AI based Automated Diagnosis of COVID-19 Patients

Sudip Mandal
ECE Department, Jalpaiguri Government Engineering College
Jalpaiguri, India

Shankhalika Mallick
ECE Department, Jalpaiguri Government Engineering College
Jalpaiguri, India

Arkaprava Roy
EE Department, Jalpaiguri Government Engineering College
Jalpaiguri, India

Arghyadip Paul
EE Department,
Jalpaiguri Government Engineering College
Jalpaiguri, India

ABSTRACT
The recent Corona Virus Disease 2019 (COVID-19) pandemic has placed severe stress on healthcare systems worldwide, which is amplified by the critical shortage of COVID-19 tests. Effective screening of SARS-CoV-2 enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. In this study, a satisfactory accurate automated diagnosis model of COVID-19 based on patient symptoms has been proposed by applying several Artificial Intelligence (AI) models. For training, COVID-19 data has been collected from Israeli Ministry of Health publicly released data. We have used Artificial Neural Network and Decision Tree for classification or prediction purpose. The proposed model predicted COVID-19 test results with satisfactory accuracy. Hence, this AI based diagnostic framework can be used, among other considerations, to prioritize testing for COVID-19 when testing resources are limited.

Keywords
COVID-19; Artificial Intelligence (AI); Artificial Neural Network (ANN); Decision Tree; Automated Disease Diagnosis; Classification

1. INTRODUCTION
The novel corona virus (COVID-19 or SARS-COV-2) epidemic has been acquainted as a global pandemic. More than 21.9 crore individuals have been infected worldwide, leading to more than 45.5 lakhs deaths as of 7th October 2021 [1]. Rapid human-to-human transmission accompanied by unrevealed nature of the virus had led to this tremendous outbreak of COVID-19. It affects people in various ways. The main symptoms are main symptoms are fever, cough, myalgia or fatigue, sputum, and dyspnea [2]. However, 80% of patients get well from the disease without facing any serious complication. It has been observed that one out of every six affected people get seriously sick and develop difficulty in breathing due to infection and inflammation in lungs caused by the Corona virus [3].

This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment, while many healthcare workers have themselves been infected. Thus, the capacity for immediate clinical decisions and effective usage of healthcare resources is crucial. The most validated diagnosis test for COVID-19, using reverse transcriptase polymerase chain reaction (RT-PCR), has long been in shortage in developing countries. This contributes to increased infection rates and delays critical preventive measures.

Effective screening enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed, in the hope of assisting medical staff worldwide in triaging patients, especially in the context of limited healthcare resources. To defeat the COVID-19 outbreak [4], appropriate and evidence-based actions must be taken worldwide. For this purpose, prediction models can help not only allocating medical resources but also raising the preparedness of healthcare systems involved. In this regard, mathematical, computational and statistical methods have been utilized to predict if a person is affected by COVID-19 or not using Artificial Intelligence (AI) techniques.

In literature, several AI or machine learning models have already proposed for automated diagnosis of the COVID-19. Mainly, two different approaches were being used: COVID affected Lungs X-ray / CT scan images analysis [5], [6] and symptoms based database [7], [8] analysis using different AI techniques. In case of X-ray / CT scan images analysis, different researchers have contributed in this domain using image processing techniques. Li et al. [9] used CT scan images to classify the images in either COVID-19 or normal and they also calculate severity of infection. Hassanien et al. [10] used support vector machine and multilevel thresholding to classify the X-ray images of lungs. On the hand, Sing et al. [11] use multi-objective differential evolution–based convolutional neural networks to classify CT scan images of chest affected by COVID-19. Yan et al. [12] used image processing tools to identify the severity of infection for COVID-19. Rajinikanth et al. [13] used Harmony Search (HS) algorithm along with multilevel thresholding to identify the infection severity in lungs CT scan images that created due COVID-19. S. Mandal [14] used Elephant Swarm Water Search Algorithm (ESWSA) for multilevel thresholding based on Otsu's and Kapur's method and further calculation of severity of infections of lungs images.

On the other hand, several authors also proposed different AI based models for diagnosis of COVID-19 with the help of symptoms based clinical data of COVID-19 patients. Khanday et al. [15] have used different machine learning techniques for the classification of COVID clinical data. Iwendi et al. [16] have used boosted random forest algorithm for COVID-19

patient health prediction. Chen et al. [17] used machine learning techniques for early prediction of mortality rate for COVID-19. Sudre et al. [18] have developed an app that helps to predict if a person is affected by COVID-19 or not using by analyzing the COVID cases and their symptoms. Zoabi et al. [19] used Gradient boosting predictor for automated diagnosis of COVID-19 patients using Israeli Ministry of Health data [28] which is publicly available for the researchers. The obtained results showed very satisfactory accuracy. Dutta et al. [20] utilized three different machine learning models namely bagging algorithm, k-nearest neighbour, and random forest for prediction. They have used the real-time COVID-19 dataset from India Government website [29] and obtained satisfactory accuracy. Mei et al. [20] used AI based techniques for rapid diagnosis of COVID-19 cases to reduce the burden of health workers.

Through this study, an automated prediction model based on Artificial Intelligence models are proposed to predict the COVID-19 through a given set of data comprising of eight different symptoms. Artificial Neural Network (ANN) and Binary Decision Tree are employed for classification purpose. Firstly, a record of confirmed cases of a desired place is taken. Then, the corresponding data were divided into two parts as training and test data. The former was used to train the models, while the latter was used for validation purposes. Thus, the estimated accumulative confirmed cases of the test data were compared with those of actual target values. The rest of the manuscript is structured as follows. Preliminary background of Artificial Neural Network and Binary Decision Tree are discussed in next section. Propose methodology and data collections processes have been elaborated in Section 3. Results using ANN and Decision tree for classification of COVID-19 data have been shown in Section 4 followed by the Conclusion section.

2. THEORETICAL BACKGROUND

In this section, some preliminary concept on Artificial Neural Network and Decision Tree are discussed those are required to understand this research work.

2.1 Artificial Neural Network

An ANN [22] is an information processing paradigm that is inspired by the way the biological nervous system (such as the brain) that process information. An ANN is configured for a specific application such as pattern recognition or data classification [23] through a learning process. An ANN is typically defined by three types of parameters:

1. The interconnection pattern between the different layers of neurons. Three types of layers are observed: input, hidden and output layer.

2. The activation function that converts a neuron's weighted input to its output activation.

3. The learning process for updating the weights of the interconnections.

The main characteristic of ANN is self-learning without prior knowledge of the complex non-linear relationships that exist between the input and output variables. Another advantage is that this type of approach also makes it possible to use several predictor variables simultaneously.

2.2 Decision Tree

Decision tree [24] is one of the most popular and efficient technique in data mining which is established and well-explored by many researchers. Decision trees are categorized as a supervised method that trying to find the relationship between input attributes and target attributes which represent the relationship in structure as a model. The model constructed by using input attributes to predict target However, some decision tree algorithms may produce a large structure of tree size such as J48 which is an implementation of C4.5 algorithm [25]. C4.5 was a version earlier algorithm developed by J. Ross Quinlan.

3. PROPOSED METHODOLOGY

The overall work has two parts:

1. Data collection, preprocessing and preparation of database.

2. Application of AI based model to the dataset for validation of the used techniques to predict the COVID-19 status i.e. either the person is affected or not.

The publicly released COVID data form Israeli Ministry of Health has been utilized for the preparation of the required dataset. Next, this dataset has been used for both training and validation for the Artificial Neural Network and Decision tree on this classification problem. All these AI based models are implemented using MATLAB 2018 in a laptop containing i3 processor and 4 GB processor. The details of the proposed methodology have been elaborated in following sub-sections.

3.1 Data Collection and Preparation

The Israeli Ministry of Health [26], [27], [28] publicly released data of individuals who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab. The initial dataset contains daily records of all the residents who were tested for COVID-19 nationwide. Various information that was provided in the datasheet includes clinically tested symptoms like cough, fever, sore throat, shortness of breath, headache. Based on these data, Artificial Neural Network and Decision Tree model will be trained that will help to predict COVID19 test results using eight binary features: gender, age 60 years or above, known contact with an infected individual, and five initial clinical symptoms. The original data collected for the months of March and April-2020. Then it was processed to make it feasible for further analysis. A few steps were executed to convert the large data into a clean data set. The steps that were followed are as follows:

1. Data for a time period was segregated
2. Rows containing missing data were eliminated
3. Erroneous and wrong data were removed
4. Non binary data was converted to 0 and 1 in the following manner
    i. Gender: male -1; female -0
    ii. Age above 60: yes-1; no-0
    iii. Other information: contact with confirmed-1; others-0
    iv. Corona result:positive-1; negative-0

The modified training validation data set consisted of records from 7,968 tested individuals (of whom 3214 were confirmed to have COVID-19), from the period March 22th, 2020 through March 31st, 2020. The test set contained data from the subsequent week, April 1st through April 7th (5,732 tested individuals, of whom 1952 were confirmed to have COVID-19).

The following list describes each of the dataset's features used by the model for training and testing:

A. Basic information:

1. Sex (male/female).
2. Age ≥60 years (true/false)

B. Symptoms:

3. Cough (true/false).

4. Fever (true/false).

5. Sore throat (true/false).

6. Shortness of breath (true/false).

7. Headache (true/false).

C. Other information:

8. Known contact with an individual confirmed to have COVID-19 (true/false).

3.2 Artificial Neural Network Based Model

To classify the COVID-19 data using ANN based model, Multilayer Feed Forward Artificial Neural Network with one hidden layer which consist of 20 nodes has been used. On the other hand, the input layer consists of 8 nodes which are different symptoms and features of COVID-19 patients. Output layer of ANN model consists of only 1 node that indicates the predicted value (COVID 1 or 0) i.e. the status of the patients.

Initially, the training dataset is used to train the ANN model with the use Back Propagation Algorithm. The trained ANN model is then cross validated using the training data itself and the performance and accuracy was noted. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. Next, the trained model is tested against a new dataset that contained inputs for the period 1st April through 7th April, 2020. The performance and accuracy are also noted for testing new case. All the results are given in results section.

3.3 Decision Tree Based Model

C4.5 algorithm produces decision tree classification for a given dataset by recursive division of the data. The decision tree is grown using Depth-first strategy. On data testing, this algorithm will emphasized on splitting dataset and by selecting a test that will give best result in information gain. With no of fold 10, and confidence factor 0.25, the algorithm is implemented for construction of decision tree. Decision Tree provides vary fast prediction but it has least accuracy than other approaches. Decision Tree shows that one directional path or inference for classification of new data based on different attribute value. Results related with cross validation and testing new data are given in next section.

4. RESULTS AND DISCUSSION

In this section, the detailed results corresponding to two cases of AI models i.e. using ANN and Decision Tree to detect if a person is affected be COVID or not are shown and discussed. Following figure shows the Neural Network model (using MATLAB) to detect if any person is COVID Positive or not, based on some input symptoms and features.

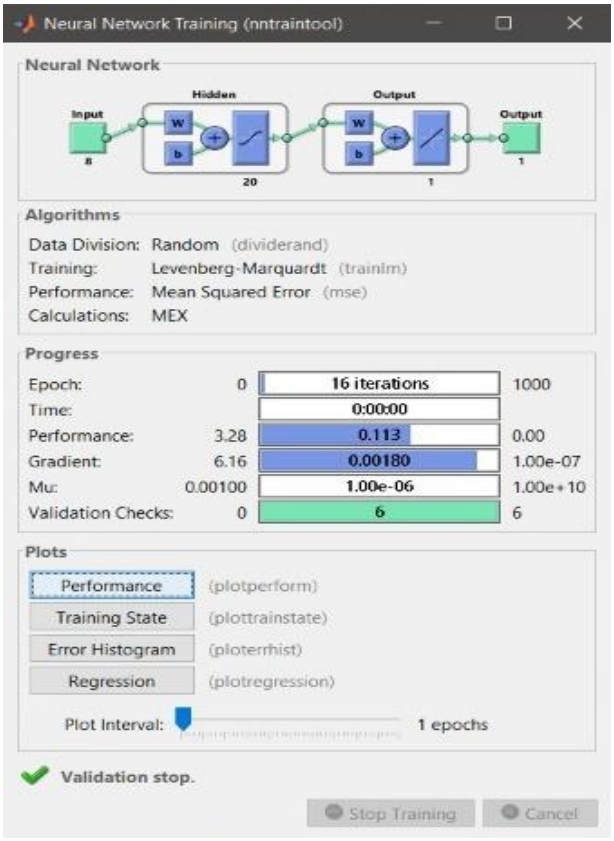

Fig. 1: ANN training model for COVID detection

The Table 1 summarizes the classification accuracy for two tested cases i.e. cross validation and testing new cases for ANN.

Table 1: Accuracy for ANN model

Process	Percentage Accuracy
Cross Validation	84.80%
New Dataset	85.80%

The Table 2 shows the classification accuracy of decision tree the two tested cases i.e. cross validation and testing new cases.

Table 2: Accuracy for Decision Tree model

Process	Percentage Accuracy
Cross Validation	83.75%
New Dataset	85.79%

From above two tables, it has been observed that classification accuracy is better for ANN over the decision tree for both Cross validation and testing new cases. Hence, it can be stated that the ANN is more suitable over decision tree in terms of classification accuracy of COVID-19 dataset.

Figure 2 shows the output Decision Tree model for to detect if any person is COVID-19 Positive or not.

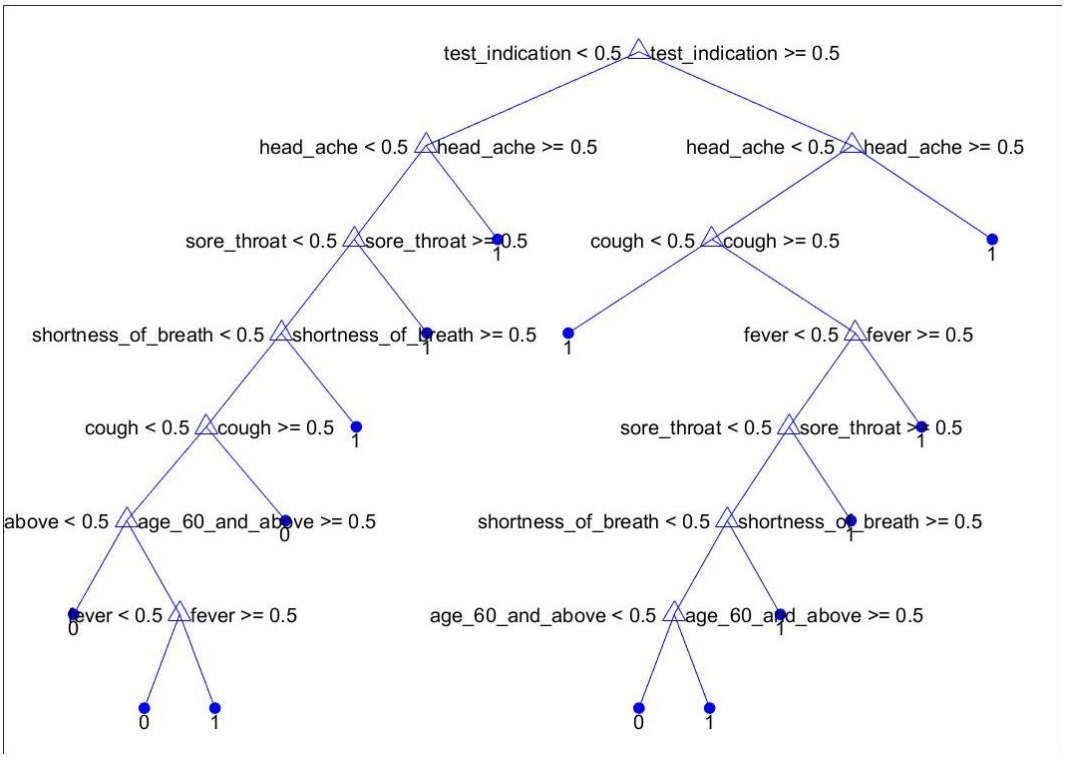

Fig. 2: Decision Tree model for COVID detection

Now, a comparison with respect to runtime of the ANN and Decision Tree for this particular problem of classification of COVID-19 has been observed. From Table 3, it can be clearly noted that Decision Tree is very fast classification algorithm compare to the ANN as it required only 38.18 sec for decision tress training whereas ANN required almost 10 times.

Table 3: Runtime for ANN and Decision Tree

Process	Runtime (Sec)
ANN	363.61
Decision Tree	38.18

However, percentage of error for decision tree is slightly higher than the ANN for both cross validation and testing new case. Following Fig. 3 shows the percentage of error for different cases during classification.

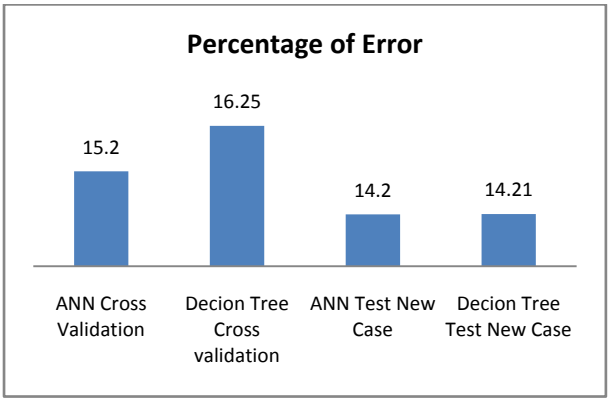

Fig. 3: Percentage of Error for different cases

Next, we shall show confusion matrix for the different case of testing using ANN and Decision Tree. Table 4 and 5 shows the confusion matrix for cross validation and testing new case using ANN respectively.

Table 4: Confusion Matrix for Cross Validation Using ANN

		Target Value	
		0	**1**
Predicted Value	**0**	4321	777
	1	433	2437

Table 5: Confusion Matrix for Testing New Case Using ANN

		Target Value	
		0	**1**
Predicted Value	**0**	3332	368
	1	448	1584

Table 6 and 7 shows the confusion matrix for cross validation and testing new case using decision tree respectively.

Table 6: Confusion Matrix for Cross Validation Using Decision Tree

		Target Value	
		0	**1**
Predicted Value	**0**	4336	418
	1	877	2337

Table 7: Confusion Matrix for Testing New Case Using Decision Tree

		Target Value	
		0	**1**
Predicted Value	**0**	3331	449
	1	365	1587

From the above confusion matrix, it is very clear to us that both ANN and Decision tree both not able to detect COVID cases for few cases (i.e. False Negative or FN) and also they detect few case as COVID positive by mistake (i.e. False Positive or FP). As an example, for cross validation using ANN, 777 numbers of FN and 433 numbers of FP are detected which are not desirable. For this reason, classification accuracy has been reduced to 84.80%. The reason behind this kind of outputs is the asymptotic case for COVID-19. As it is already known to us that many person do not show any symptoms but they are affected by the COVID. Hence, the proposed models also detect them as normal case i.e. FNs are detected which of obvious a drawback of the AI based automated diagnosis process. Similarly, it is also possible that showing symptoms like fever; cough etc. does not guarantee the COVID. The person may suffer due to others disease. In this scenario, computer can predict them as COVID positive case i.e. FPs are detected. However, the AI based COVID-19 diagnosis is still very promising techniques where medical facility like RTPCR test has limited availability.

5. CONCLUSION

Here, the aim is to develop an automated diagnosis system for COVID-19 using two popular AI based models namely Artificial Neural Network and Decision Tree by observing the symptoms and features of the respective patients. This paper focuses on how ANN and decision tree can be used to detect if any person is COVID positive or not, based on eight input parameters. The dataset has been collected from Israeli Ministry of Health website where all data are available publicly for further analysis. In this study, both ANN and decision tree model have been trained and tested for validation of the proposed methodology. It has observed that performance of the ANN is superior over decision tree in terms of classification accuracy for both cross validation and testing new cases. However, the runtime for decision tree is very small compare to ANN. It is an advantage of using decision tree for classification problem. It is possible to achieve satisfactory classification accuracy for both models although other advanced AI model or algorithm may be applied to improve the performance in future. If healthcare facility (i.e. RTPCR test) is limited, this AI based diagnostic framework can be used where automated decision can be taken by the computer or machine instantly without waiting for the report from the pathologist.

6. REFERENCES

[1] World Health Organization. Report of the WHO-China Joint Mission on Corona virus Disease 2019 (COVID-19) Geneva: World Health Organization; 2020 https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-finalreport.

[2] Ministry of Health & Family Welfare Government of India. https://www.mohfw.gov.in/pdf/PreventionandManagementofCOVID19FLWEnglish.pdf [Accessed on 20th May 2020].

[3] Liu J, Liao X, Qian S et al. Community transmission of severe acute respiratory syndrome coronavirus 2, Shenzhen, China, 2020. Emerg Infect Dis 2020 doi.org/10.3201/eid2606.200239

[4] Novel Corona Virus Map, https://infographics.channelnewsasia.com/covid-19/map.html [Accessed on 24th May 2020]

[5] Chung M, Bernheim A, Mei X, et al. CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV). Radiology. 2020;295(1):202-207. doi:10.1148/radiol.2020200230.

[6] Pan, Feng, et al. "Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia." Radiology (2020): 200370.

[7] Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020; 395: 497–506.

[8] Liu Y, Yan LM, Wan L et al. Viral dynamics in mild and severe cases of CVOID-19. Lancet Infect Dis doi.org/10.1016/S1473-3099(20)30232-2

[9] Li, K., Fang, Y., Li, W. et al. CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). Eur Radiol (2020). https://doi.org/10.1007/s00330-020-06817-6.

[10] Aboul Ella Hassanien Sr., Lamia Nabil Mahdy Jr., Kadry Ali Ezzat Jr., Haytham H. Elmousalami Jr.,Hassan Aboul Ella Jr. Automatic X-ray COVID-19 Lung Image Classification System based on Multi-Level Thresholding and Support Vector Machine, doi: https://doi.org/10.1101/2020.03.30.20047787.

[11] Singh, D., Kumar, V., Vaishali et al. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks. Eur J Clin Microbiol Infect Dis (2020). https://doi.org/10.1007/s10096-020-03901-z.

[12] Yan, R. et al. Chest CT Severity Score: An Imaging Tool for Assessing Severe COVID-19. Radiology: Cardiothoracic Imaging 2020, 2(2). https://doi.org/10.1148/ryct.2020200047.

[13] Rajinikanth, V., Nilanjan Dey, Alex Noel Joseph Raj, Aboul Ella Hassanien, K. C. Santosh, and N. Raja. "Harmony-search and otsu based system for coronavirus disease (COVID-19) detection using lung CT scan images." arXiv preprint, arXiv:2004.03431 (2020).

[14] S. Mandal, "Identification of Severity of Infection for COVID-19 Affected Lungs Images using Elephant Swarm Water Search Algorithm" International Journal of Modelling and Simulation, 2021, doi: 10.1080/02286203.2021.1934797.

[15] A.M.U.D. Khanday et al., Machine learning based approaches for detecting COVID-19 using clinical text data, Int. J. Inf. Technol. (2020), https://doi.org/10.1007/s41870-020-00495-9.

[16] C. Iwendi, et al., COVID-19 patient health prediction using boosted random forest algorithm, Front. Public Health 8 (2020), https://doi.org/10.3389/fpubh.2020.00357.

[17] X. Chen, Z. Liu, Early prediction of mortality risk among severe COVID-19 patients using machine learning, preprint, Epidemiology (2020), https://doi.org/10.1101/2020.04.13.20064329.

[18] Carole H. Sudre et al, Attributes and predictors of Long-COVID: analysis of COVID cases and their symptoms collected by the COVID Symptoms Study App, medRxiv preprint (2020) doi: https://doi.org/10.1101/2020.10.19.20214494.

[19] Yazeed Zoabi et al, Machine learning-based prediction of COVID-19 diagnosis based on symptoms, npj Digital Medicine (2021) 4:3 ; https://doi.org/10.1038/s41746-020-00372-6.

[20] Pijush Dutta, Shobhandeb Paul, Asok Kumar, Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19, Electronic Devices, Circuits, and Systems for Biomedical Applications (2021). https://doi.org/10.1016/B978-0-323-85172-5.00020-4.

[21] Mei, X. et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat. Med. 26, 1224–1228 (2020).

[22] S. Mandal, G. Saha, and R. K. Pal, "A Comparative Study on Disease Classification Using Different Soft Computing Techniques", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), vol. 1(3), pp. 59-66, 2014

[23] S. Mandal, G. Saha, and R. K. Pal, "An Approach towards Automated Disease Diagnosis & Drug Design Using Hybrid Rough-Decision Tree from Microarray Dataset", Journal of Computer Science and System Biology, vol. 6(6), pp. 337-343, 2013, DOI:10.4172/jcsb.1000130.

[24] S. Mandal, and I. Banerjee, "Cancer Classification Using Neural Network", International Journal of Emerging Engineering Research and Technology, vol. 3(7), pp. 172-178, 2015.

[25] S. Mandal, G. Saha, and R. K. Pal, "Neural Network Training Using Firefly Algorithm", Global Journal on Advancement in Engineering and Science, vol. 1(1), pp. 07-11, 2015.

[26] COVID-19-Government Data. https://data.gov.il/dataset/covid-19 (2020).

[27] The Novel CoronavirusIsrael Ministry of Health. https://govextra.gov.il/ministry-ofhealth/corona/corona-virus-en/ (2020).

[28] COVID-19-Government Data Information. https://data.gov.il/dataset/covid-19/resource/3f5c975e-7196-454b-8c5b-ef85881f78db/download/-readme.pdf (2020).

[29] Covid-19 India data information. https: www.covid19india.org (2020)