Journal of Engineering and Applied Science Vol. 07, No. 01, pp. 71–79, December 2023

Journal of Engineering and Applied Science

Contents are available at www.jeas.ruet.ac.bd

Developing Generalized Models for COVID-19 Detection and Outbreak Prediction by AI Approach

Partha Pratim Debnath[1*], Erona Moumita, Mursheda Nusrat Della[1], Md. Lincon Hasan[1]

[1] *Information and Communication Engineering (ICE), Bangladesh Army University of Engineering & Technology (BAUET), Bangladesh*

ARTICLE INFORMATION

Received date: 20th Oct 2023
Revised date: 24th Dec 2023
Accepted date: 31st Dec 2023

Keywords

Convolutional Neural Network
Artificial Neural Network
Data Engineering

ABSTRACT

The target of the research is to evaluate the performance of convolutional neural network architectures for medical image classification and artificial neural network for future prediction. Our focus is to create a generalized model from our own dataset by collecting it from different sources. We have also represented an artificial neural network (ANN) model which can estimate and forecast expansion of COVID-19 in a region based on several parameters. The outbreak prediction model worked with the accuracy of 98%. For CNN, system demonstrates an accuracy of 99.4%. Although, our system still has the scope to further improvement when more COVID-19 images become available. Our proposed model can be used for fast and reliable identification of COVID-19 from patient's chest X-rays. Primary health workers in remote places where the proper diagnostic laboratory and expert doctor are not available, can use our automatic COVID-19 detection system. In addition to that, our ANN model can predict the future outbreak of COVID-19 for a country which can help any country to take precaution according to it.

1. Introduction

COVID-19 is a dreadful contagious disease. The mortality rate due to this disease is not too high, yet its expansion rate is very rapid. This severe disease definitely results in death as no effective medicine has been invented. However, a lot of variations of the virus have already been discovered [1]. Although because of the great efforts of the scientists, diagnosis has become relatively swift, in spite of that there is financial issues arising from the cost of diagnostic tests – which is crucial, especially developing countries [2].

Since 31 December 2019, 53,51,43,050 cases of COVID-19 have been announced, among them depressingly 63,28,694 is losing of life case [3]. As per the news report of May 1, 2020 around 23,430 Covid patients have died out in New York City. This proportionate to a 0.28% crude mortality rate or 279 deaths per 100,000 populations, or in other words 1 death every 358 people. The Crude Mortality Rate continued to arise as time passed [4]. Considering the present pandemic situation, it becomes crucial to uniquely identify the patients. And fortunately there is relationship between the classification of COVID-19 cases and chest X-ray image analysis. In this work, an automatic diagnostic system has been developed using CNN which uses chest X-ray to predict whether a person is COVID-19-affected or normal. This research work has shown excellent performance in terms of its accuracy and other evaluating parameters to define the infected one in easily manner and within short period of time. If the proposed CNN model results in a good accuracy and validated with real data, it will be a milestone in the field of telemedicine [5].

* Corresponding authors: Information and Communication Engineering (ICE), Bangladesh Army University of Engineering & Technology (BAUET), Bangladesh
E-mail addresses: parthapratim.ice10@gmail.com (Partha Pratim Debnath)

To construct public health policies and decisions it is essential to acknowledge Coronavirus spread information. These forecast models use different features such as age, hospitalization rates, morality, gender etc. to give an idea how the virus may outbreak in the future. This information is crucial for governments as well as people to fight against this pandemic. In our research work, we have incorporated the COVID-19 outbreak prediction along with its detection. The joint approach will be a new milestone to fight against this deadly disease.

2. Literature Review

The approaches can be classified into two category-manual detection and computer vision-based detection. There has been a lot of prior research with the models, and Convolutional Neural Networks have a far higher level of accuracy. From literature review we can say that, maximum researchers use different type of deep learning methods [6-10]. Some use Swab-based Reverse Transcription Polymerase Chain Reaction (RT-PCR) test, Blood sample-based antibody test and Transfer learning [11-14]. The accuracy level remains high even if the target is changed.

The following table summarizes the overall performance of different models-

Table 1: The models and the accuracy level.

Model	Accuracy
CNN-LSTM	99.4%
Fusion of CNN model	98.93%
Simple CNN	98.28%
Chanaal boasting	97.43%
R-CNN	97.36%
J48 ALGORITHM	97.18%
PDCOVIDNet	96.58%
HOG	96.47%
Hybrid System	95.2%
Covid GAN	95%

We can see that, the model with the best performance is CNN-LSTM, because we get the best accuracy which is 99.4% out of it, the second-best performance holder is Fusion of CNN model having accuracy of 98.93% [15-20].

We see that, most of the model works with optimal accuracy but are not generalized model. To deploy any artificial system in real situation, it should be generalized. In our work we aim to build a generalized model by tuning its parameters.

3. Methodology

The convolutional neural network (CNN), a class of artificial neural networks that has been a well-reasoned method for computer vision tasks as it has shown surprising result on the object recognition competition, pattern recognition and image analysis.

Concept of CNN:

CNN uses a special technique called Convolution. CNN is a type of deep learning model especially effective while dealing with grid type data such as image, which operate based on feature extraction and designed to generate automatically feature map, that assists to recognize patterns latter. Any complex CNN architecture can be decomposed into three fundamental layers: convolution, pooling, and fully connected layers.

In simpler word, an artificial neuron network (neural network) is just reassembles human nerve system and work in the similar way like human brain [19]. A neural network is nothing more than a large collection of artificial neurons, which technically arranged in a sequence of layers.

The basic blocks of any artificial Neural Network is three layers: Input Layer, Hidden Layer, and Output Layer

Our Proposed Model:

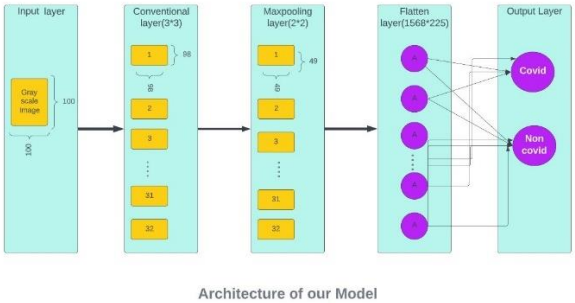

Figure 1: Architecture of our Model-1 (CNN)

In our model-1 there is one input layer, one Convolutional layer, one maxpoling layer and one flatten layer. In our output section there are two neurons by which we get two results- one is Covid and another is non Covid. In input section we take an input image of size 100X100 and in Convolutional layer we convolve it with filter of size 3X3. After that in maxpoling section, we maypole it with kernel (size 2X2). Next, we pass it through flatten layer in which image size is converted into 1568X225. Finally at the output section we get two output results tracking Covid positive or negative.

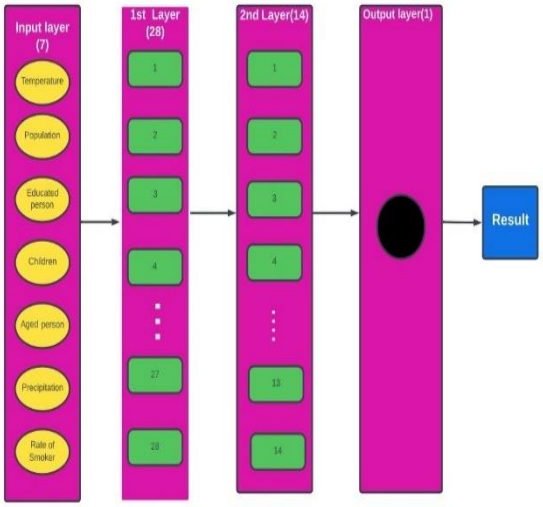

Figure 2: Architecture of our Model-2 (ANN).

In our model-2 we take 7 neurons in input layer. In 1st hidden layer we take 28 neurons and in 2nd hidden layer we take 14 input neurons respectively. Finally, we get an output through a single neuron. Different parameters of the proposed model are summarized as follows-

Table 2: Different Parameters.

Parameter Name	Parameter Value
Total Number of Rows	10000
Epochs Number	200
Number of Layer	2
Number of Features	9

4. Data Collection

Typically, the most challenging part of the whole process is to collect significant data relevant to the project. During the COVID-19 pandemic situation it was impossible for us to collect data from field that's why we had to collect the dataset from online sources and combine them into a suitable on after performing data engineering on it. The datasets exert in our experiment are classified into two categories- image data and text data.

Primarily, an ingathering of 7232 X-ray images inclusively 3616 images with verified COVID-19 infection and 3616 images of normal conditions was used to train and validate our CNN model.

Figure 3: Some sample images of dataset-1

We gather the data from Kaggle. There are 112458 data of 100 countries. The format of dataset is csv, where each column represents a features and attributes about the states affected by Covid.

M	N	O	P	Q
precipitation	temperature	total_population	female_percent	area
0		55601	0.51448715	594.44
0	2	218022	0.515383769	1589.78
0	1.4	24881	0.472167517	884.88
0	2.5	22400	0.4678125	622.58
0	1.4	57840	0.507261411	644.78
0	0.8	10138	0.454823437	622.81
0	3.6	19680	0.53429878	776.83
0	3.2	114277	0.519465859	605.87
0	-0.8	33615	0.521255392	596.53
		26032	0.504609711	553.7
0	2.4	44153	0.509251919	692.85
0		12841	0.527295382	913.5
0	0.8	23920	0.526421405	1238.47
		13275	0.510960452	603.96
0	-0.6	14987	0.504837526	560.1
0	2.5	51909	0.508351153	678.97
0	1	54762	0.52048866	592.62
0	4	12277	0.520974179	850.16
0		10715	0.496126925	650.93
0	5.3	36986	0.516141243	1030.46
0	3.2	13824	0.517216435	608.84
0	0.2	83442	0.505704561	734.84
0	4.8	48956	0.50929406	561.15
0	0	38310	0.538371182	978.7

dataset-1

Figure 4: Collected Dataset-2

Data Preprocessing

Data Cleaning: Data cleaning is emergent move if we want to evaluate more perfectly. It is more like eliminating defective data. It has main 4 steps:

1. Remove similar observations
2. Fix structural errors
3. Handle missing data
4. Replacing with Mean/Median/Mode

Data Scaling:

We use mainly two kinds of data scaling in this research work: 1. Normalization 2. Standarzation.

Data Normalization:
We can define normalization by rescaling of the data from the original range which enable us to accurately estimate the minimum and maximum observable values.

The formula of normalization as follows:
N=(b - Minimum) / (maximum - minimum)……..(1)

Where the minimum and maximum values related to the value b being normalized.

Data Standardization
Standardizing can be achieved by rescaling the distribution of values which ensure that the mean must be 0 and the standard deviation 1. Standardization needed for accurately calculation of the mean and standard deviation of observable values.

The cue to count standardized of a value as follows:

y = (x - mean) / standard-deviation ……..…(2)

Where the *mean* is determined as:

Mean = sum(x) / count(x)……..…..(3)

Image Resizing

"numpy.resize (a, new_shape) "is a python function that Return a new array with the specified shape. It can be found in PIL which is the Python Imaging Library that provides us the facility of image editing. Along with that, some other amenability takes an example of factory functions to load images from files, and to create new images. Image and Resize () Returns a resized copy of this image.

5. Model Evaluation

Covid-19 Detection Model

We trained our model using simple CNN. In the dataset the test train split was kept 8:2. At last we get training accuracy 99% and testing accuracy 95%. After fine tuning our model with different hyper parameters we measure F1 score, precision, accuracy and recall to see the performance of our model. Figure 5 shows confusion matrix of our model with testing data. From this confusion matrix *precision*, *recall*, *Accuracy* etc. are calculated.

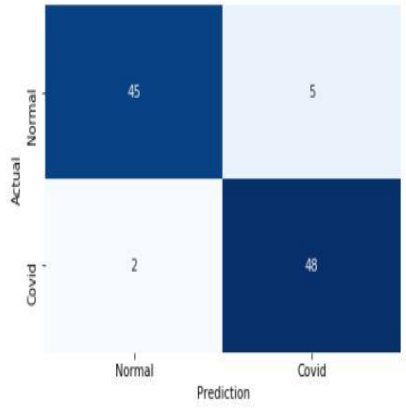

Figure 5: Confusion matrix for Covid-19 detection.

From the confusion matrix we can say that,

True Positive (TP) = 48
False Positive (FP) = 5
True Negative (TN) = 45
False Negative (FN) = 2

Accuracy: Accuracy yields the number of correctly predicted data out of total amount of data.

Accuracy =
(TP+TN)/(TP+TN+FT+FN)……(4)

Precision: Precision can be measure from total number of positively predictive data out of total amount of data in a dataset.

Precision =TP/(TP+FP)……..(5)

Recall: Recall in other word called sensitivity or true positive rate. Good classifier must have recall 1.

Recall = TP/(TP+FN)……..……(6)

F1 score: *F1 score* can be found from the mean of precision and recall. It is a better way to measure accuracy. *F1 score* is defined as follows-

F1 score=
2((precision*recall)/ (precision + recall)……(7)*

In our case,

Accuracy = (48 + 45)/ (48 + 5+ 2 +45) = 0.95 or 95%.
Precision = 48/ (5+48) = 0.9 or 90%.
Recall = 48/ (48+ 2) = 0.93 or 93%.
F1 Score = 2* (0.9* 0.93)/ (0.9+ 0.93) = 0.96 or 96%.

Table 3: Performance evaluation of the proposed model.

Parameter	Value
Training accuracy	95%
F1 score	96%
Precession	93%
Recall	90%

Model Efficiency in Terms of Different Parameters

For any computer vision model, the efficiency changes with different parameters like learning rate filter size, total number of filters, total number of images, total number of layers etc. So, we need to perform experiments by varying the values of different parameters and track the best value.

Figure 6 shows the testing accuracy. At first when we increase iteration the accuracy also increases. Then after a certain point it started to decrease. When X label value is '30' Y label value is the highest which '93%'.

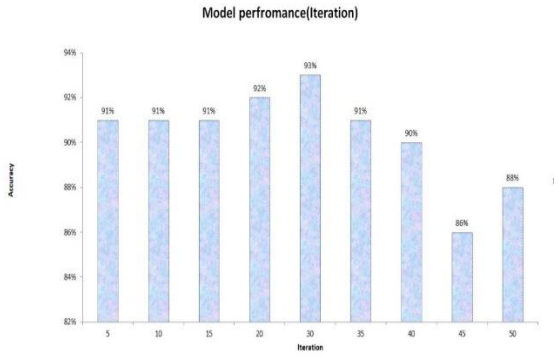

Figure 6: Iteration Vs testing accuracy graph.

Figure 7 depicts the accuracy and iteration relationship. But, for figure 7 accuracy means accuracy for training data. In the graph, we found that at the beginning the accuracy increased with iteration. When the accuracy reached at highest point, then it became almost constant.

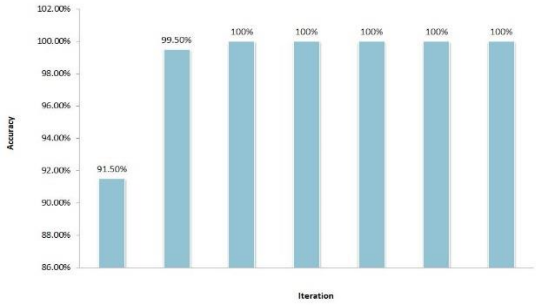

Figure 7: Iteration vs. Training accuracy graph.

From figure 8 we are able to understand, how accuracy change with the size of filters. For figure 8 accuracy means testing accuracy. At first when we increase filter size, the accuracy increased. Then after a certain point it started to decrease. When X label value is '3X3' and Y label value is highest which is '91%'.

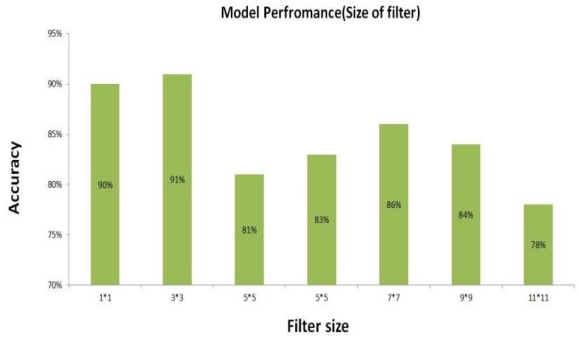

Figure 8: Filter Size vs. Accuracy graph.

Figure 9 is a graph of our achieved accuracy at different number of filters used in different layers of our model. When the number of filters increased, the model can able to extract feature more accurately from image data. As a result, accuracy increased with the increment of number of filters in each layer. We achieved the highest accuracy 93% when x label value is 90.

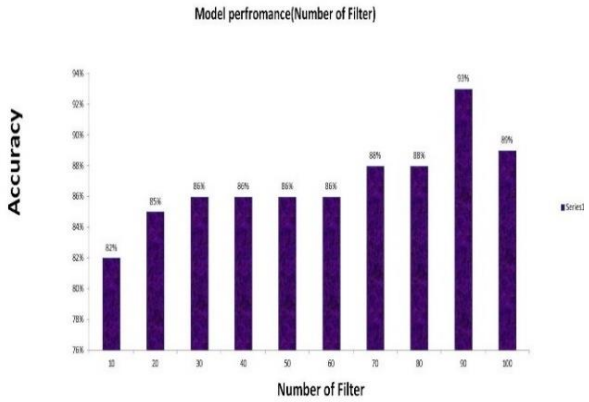

Figure 9: Total number of filters vs. accuracy graph.

In figure 10 we can see, how accuracy changes with total number of images. For figure 10, accuracy means testing accuracy. Total Number of images of a dataset can affect the accuracy. If we have very small number of datasets, it is difficult to have an efficient accuracy. Because, at first when we increased total number of images, the accuracy also increased. Then after a certain point it started to decrease. So, total number of images in a dataset becomes a crucial parameter. For 10,000 images we will get the highest accuracy which is 93%. 10,000 images are enough for the dataset with respect to our model.

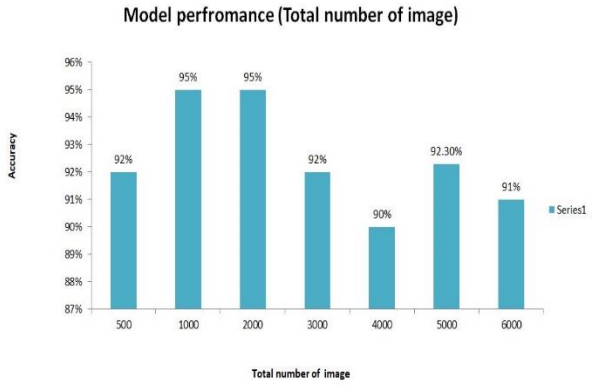

Figure 10: Total number of images in the dataset vs accuracy graph.

Figure 11 is a graph of our model accuracy for different number of layers in our model. At first, we build our model with one convolution layer and one maxpooling layer- that means total two layers. There was enough number of layers in our model. But we still tried to raise the number of layers more than we needed to get the

effects and our validation accuracy fall down gradually in every iteration. The highest accuracy 95%.

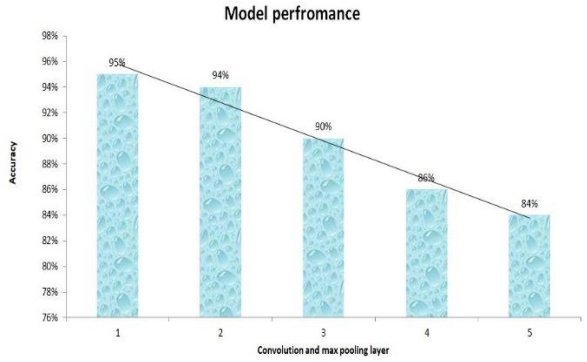

Figure 11: Total number of layers vs accuracy graph.

Now we are going to summarize how accuracy changes with the change of different parameters in a nutshell. Table iv shows the best value of the different parameters that we kept to fine tune our model in terms of highest accuracy.

Table 4: Different parameters of our model.

Parameter	Best Value	Highest Accuracy
Learning rate	30	93%
Filter size	3X3	91%
Total number of filters	90	95%
Total number of images	10,000	95%
Total number of layers	1	95%

Covid-19 Outbreak Prediction Model

We trained our model using simple neural network. The dataset was comprised of 70% training data 30% testing data. At last, we obtained training accuracy 97% and validation accuracy 96%. We also measured F1 score, precision, accuracy and recall from confusion matrix to see the performance of our proposed model. Figure 12 shows confusion matrix for of our model testing data.

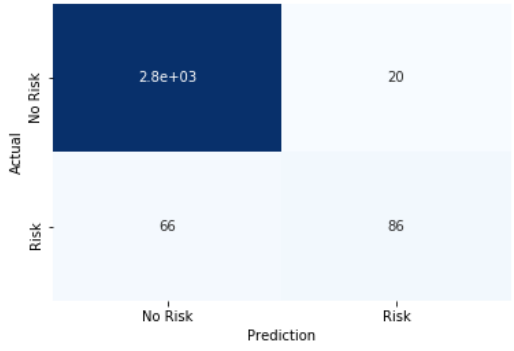

Figure 12: Confusion matrix for outbreak detection.

From the confusion matrix we can say that,

True positive=86
False positive=20
True negative=$2.8*10^3$
False negative=66
For proposed model,

Accuracy = $(66+2.8*10^3)/ (86 +20+ 2.8*10^3 + 66)$ = 0.96 or 96 %. (From equation-4)

Precision = (86)/ (20+86) = 0.97 or 97 %. (From equation-5)

Recall =86/ (86+ 66) = 0.81 or 81 %. (From equation-6)

F1 Score =2* (0.97 * 0.81)/ (0.97 + 0.81) = 0.88 or 88 %. (From equation-7)

Table 5: Performance evaluation of outbreak prediction model.

Parameter	Value
Training accuracy	96%
F1 score	88%
Precession	97%
Recall	81%

Table-5 shows the efficiency of our proposed model.

Model Efficiency in Terms of Different Parameters

For any Artificial Neural Network (ANN) different parameters like Learning rate, Total number of neurons, Total number of features, Total number of data, Total number of layers plays an important role on accuracy. We are going to discuss the effects of these parameters in the following section.

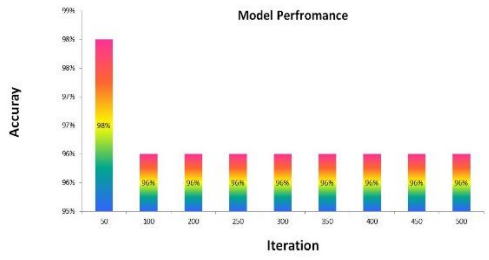

Figure 13: Iteration Vs accuracy graph.

In figure 13 we can see, how accuracy changes for iteration. For figure 13 accuracy means testing accuracy. Iteration also called learning rate. Choosing the perfect learning rate is very challenging. At first when we increased iteration the accuracy started to increase. Then after a certain point it started to decrease.

When X label value is ‘50’ Y label value is highest which is ‘97%’. After 50 iteration we got the highest accuracy which is 97%.

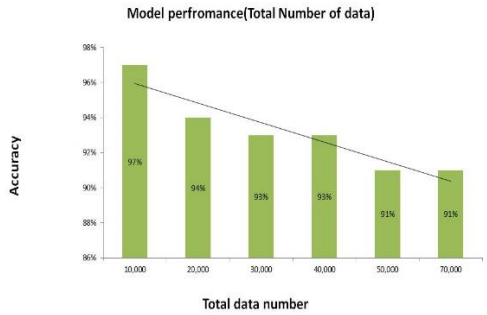

Figure 14: Total number of data vs accuracy graph.

Figure 14 is a pictorial narration of how accuracy changes with total number of data change. For figure 14 accuracy means testing accuracy. Total number of rows in a dataset can affect the accuracy. If we have very small number of datasets, it is difficult to have an efficient accuracy. Normally accuracy increased when number of rows in the dataset was increased. But in our case when we increased total number of rows the accuracy decreased. When X label value lowest which is 10,000. Y label value is highest which is ‘97%’. For 10,000 data we got the highest accuracy which is 97%.

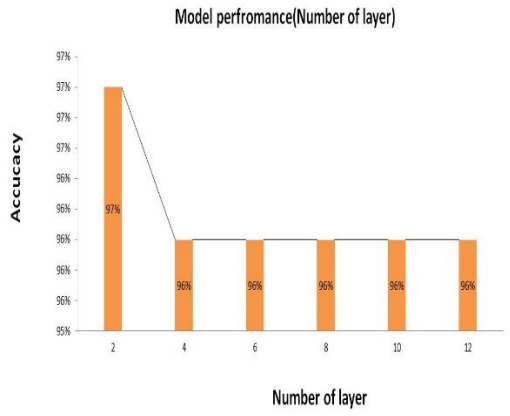

Figure 15: Number of layers vs accuracy graph.

Figure 15 is a graph of our gained accuracy at different number of layers in our model. Initially there were 2 layers in our model. There is the enough number of layers in our model. But we try to raise the number of layers much more than needed and our validation accuracy fall down. We got the highest accuracy 97% when x label value is lowest which is ‘2’.

Figure 16 represents a graph that describes how accuracy changes with the number of features we take. In this case, accuracy means testing accuracy. Feature means how many columns we take as input and these columns provide specific information about output.

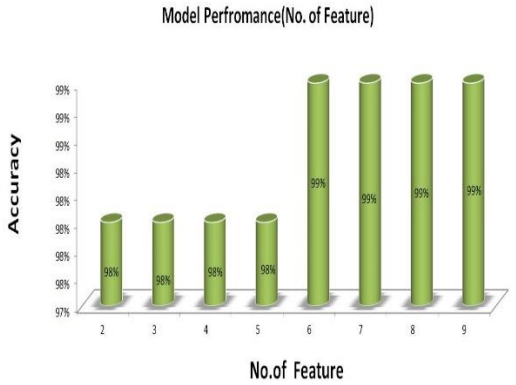

Figure 16: Total number of features vs accuracy graph.

When we increased total number of inputs, the model has more information about the output. So, the accuracy improved with the increment of number of features. When X label value is at peaked point ‘9’, we gain the highest accuracy which is 93%. Model efficiency can be enhanced by raising the number of inputs for a model.

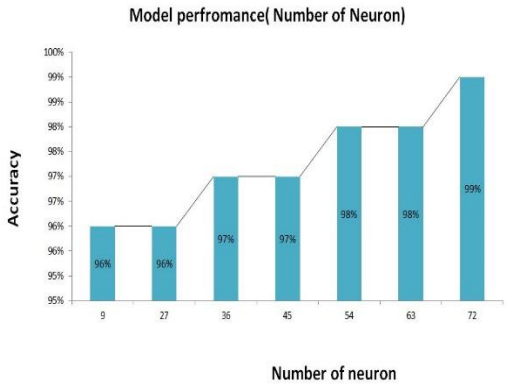

Figure 17: Total number of neurons vs accuracy graph.

Figure 17 shows how accuracy changes when total number of neurons change. For 17 figures accuracy means testing accuracy. Neuron represents small individual units of layers. When we increased total number of neurons of our model, then the complexity of hidden layer also increased. As a result, efficiency increased with the rise of total number of neurons. In figure 16 when X label value is highest (72), Y label value have also reached at highest position (‘99 %’). So we can say that, Accuracy increases with the increment of total amount of neurons.

Now we look forward to summarize the change in accuracy with the change of different parameters. Table

6 shows the parameters of our model that we needed to update to fine tune our model.

Table 6: Different parameters of our model.

Parameter	Best Value	Highest Accuracy
Learning rate	50	97%
Total number of neurons	72	99%
Total number of features	7	97%
Total number of data	10,000	97%
Total number of layers	2	97%

6. Conclusion

The necessity for automatic diagnostic process is at the peak now as there are no effective manual toolkits found economical. Our focus is to detect COVID-19 automatically using AI. In our research paper, a model using simple CNN (Convolutional Neural Network) for the detection of COVID-19 infected person from their chest X-ray is illustrated. We were able to gain a training accuracy of 99.4% and testing accuracy of 95%.

We also tried to invent a model that can foretell future outbreak of COVID-19 for a country. Our model is based on Artificial Neural Network. Future risk analysis model achieved testing accuracy 98% and testing accuracy 97% -which is optimum. Since our model has been trained and tested on the data obtained from online sources, still it needs to be checked with manually driven real data.

7. References

[1] Salama, A., Darwsih, A., & Hassanien, A. E. (2021). Artificial intelligence approach to predict the covid-19 patient's recovery. Digital Transformation and Emerging Technologies for Fighting COVID-19 Pandemic: Innovative Approaches, 121-133.

[2] Visit coronavirus.gov for the latest Coronavirus Disease (COVID-19) updates. [Online]. Available: https://www.healthypeople.gov/2020/leading-health-indicators/2020-lhi-topics/Access-to-Health-Services

[3] Nuttall, N. (2001). Clinical decision-making–can a computer aided learning package help?. British Dental Journal, 190(10), 545-545.

[4] "Numerous Unknowns and Uncertainties About Monkeypox, WHO Acknowledges", [Online]. Available: https://healthpolicy-watch.news/numerous-unknowns-and-uncertainties-about-monkeypox/

[5] Quiroz-Juárez, M. A., Torres-Gómez, A., Hoyo-Ulloa, I., León-Montiel, R. D. J., & U'Ren, A. B. (2021). Identification of high-risk COVID-19 patients using machine learning. Plos one, 16(9), e0257234.

[6] Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., & Hassanien, A. E. (2020). Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art. SN computer science, 1, 1-15.

[7] Organización para la Cooperación y el Desarrollo Económico. (2020). The territorial impact of COVID-19: Managing the crisis across levels of government. OECD Policy Responses to Coronavirus (COVID-19).

[8] Tupetz, A. (2019). Scarf Injuries in Bangladesh: Exploring the Impact on Females who live with Spinal Cord Injuries.

[9] Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, 29(2), 102-127.

[10] Quiroz-Juárez, M. A., Torres-Gómez, A., Hoyo-Ulloa, I., León-Montiel, R. D. J., & U'Ren, A. B. (2021). Identification of high-risk COVID-19 patients using machine learning. Plos one, 16(9), e0257234.

[11] Budd, J., Miller, B. S., Manning, E. M., Lampos, V., Zhuang, M., Edelstein, M., ... & McKendry, R. A. (2020). Digital technologies in the public-health response to COVID-19. Nature medicine, 26(8), 1183-1192.

[12] Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 337-339.

[13] Kassania, S. H., Kassanib, P. H., Wesolowskic, M. J., Schneidera, K. A., & Detersa, R. (2021). Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach. Biocybernetics and Biomedical Engineering, 41(3), 867-879.

[14] Patel, P. (2023). Information for persons who are immunocompromised regarding prevention and treatment of SARS-CoV-2 infection in the context of currently circulating omicron sublineages—United States, January 2023. MMWR. Morbidity and Mortality Weekly Report, 72.

[15] Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A., & Kozlakidis, Z. (2021). Considerations for diagnostic COVID-19 tests. Nature Reviews Microbiology, 19(3), 171-183.

[16] "The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2." Nature microbiology 5, no. 4 (2020): 536-544.

[17] Loeffelholz, M. J., & Tang, Y. W. (2020). Laboratory diagnosis of emerging human coronavirus infections–the state of the art. Emerging microbes & infections, 9(1), 747-756.

[18] Islam, M. Z., Islam, M. M., & Asraf, A. (2020). A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Informatics in medicine unlocked, 20, 100412.

[19] MacKay, M. J., Hooker, A. C., Afshinnekoo, E., Salit, M., Kelly, J., Feldstein, J. V., ... & Mason, C. E. (2020). The COVID-19 XPRIZE and the need for scalable, fast, and widespread testing. Nature biotechnology, 38(9), 1021-1024. [20] Hira, S., Bai, A., & Hira, S. (2021). An automatic approach based on CNN architecture to detect Covid-19 disease from chest X-ray images. Applied Intelligence, 51, 2864-2889.