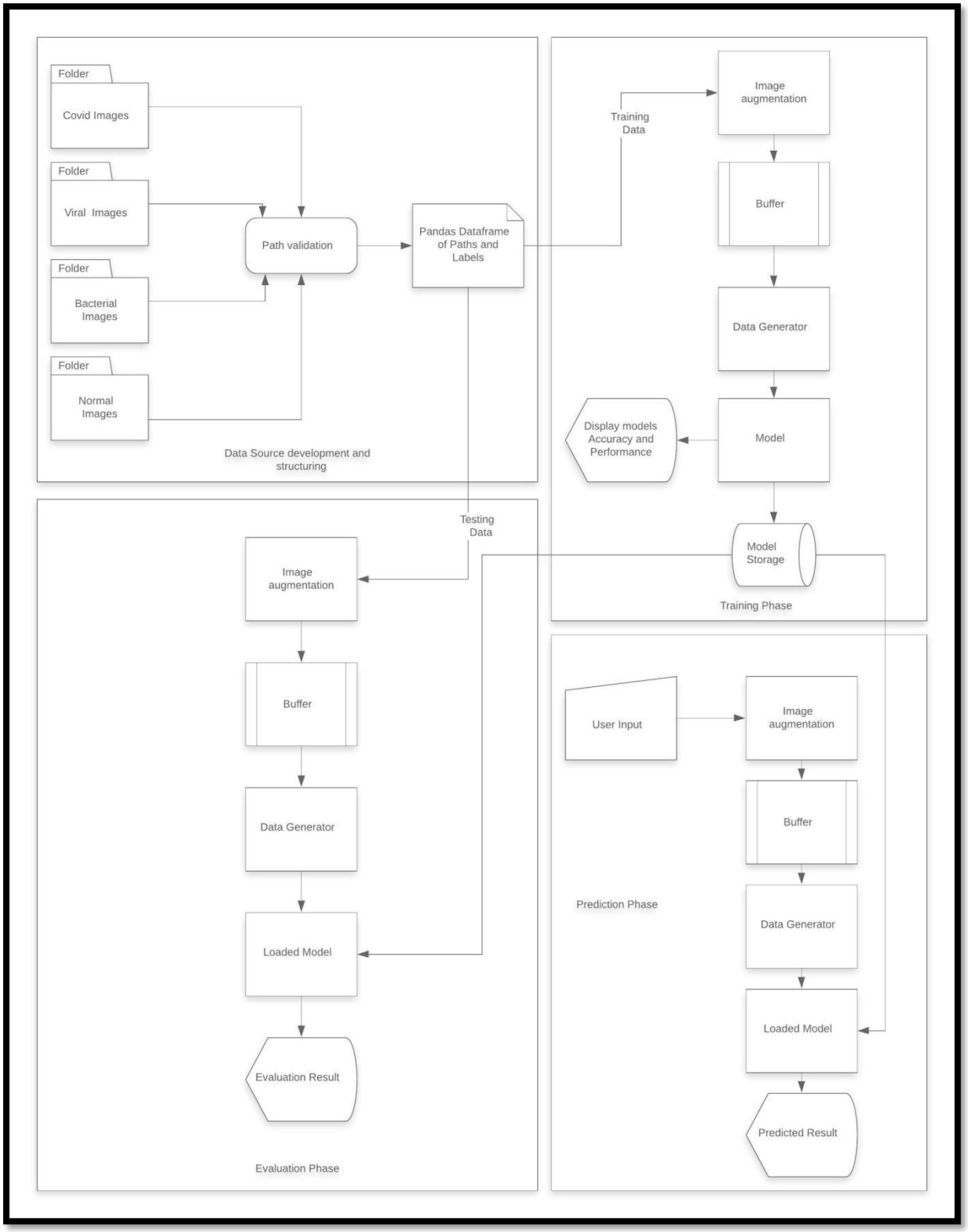


Figure 2: Architecture of the system used to develop the neural network-based prediction model.

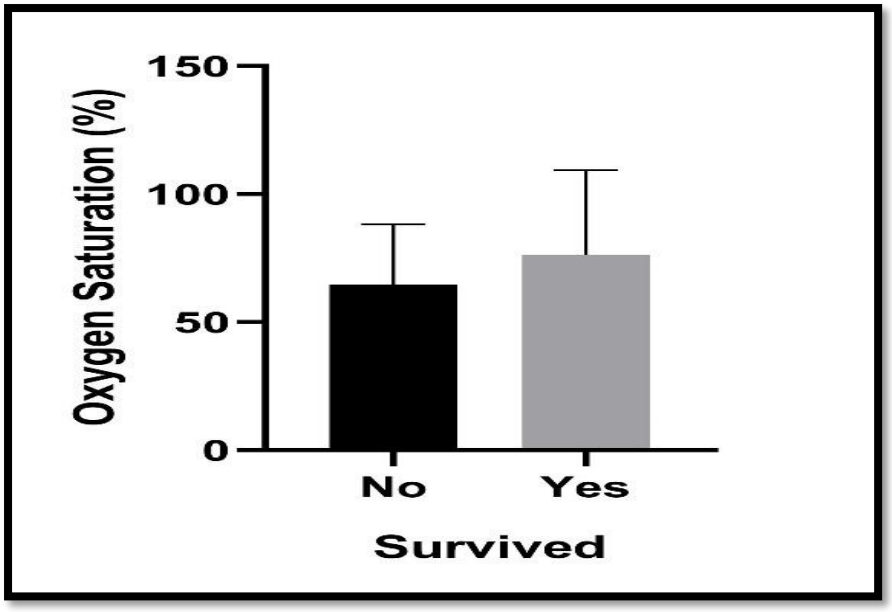


Figure 3: Bar Chart representing that there is no statistically significant difference observed in SpO2 level between the subjects that died due to COVID -19 and those who survived after being infected.