Original Article

Prediction equations for detecting COVID-19 infection using basic laboratory parameters

Shirin Dasgupta[1], Shuvankar Das[2], Debarghya Chakraborty[2]

[1]Dr. B. C. Roy Multi Speciality Medical Research Centre, Indian Institute of Technology Kharagpur, West Bengal, India,
[2]Department of Civil Engineering, Indian Institute of Technology Kharagpur, West Bengal, India

Abstract

Objectives: Coronavirus disease 2019 (COVID-19) emerged as a global pandemic during 2019 to 2022. The gold standard method of detecting this disease is reverse transcription-polymerase chain reaction (RT-PCR). However, RT-PCR has a number of shortcomings. Hence, the objective is to propose a cheap and effective method of detecting COVID-19 infection by using machine learning (ML) techniques, which encompasses five basic parameters as an alternative to the costly RT-PCR. **Materials and Methods:** Two machine learning-based predictive models, namely, Artificial Neural Network (ANN) and Multivariate Adaptive Regression Splines (MARS), are designed for predicting COVID-19 infection as a cheaper and simpler alternative to RT-PCR utilizing five basic parameters [i.e., age, total leucocyte count, red blood cell count, platelet count, C-reactive protein (CRP)]. Each of these parameters was studied, and correlation is drawn with COVID-19 diagnosis and progression. These laboratory parameters were evaluated in 171 patients who presented with symptoms suspicious of COVID-19 in a hospital at Kharagpur, India, from April to August 2022. Out of a total of 171 patients, 88 and 83 were found to be COVID-19-negative and COVID-19-positive, respectively. **Results:** The accuracies of the predicted class are found to be 97.06% and 91.18% for ANN and MARS, respectively. CRP is found to be the most significant input parameter. Finally, two predictive mathematical equations for each ML model are provided, which can be quite useful to detect the COVID-19 infection easily. **Conclusion:** It is expected that the present study will be useful to the medical practitioners for predicting the COVID-19 infection in patients based on only five very basic parameters.

Keywords: Artificial neural network, COVID-19, laboratory parameters, multivariate adaptive regression splines, predictive models

Introduction

The first four human coronavirus strains discovered, NL63 (HCoV-NL63), 229E (HCoV-229E), OC43 (HCoV-OC43), and HKU1 (HCoV-HKU1), were identified to cause common cold in otherwise healthy individuals. The year 2003 witnessed a new strain, the severe acute respiratory syndrome coronavirus (SARS-CoV), which was first isolated in China, followed by another, the Middle East respiratory syndrome coronavirus (MERS-CoV), in 2012 detected in Saudi Arabia. Both these strains caused severe respiratory tract infections, MERS having an extremely high death rate. The outbreaks were a result of zoonosis possibly originating from bats. Both SARS-CoV and MERS-CoV caused epidemics, thus exposing coronaviruses as a risk to human health and thus necessitating scientific research into the coronaviruses. In 2019, a new strain again initiated from China, which too probably originated from bats. The virus was established as SARS-CoV-2 [a non-enveloped ribonucleic acid (RNA) beta coronavirus]; the World Health Organization declared the outbreak as pandemic on March 11, 2020 due to its worldwide potential and fatal outcomes.[1] Within a short span,

Address for correspondence: Dr. Shirin Dasgupta, Dr. B. C. Roy Multi Speciality Medical Research Centre, Indian Institute of Technology Kharagpur, Kharagpur - 721 302, West Bengal, India.
E-mail: shirin@bcrmrc.iitkgp.ac.in

Received: 23-11-2023 **Revised:** 06-01-2024
Accepted: 04-02-2024 **Published:** 28-06-2024

Access this article online

Quick Response Code:

Website: http://journals.lww.com/JFMPC

DOI: 10.4103/jfmpc.jfmpc_1862_23

How to cite this article: Dasgupta S, Das S, Chakraborty D. Prediction equations for detecting COVID-19 infection using basic laboratory parameters. J Family Med Prim Care 2024;13:2683-91.

the SARS-CoV-2 pandemic managed to cause about 360 million infections with over 5 million casualties globally.[2] This infection is transmitted through respiratory droplets of infected people; besides, the virus's presence has been traced in the stool and urine of infected individuals.[3] It is the most terrible pandemic that tormented humanity since the Spanish flu of 1918. Contrary to other coronaviruses, SARS-CoV-2-infected patients can remain asymptomatic or manifest mild symptoms; thus, the patient remains unaware of the condition. This upsurges the risk of the disease getting spread from one to another, thus highlighting the need for early identification of infection.[4]

COVID-19 patients can produce a plethora of symptoms with highly variable clinical features ranging from the asymptomatic state to respiratory failure and septic shock along with ground-glass opacities, consolidations, and reticular patterns on chest computed tomography (CT).[5]

The complete blood count (CBC), one of the first-line investigations, is a rapid, non-complicated, and easily available test which offers valuable evidence about the patient's disease condition. Hematological and biochemical markers are central elements in the investigation of suspected COVID-19 cases as their abnormal values can provide insight to the clinicians regarding risk stratification and prognosis.[6] The importance of CBC parameters can be emphasized by Zheng *et al.*,[7] where it was found that COVID-19 patients at the time of admission typically presented with a platelet count and total leucocyte count in the normal range but developed lymphopenia right after the onset of symptoms, which became more distinct as the disease progressed. This study also showed that the neutrophil, lymphocyte, and platelet counts are clinically useful in stratifying patients with COVID-19.

The reverse transcription-polymerase chain reaction (RT-PCR) is the current gold standard tool for detecting this COVID-19 infection. However, it is costly and requires dedicated equipment and reagents, trained personnel for the sample collection, and proper genetic conservation of the RNA sequences used for annealing the primers.[8] Additionally, RT-PCR is difficult to perform in the peripheral rural areas because of its requirement for infrastructure, whereas it is more feasible to conduct a simple CBC test in these circumstances.

In this paper, two machine learning (ML)-based predictive models for predicting COVID-19 infection are implemented as a cheaper and simpler alternative to RT-PCR based on five parameters, that is, age, total leucocyte count (TLC), red blood cell (RBC) count, platelet count, and C-reactive protein (CRP). The obtained equations will greatly help the general primary care providers and family physicians for a rapid and effective COVID diagnosis; this diagnosis will be based only on the input of these five basic parameters.

Recently, several ML model-based studies[9,10] have been conducted; however, most studies[8] considered CT scans and chest X-rays as input parameters in the ML framework. However, both tests give high false-negative results and large exposure to CT radiation may be carcinogenic. In addition, both tests are very costly. On the other hand, laboratory-based clinical tests are easily accessible in most rural medical facilities. Different types of ML models were applied by several researchers to identify the COVID-19 infection using laboratory parameters.[2,11]

Recently, Chadaga *et al.*[10] utilized 13 laboratory parameters in four ML models (i.e., logistic regression, K nearest neighbors, random forest, and XGBoost) to detect COVID-19 infection. By considering 30 input laboratory parameters, Lin *et al.*[12] predicted the COVID-19 infection with Artificial Neural Network (ANN). By using deep neural networks, Babaei Rikan *et al.*[13] predicted the COVID-19 infection utilizing minimum 15 laboratory parameters. In a recent study, four input laboratory parameters [i.e., age, white blood cell (WBC) count, monocytes, and lymphocytes] have been considered in a stacking machine learning model (SML) to predict the COVID-19 infection.[2] A nomogram-based prediction model has been presented with three ML model (i.e., gradient boosting, random forest, and XGBoost) scores in terms of probability.

From the available literature, it is observed that most of the studies considered more than 15 laboratory parameters to predict the COVID-19 infection. Moreover, as per authors' knowledge, except Rahman *et al.*[2], no study has provided any prediction equation. The prediction equation provided by Rahman *et al.*[2] was based on an SML model, which is a bit complicated and difficult to reproduce. On the other hand, the present study considers only five basic parameters. Most importantly, simple and reproducible predictive equations are presented by two ML methods, namely, ANN and MARS.

Materials and Methods

Laboratory parameters considered for the study

As mentioned above, five basic parameters, that is, age, TLC, RBC count, platelet count, and CRP, are considered as the inputs for the present machine learning-based predictive models. This section provides a detailed discussion on the significance of considering each of these parameters.

Age

SARS-CoV-2 was found to be more prevalent and have a more serious outcome in older subjects. Generally, aged individuals are more susceptible to any infection compared to their younger counterparts. The possible reason could be their relative immune compromised state and associated co-morbidities. COVID-19 being a novel virus, not been previously encountered, it is necessary to have more immune cells to fight against it, which is lacking in older individuals. The availability of naive T-cells along with the ratio of CD4/CD8 T-cells is inversely proportional to increasing age; thus, the ability to address any new pathogen becomes depleted, leading to poor prognosis to COVID-19 in the aged.[14]

Typically, the lung conducts defense mechanisms like cough reflex, the barrier function of the epithelium and mucus, and mucociliary clearance, which acts with the innate immune system in harmony to remove inhaled or aspirated substances, including infectious agents. These defense mechanisms decrease with age. As a person ages, the thymus becomes atrophic and is eventually replaced by fibrotic tissue, resulting in a reduced number or even complete nullification of exiting naive T-cells. This exhausts the T-cells, thereby reducing the immunity, making the older individuals more susceptible to the infection.[15]

A study to discover the association between the age and gender of the whole population in various geographical areas and the epidemic characteristics of COVID-19 globally showed that the incidence rate, case fatality rate, and mortality rate of COVID-19 were high in regions where the maximum population was composed of people aged 65 years and above. Conversely, places with a higher proportion of young population (under 25 years old) showed less COVID-19 rates.[16]

A review reported that elderly populations were a higher-risk group of developing adverse complications to COVID-19 (i.e., acute liver injury, acute kidney injury, acute cardiac injury) when compared to their younger counterparts.[17]

Total Leukocyte Count

Leukocytes are cells of the immune system that protect the body against infectious disease and foreign agents. Since the immune system is altered in COVID-19 infection, TLC has a significant role in the propagation of this infection. The dysregulation of immune responses and subsequently immunologic abnormality played important roles in the severity of viral diseases.[18]

A retrospective study conducted to correlate the TLC of confirmed COVID-19 patients with a definite clinical outcome (i.e., death or discharge) showed that the patients with higher TLC faced a much higher death possibility.[19] In a meta-analysis, it was reported that patients who died of COVID-19 had significantly high TLC, whereas the patients who had the infection had only mild elevation in this parameter.[20]

A systemic review and meta-analysis targeted to explore the effect of risk factors on the severity of the infection, focusing on immune-inflammatory parameters, which represent the immune status of patients, revealed that compared to COVID-19 patients with normal TLC, COVID-19 patients who presented with an increased TLC (>10,000 cells/mm^3) had about 3-fold higher risk of developing severe infection.[21]

RBC count

RBC count is the number of red blood cells per cubic millimeter (mm^3) of blood. Many studies did find a correlation with this parameter with the severity of COVID-19 infection.

A retrospective cross-sectional study stated that hemoglobin concentration, RBC count, hematocrit, and other RBC indices were all significantly decreased in COVID-19 subjects compared with controls.[22] Other studies also agree with similar findings which displayed that severely and critically ill patients of COVID-19 had a significantly decreased RBC count and hemoglobin when compared to normal controls.[23,24]

Another study concluded that the determination of RBC count and hematocrit concentration were the most significant predictors of death of COVID-19 patients, thereby putting forward the importance of these parameters.[25] Another cohort study concluded that among many other parameters, anemia is also a poor prognostic factor for COVID-19 patients.[26]

Platelet count

Platelets are tiny cell fragments (2–4 μm) which are formed and introduced into the blood stream by megakaryocytes. Their principal function is to induce a first-line cellular response to thrombosis and vascular injuries. Platelets have the potential to connect the immune system with thrombotic events through the release of various effective chemokines and cytokines, thus acting as key inflammatory mediators.[1] The interactions between endothelial cells, platelets, and leukocytes play a critical role in the pro-coagulant effect of viral infections.[27]

The heart, liver, brain, and kidneys are highly susceptible hosts to micro-thrombi formation in COVID-19 patients.[28]

However, studies have also found platelet counts as an unreliable predictor of COVID-19 mortality. One case series reported that only 5% of patients of COVID-19 had platelets counts <100,000/mm^3 of blood.[29] Another study stated that among 69 COVID-19 patients, none had platelet counts <100,000/mm^3 of blood at admission.[30] In a retrospective cohort study, there was no significant association between the platelet count on admission and disease severity or mortality of COVID-19 patients.[31]

C-reactive protein

CRP is an acute-phase reactant protein, synthesized by the liver and responsible for the removal of pathogens via the complement system and enhanced phagocytosis. Its concentration rises in various inflammatory conditions.

CRP rise was reported during SARS outbreak in 2002. A direct correlation between higher CRP concentrations and deteriorating lung lesions in COVID-19 positive patients was also demonstrated.[32] Various studies were conducted to determine CRP as a biomarker associated with poor prognosis of COVID-19 patients. A meta-analysis reported that patients who died of COVID-19 infections portrayed significantly higher CRP concentrations when compared to those who survived.[33]

In another study, receiver operating characteristic curve analysis established CRP as a valuable predictor of COVID-19 infection progression and severity. Furthermore, patients with CRP > 64.75 mg/L are more likely to have severe

complications.[34] A retrospective, single-center study with an aim to evaluate the potential of CRP in outcome prediction of patients with COVID-19 concluded that the serum CRP level upon admission is a determinate of disease severity. To the best of their knowledge, this was the first report on the prognostic value of CRP in patients with COVID-19.[35]

A systematic review discovered that the high level of CRP was observed in about 85% of severely ill COVID-19 patients, which indicates severe infection and poor outcome. CRP is the only marker which correlates to the progression of non-severe COVID-19 infection.[36]

CRP increased significantly at the initial stage in severe COVID-19 patients, while there are still no CT findings.[37] Furthermore, it was confirmed that CRP is an early biomarker which can predict the severity of COVID-19 with good performance.

Based on this detailed literature review, it is quite clear that all these five parameters (i.e., age, TLC, RBC count, platelet count, and CRP), which can be obtained quite easily even at a rural setup, may be used as the input parameters for the predictive machine learning models.

Data collection

This is a retrospective cross sectional study conducted in an in-campus semi-urban hospital at Kharagpur, India, from April 2022 to August 2022. Patients included in this study were those who attended the OPD with respiratory tract symptoms such as cough, running nose, flu-like symptoms, fever, and so on. The laboratory parameters were tested in the Pathology and Biochemistry laboratory of the hospital. Venous blood was collected from these patients by trained phlebotomists maintaining aseptic conditions. 2.5 ml of blood was collected in a K2-EDTA vacutainer and 3 ml in a clot vacutainer and run in a Hematology and Biochemical analyzer, respectively.

After collecting a nasopharyngeal swab from the patients, the COVID-19 test was done for all these patients using a rapid antigen detection kit. Among 171 patients, 88 were found to be COVID-19-negative and 83 were found to be COVID-19-positive. The present study utilized only the test results obtained from routine laboratory work, and the patients' identities are not revealed.

ML techniques

In this section, a description regarding two ML techniques, that is, ANN and MARS, are discussed.

ANN

Among all the available machine learning techniques, ANN is the most popular one. In recent years, ANN has been employed efficiently for several purposes in the medical field.[38] In the present study, by casting the problem as a classification problem, out of 171 patients' laboratory data, randomly selected 137 (80%) patients' data are kept for training, and the remaining 34 (20%) patients' data are utilized for testing. By using five input parameters (i.e., age, TLC, RBC count, platelet count, and CRP) and two output classes (i.e., COVID-19-positive and COVID-19-negative), ANN classification model is created. Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) quasi-Newton algorithm is applied in MATLAB Version 9.12 for the selected ANN classification model. For the stable performance of the ANN model, the suitable number of neurons in the hidden layer is set as 8 [refer to Figure 1(a)]. The selected ANN architecture is shown in Figure 1(b). By using the obtained weights and biases of the optimal ANN model, the mathematical prediction equation for COVID-19 infection is developed.

Additionally, to understand the effects of the input parameters on the obtained output results, sensitivity analysis is executed. Garson's algorithm is employed to test the impacts of the five input parameters on the COVID-19 Indicator (*CI*).[39] For finding out the relative importance (*RI*) of every input parameter (refer to Equation 1), the weights with respect to the corresponding input parameters are used.

$$RI_j = \frac{\sum_{m=1}^{H}\left(\left(\frac{|W_1^{mj}|}{\sum_{n=1}^{P}|W_1^{mn}|}\right)\times|W_2^{m}|\right)}{\sum_{n=1}^{P}\left\{\sum_{m=1}^{H}\left(\left|W_1^{mn}\right|\Big/\sum_{n=1}^{P}|W_1^{mn}|\right)\times|W_2^{m}|\right\}} \qquad j=1,2,3,4,5 \tag{1}$$

where RI_j indicates the index of each input parameter importance in a relative manner. P and H are the numbers of input layers and hidden neurons, respectively, and W_1 and W_2 represent the weights acquired from the selected ANN model for input-hidden and hidden-output layers.

MARS

By utilizing piecewise linear functions (also called basis functions), MARS generates a non-linear functional relationship between the output and input variables. In this regression technique, a non-linear relationship is obtained from two or more linear functions with the help of different gradients. The gradient changing locations of the linear segments are known as knots. The generalized form of the MARS model can be expressed by using Equation 2.

$$f(X) = \beta_0 + \sum_{m=1}^{M}\beta_m\lambda_m(X) \tag{2}$$

where $f(X)$ indicates the predicted magnitude from the MARS, β_0 and β_m are coefficients, M is utilized to indicate the total number of basis functions, and $\lambda_m(X)$ is the m^{th} basis function.

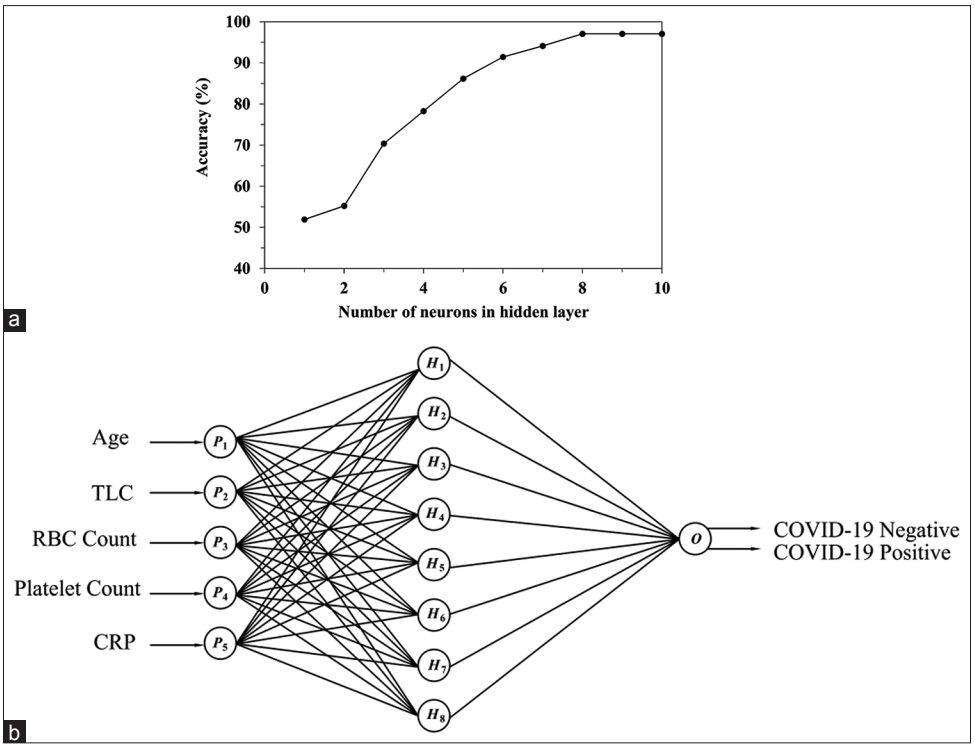

Figure 1: (a) Variation of accuracy with number of neurons in hidden layer; (b) architecture of neural network for classification

The entire methodology of MARS is divided into two main steps (i.e., the forward phase and the backward phase). In order to check the least contributing basis functions, generalized cross validation (GCV) technique is utilized. GCV is the ratio of mean squared error of the obtained model and a penalty component that accounts for the complexity of the model. The GCV can be expressed by the following equation:[40]

$$\mathrm{GCV} = \frac{\frac{1}{\mathrm{N}}\sum_{i=1}^{N}[y_i - f(X_i)]^2}{\left[1 - \frac{M + d \times \frac{(M-1)}{2}}{N}\right]^2} \quad (3)$$

where N is the total number of observed data, y_i is the observed value, $f(X_i)$ is the predicted value (from the MARS), and d is the penalty factor. Here, the value of d is considered as 2. The optimum MARS model is obtained using ARESLab Version 1.13.0 when the GCV value is the least. For a detailed mathematical background of the MARS, one can refer to Friedman.[40]

In the present MARS model, the training and testing data divisions are kept exactly the same as the ANN model. In order to find the optimum basis functions, Figure 2 is plotted. It is observed that the optimum number of basis functions remains constant at 14 when basis functions in forward phase are ≥ 107. Therefore, the present MARS model is formed with 14 basis functions.

Results

Prediction of COVID-19 infection using ANN

By using the obtained weights and biases, the prediction equation is presented as follows [refer to Figure 3(a)]:

$$CI_A = \mathrm{softmax}\left[\sum_{n=1}^{H} W_2 \mathrm{ReLU}\left(\sum_{j=1}^{P}\{W_1 NI_j\} + B_1\right) + B_2\right] \quad (4)$$

Here, NI_j designates the normalized input parameters. Normalization is carried out by using Equation (5).

$$NI_j = \left(\frac{i_j - i_j^{\min}}{i_j^{\max} - i_j^{\min}}\right) \quad (5)$$

where $i_j^{\max}$ and $i_j^{\min}$ are the maximum and minimum magnitudes of the input parameters, respectively, as presented in Table 1. Here, '*ReLU*' and '*softmax*' functions are utilized, which can be expressed as

Rectified linear unit function,

$$\mathrm{ReLU}\left(\sum_{j=1}^{P}\{W_1 NI_j\} + B_1\right) = \max\left[0, \left(\sum_{j=1}^{P}\{W_1 NI_j\} + B_1\right)\right] \quad (6)$$

$$\text{softmax}(I_{2i}) = \frac{\exp(I_{2i})}{\sum_{i=1}^{n} \exp(I_{2i})} \quad (7)$$

B_1 and B_2 indicate the biases, and as described before, W_1 and W_2 represent the weights acquired from the selected model for input-hidden and hidden-output layers, respectively [refer to Table 2]. The COVID-19 indicator (CI_4) designates the output class (i.e. COVID-19-positive and COVID-19-negative) based on the obtained probability for each class. The accuracy of the predicted class is found to be 97.06%, as shown in the confusion matrix [refer to Figure 3(b)].

The relative importance of the input parameters is determined using Equation 1 and presented in Figure 3(c). It is found that for detection of COVID-19 infection, CRP is the most sensitive parameter having *RI* as 46.85%, followed by age as 17.92%, RBC count as 15.76%, TLC as 13.36%, and platelet count as 6.11%.

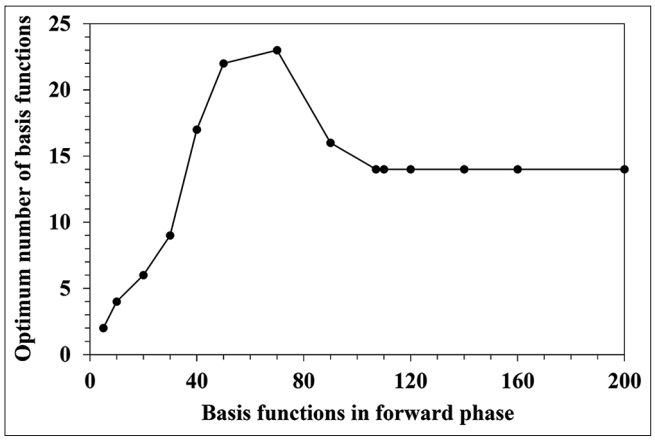

Figure 2: Selection of the optimum number of basis functions retained after pruning phase

Table 1: Maximum and minimum values of the input parameters considered in the present study

Input parameters	Maximum value	Minimum value
Age (years)	83	8
TLC (cells/mm^3)	10310	1800
RBC Count (million/mm^3)	6.1	3.19
Platelet Count (/mm^3)	441000	152000
CRP (mg/L)	67.7	1

Prediction of COVID-19 infection using MARS

The expressions of basis functions are given in Table 3. Note that here, x_1 = Age, x_2 = TLC, x_3 = RBC Count, x_4 = Platelet Count, and x_5 = CRP. The final expression of the MARS model is given by Equation 8. Also, it needs to be mentioned that in the present MARS model, the normalization of the input parameters is carried out by dividing each parameter with the maximum magnitude of the respective input parameter, which can be found in Table 1.

The performance of the MARS model is expressed in terms of accuracy, which is noticed from the confusion matrix from the testing data points. The accuracy of the predicted class is observed to be 91.18% [refer to Figure 4(a)]. Sensitivity analysis is carried out using the relative importance concept and ANOVA decomposition. Figure 4(b) shows the percentage relative importance of various input parameters (out of 100%). The percentage relative importance of any input parameter is calculated from the difference in GCV value when that parameter is discarded from the model. From the above analysis, it is clear that CRP is the most sensitive parameter having *RI* as 41.63%, followed by Age as 18.83%, RBC Count as 17.43%, TLC as 16.17%, and Platelet Count as 5.94%. Table 4, which shows the ANOVA decomposition, further confirms that CRP is the most significant parameter as it gives the highest GCV value.

Discussion

The proposed predictive models indicate that the most significant parameter is CRP, followed by Age, RBC Count, TLC, and Platelet Count for predicting COVID-19 infection. These results correlate well with several previous studies on COVID-19 diagnosis and severity of infection.

The role of each of the parameters included in our study has been well researched, and their importance in COVID-19 infections is well documented. Previous studies have established

Table 2: Obtained weights and biases for input-hidden and output-hidden layers

Hidden neuron	Weights (W1) between input-hidden layers					Weights (W2) between output-hidden layers		Biases		
	Age	TLC	RBC Count	Platelet Count	CRP	COVID-19 Negative	COVID-19 Positive	Input-hidden (B_1)	Output-hidden (B_2)	
									COVID-19 Negative	COVID-19 Positive
1	-0.53878	-0.10013	-0.21754	-0.41618	-0.31958	0.59416	0.61573	0.00000	-2.24037	2.24037
2	-1.59110	-2.00667	-1.18135	-1.27932	0.58080	1.07475	-1.34048	-2.49985		
3	0.73741	0.29942	0.53460	0.18354	2.09157	-2.60927	1.21355	1.47500		
4	0.43699	0.50722	0.46116	0.01140	-2.27116	4.96820	-5.03872	0.90528		
5	2.10788	1.22257	0.54100	0.10648	-2.42036	0.54710	-1.53634	-0.44693		
6	0.23700	-0.67835	-0.25276	-0.03518	0.25805	-0.00073	0.37486	-0.20694		
7	-0.36300	-0.36125	1.25278	-0.18778	0.76019	-1.06046	0.60625	1.77528		
8	0.51724	-0.18549	-0.20850	0.28777	-0.71397	0.93502	0.46072	0.23099		

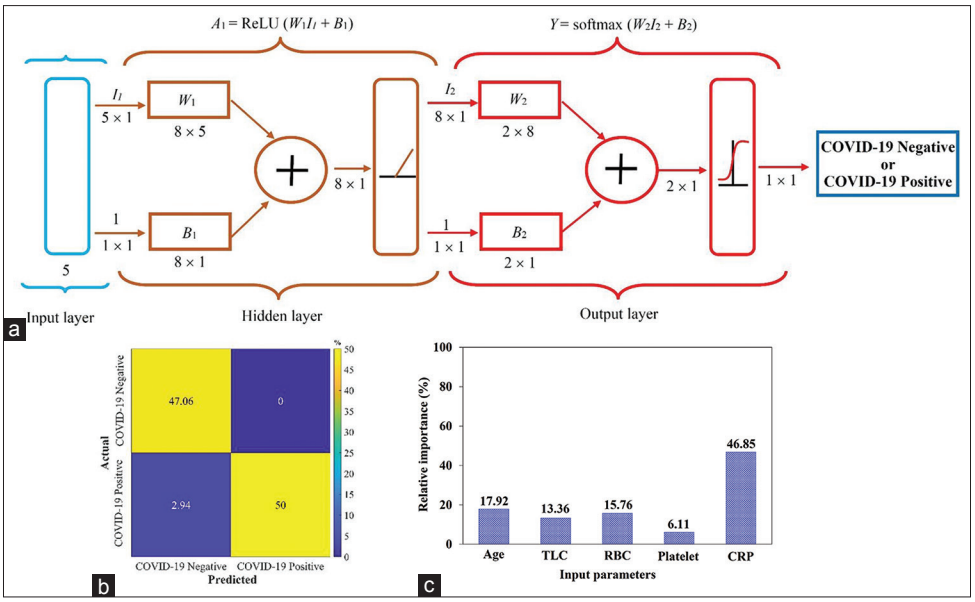

Figure 3: (a) Network with a detailed weight matrix; (b) confusion matrix; (c) relative importance of input variables for detecting COVID-19 infection as per selected ANN model

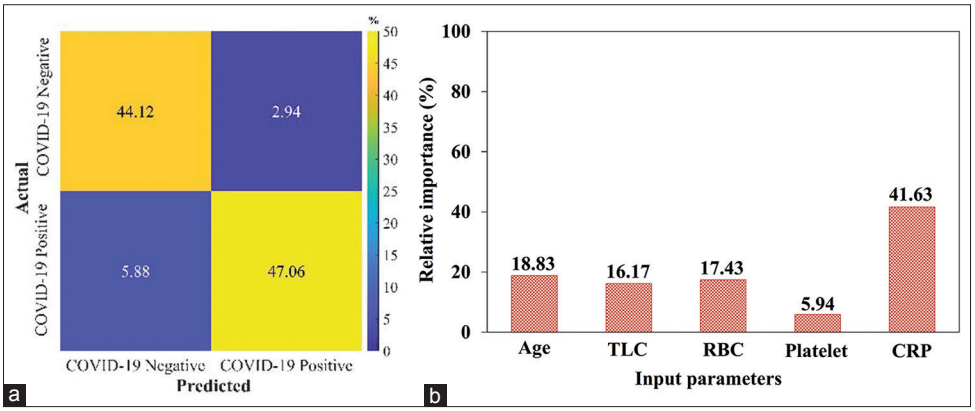

Figure 4: (a) Confusion matrix; (b) relative importance of input variables for detecting COVID-19 infection as per selected MARS model

that CRP [32-37], RBC [22-26], TLC [18-21], and increasing age [14-17] have a significant effect on the severity of COVID-19 infection. Platelets are less likely to have a significant role [27-31], which also matches with the results of the proposed ML models. Hence, the results obtained from the predictive models are medically justified. These predictive models shall prove to be of great aid to family physicians and primary care providers as they will be able to easily identify COVID infection in the patients with the help of these basic parameters which can be readily obtained.

Conclusion

The present study deals with two ML-based classification models, namely, ANN and MARS, to predict the COVID-19 infection cheaply and quickly. The study was carried out by utilizing laboratory data from an in-campus semi-urban hospital. The developed design models predict the COVID-19 infection with a reasonable accuracy of 97.06% and 91.18% for ANN and MARS, respectively. Additionally, two very simple predictive mathematical equations are developed for each ML model. Sensitivity analysis of input laboratory parameters indicates that CRP is the most sensitive parameter. It is expected that the present study will be useful to the medical practitioners.

Acknowledgment

Authors would like to thank the authorities of the B. C. Roy Technology Hospital, Indian Institute of Technology Kharagpur for allowing to conduct the study using the patients' data. Authors would also like to thank the staffs and laboratory technicians of the hospital for helping us gather all the test data.

Ethical approval

Retrospective data were utilised for this study, patients' identities are not revealed and for carrying out the study permission was taken from the Head of the hospital.

Table 3: Equations of basis functions of the developed MARS model

Basis Function	Equation
λ_1	max (0, x_5-0.15953)
λ_2	max (0, 0.15953 - x_5) × max (0, x_2-0.5257) × max (0, 0.55422 – x_1) × max (0, x_3 – 0.64754)
λ_3	λ_2 × max (0, x_4 – 0.70522)
λ_4	λ_2 × max (0, 0.70522 – x_4)
λ_5	max (0, 0.15953 – x_5) × max (0, 0.55422 – x_1) × max (0, x_4 – 0.56009) × max (0, x_2 – 0.6935) × max (0, 0.74098 – x_3)
λ_6	max (0, 0.15953 – x_5) × max (0, x_2 – 0.5257) × max (0, 0.55422 – x_1) × max (0, x_4 – 0.62812)
λ_7	max (0, 0.15953 – x_5) × max (0, 0.5257 – x_2) × max (0, x_1 – 0.54217)
λ_8	max (0, 0.15953 – x_5) × max (0, 0.55422 – x_1) × max (0, 0.80656 – x_3) × max (0, 0.82929 – x_2)
λ_9	max (0, 0.15953 – x_5) × max (0, 0.55422 – x_1) × max (0, 0.56009 – x_4) × max (0, x_2 – 0.70514) × max (0, x_3 – 0.72951)
λ_{10}	max (0, 0.10635 – x_5)
λ_{11}	max (0, x_5-0.050222)
λ_{12}	max (0, 0.15953 – x_5) × max (0, x_4 – 0.46712) × max (0, 0.36145 – x_1) × max (0, 0. 64985 – x_2) × max (0, x_3 – 0.80328)
λ_{13}	λ_{10} × max (0, 0. 31325 – x_1)
λ_{14}	max (0, 0.15953 – x_5) × max (0, 0.5257 – x_2) × max (0, 0.54217 – x_1) × max (0, x_3 – 0.62295)

Note: Here, x_1=Age, x_2=TLC, x_3=RBC Count, x_4=Platelet Count, x_5=CRP

$$CI = 0.36069 - 5.8224\lambda_1 - 454.68\lambda_2 - 18316\lambda_3 + 3182.5\lambda_4 + 3.3779e^5\lambda_5 + 1372.8\lambda_6 + 225.08\lambda_7 + 183.31\lambda_8 - 19381\lambda_9 - 4.8511\lambda_{10} + 5.8272\lambda_{11} + 1.6884e^5\lambda_{12} + 47.76\lambda_{13} + 1241.2\lambda_{14} \quad (8a)$$

if $CI < 0.5$

CI_M = COVID - 19 Negative

else

CI_M = COVID - 19 Positive (8b)

Table 4: ANOVA decomposition of the present MARS model

Function	GCV	Variable (s)
1	0.287453	x_5
2	0.093000	x_1, x_5
3	0.100862	x_1, x_2, x_5
4	0.147816	x_1, x_2, x_3, x_5
5	0.105504	x_1, x_2, x_4, x_5
6	0.113684	x_1, x_2, x_3, x_4, x_5

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Rohlfing AK, Rath D, Geisler T, Gawaz M. Platelets and COVID-19. Hamostaseologie 2021;41:379–85.
2. Rahman T, Khandakar A, Abir FF, Faisal MAA, Hossain MS, Podder KK, *et al.* QCovSML: A reliable COVID-19 detection system using CBC biomarkers by a stacking machine learning model. Comput Biol Med 2022;143:105284.
3. Gupta R, Ghosh A, Singh AK, Misra A. Clinical considerations for patients with diabetes in times of COVID-19 epidemic. Diabetes Metab Syndr 2020;14:211-2.
4. Oran DP, Topol EJ. Prevalence of asymptomatic SARS-CoV-2 infection. Ann Intern Med 2020;174:362–8.
5. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, *et al.* Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: A descriptive study. Lancet Infect Dis 2020;20:425–34.
6. Letícia de Oliveira Toledo S, Sousa Nogueira L, das Graças Carvalho M, Romana Alves Rios D, de Barros Pinheiro M. COVID-19: Review and hematologic impact. Clin Chim Acta 2020;510:170–6.
7. Zheng Y, Zhang Y, Chi H, Chen S, Peng M, Luo L, *et al.* The hemocyte counts as a potential biomarker for predicting disease progression in COVID-19: A retrospective study. Clin Chem Lab Med 2020;58:1106–15.
8. Vogels CBF, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC, *et al.* Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT–qPCR primer–probe sets. Nat Microbiol 2020;5:1299–305.
9. Rasheed J, Jamil A, Hameed AA, Al-Turjman F, Rasheed A. COVID-19 in the age of artificial intelligence: A comprehensive review. Interdiscip Sci Comput Life Sci 2021;13:153–75.
10. Chadaga K, Prabhu S, Vivekananda Bhat K, Umakanth S, Sampathila N. Medical diagnosis of COVID-19 using blood tests and machine learning. J Phys Conf Ser 2022;2161:012017.
11. Khanna VV, Chadaga K, Sampathila N, Prabhu S, Chadaga R, Umakanth S. Diagnosing COVID-19 using artificial intelligence: A comprehensive review. Netw Model Anal Health Inform Bioinform 2022;11:25.
12. Lin JK, Chien TW, Wang LY, Chou W. An artificial neural network model to predict the mortality of COVID-19 patients using routine blood samples at the time of hospital admission: Development and validation study. Medicine (Baltimore) 2021;100:E26532.
13. Babaei Rikan S, Sorayaie Azar A, Ghafari A, Bagherzadeh Mohasefi J, Pirnejad H. COVID-19 diagnosis from routine blood tests using artificial intelligence techniques. Biomed Signal Process Control 2022;72:103263.

14. Aviv A. Telomeres and COVID-19. Faseb J 2020;34:7247.
15. Naylor K, Li G, Vallejo AN, Lee W-W, Koetz K, Bryl E, *et al.* The influence of age on T cell generation and TCR diversity. J Immunol 2005;174:7446–52.
16. Hu D, Lou X, Meng N, Li Z, Teng Y, Zou Y, *et al.* Influence of age and gender on the epidemic of COVID-19: Evidence from 177 countries and territories—An exploratory, ecological study. Wien Klin Wochenschr 2021;133:321–30.
17. Tiruneh SA, Tesema ZT, Azanaw MM, Angaw DA. The effect of age on the incidence of COVID-19 complications: A systematic review and meta-analysis. Syst Rev 2021;10:1–9.
18. Channappanavar R, Perlman S. Pathogenic human coronavirus infections: Causes and consequences of cytokine storm and immunopathology. Semin Immunopathol 2017;39:529–39.
19. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, *et al.* A novel Coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;382:727–33.
20. Henry BM, De Oliveira MHS, Benoit S, Plebani M, Lippi G. Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): A meta-analysis. Clin Chem Lab Med 2020;58:1021–8.
21. Feng X, Li S, Sun Q, Zhu J, Chen B, Xiong M, *et al.* Immune-inflammatory parameters in COVID-19 cases: A systematic review and meta-analysis. Front Med 2020;7:231.
22. Elderdery AY, Elkhalifa AME, Alsrhani A, Zawbaee KI, Alsurayea SM, Escandarani FK, *et al.* Complete blood count alterations of COVID-19 patients in Riyadh, Kingdom of Saudi Arabia. J Nanomater 2022;2022:6529641.
23. Yuan X, Huang W, Ye B, Chen C, Huang R, Wu F, *et al.* Changes of hematological and immunological parameters in COVID-19 patients. Int J Hematol 2020;112:553–9.
24. Berzuini A, Bianco C, Migliorini AC, Maggioni M, Valenti L, Prati D. Red blood cell morphology in patients with COVID-19-related anaemia. Blood Transfus 2021;19:34-6.
25. Atnaf A, Shiferaw AA, Tamir W, Akelew Y, Toru M, Tarekegn D, *et al.* Hematological profiles and clinical outcome of COVID-19 among patients admitted at Debre Markos Isolation and Treatment Center, 2020: A prospective cohort study. J Blood Med 2022;13:631–41.
26. Lanini S, Montaldo C, Nicastri E, Vairo F, Agrati C, Petrosillo N, *et al.* COVID-19 disease—Temporal analyses of complete blood count parameters over course of illness, and relationship to patient demographics and management outcomes in survivors and non-survivors: A longitudinal descriptive cohort study. PloS One 2020;15:e0244129.
27. Giannis D, Ziogas IA, Gianni P. Coagulation disorders in coronavirus infected patients: COVID-19, SARS-CoV-1, MERS-CoV and lessons from the past. J Clin Virol 2020;127:104362.
28. Wang T, Chen R, Liu C, Liang W, Guan W, Tang R, *et al.* Attention should be paid to venous thromboembolism prophylaxis in the management of COVID-19. Lancet Haematol 2020;7:e362–3.
29. Xu XW, Wu XX, Jiang XG, Xu KJ, Ying LJ, Ma CL, *et al.* Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: Retrospective case series. BMJ 2020;368:m606.
30. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. Lancet 2020;395:507–13.
31. Hana C, Aboulenain S, Dewaswala N, Narendran V. Does thrombocytopenia truly correlate with COVID-19 severity? Blood 2020;136:39–40.
32. Wang L. C-reactive protein levels in the early stage of COVID-19. Med Mal Infect 2020;50:332–4.
33. Sahu BR, Kampa RK, Padhi A, Panda AK. C-reactive protein: A promising biomarker for poor prognosis in COVID-19 infection. Clin Chim Acta 2020;509:91–4.
34. Sadeghi-Haddad-Zavareh M, Bayani M, Shokri M, Ebrahimpour S, Babazadeh A, Mehraeen R, *et al.* C-Reactive protein as a prognostic indicator in COVID-19 patients. Interdiscip Perspect Infect Dis 2021;2021:5557582.
35. Luo X, Zhou W, Yan X, Guo T, Wang B, Xia H, *et al.* Prognostic value of C-Reactive protein in patients with coronavirus 2019. Clin Infect Dis 2020;71:2174–9.
36. Yitbarek GY, Walle Ayehu G, Asnakew S, Ayele FY, Bariso Gare M, Mulu AT, *et al.* The role of C-reactive protein in predicting the severity of COVID-19 disease: A systematic review. SAGE Open Med 2021;9:20503121211050755.
37. Tan C, Huang Y, Shi F, Tan K, Ma Q, Chen Y, *et al.* C-reactive protein correlates with computed tomographic findings and predicts severe COVID-19 early. J Med Virol 2020;92:856–62.
38. 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.
39. Garson GD. Interpreting neural-network connection weights. AI Expert 1991;6:46–51.
40. Friedman JH. Multivariate adaptive regression splines. Ann Stat 1991;19:1–67.