Research on Biomedical Engineering
https://doi.org/10.1007/s42600-020-00103-6

ORIGINAL ARTICLE

Machine learning based prognostic model and mobile application software platform for predicting infection susceptibility of COVID-19 using healthcare data

R. Srivatsan[1] · Prithviraj N. Indi[1,2] · Swapnil Agrahari[3] · Siddharth Menon[1] · S. Denis Ashok[4]

Received: 11 July 2020 / Accepted: 6 October 2020

Abstract

Introduction From public health perspectives of COVID-19 pandemic, accurate estimates of infection severity of individuals are extremely valuable for the informed decision-making and targeted response to an emerging pandemic. This paper presents machine learning based prognostic model for providing early warning to the individuals for COVID-19 infection using the healthcare dataset. In the present work, a prognostic model using random forest classifier and support vector regression is developed for predicting the infection susceptibility probability (ISP) score of COVID-19, and it is applied on an open healthcare dataset containing 27 field values. The typical fields of the healthcare dataset include basic personal details such as age, gender, number of children in the household, and marital status along with medical data like coma score, pulmonary score, blood glucose level, HDL cholesterol, etc. An effective preprocessing method is carried out for handling the numerical and categorical values (non-numerical) and missing data in the healthcare dataset. The correlation between the variables in the healthcare data is analyzed using the correlation coefficient, and heat map with a color code is used to identify the influencing factors on the infection susceptibility probability (ISP) score of COVID-19. Based on the accuracy, precision, sensitivity, and F-scores, it is noted that the random forest classifier provides an improved classification performance as compared to support vector regression for the given healthcare dataset. Android-based mobile application software platform is developed using the proposed prognostic approach for enabling the healthy individuals to predict the susceptibility infection score of COVID-19 to take the precautionary measures. Based on the results of the proposed method, clinicians and government officials can focus on the highly susceptible people for limiting the pandemic spread.

Methods In the present work, random forest classifier and support vector regression techniques are applied to a medical healthcare dataset containing 27 variables for predicting the susceptibility score of an individual towards COVID-19 infection, and the accuracy of prediction is compared. An effective preprocessing is carried for handling the missing data in the healthcare dataset. Correlation analysis using heat map is carried on the healthcare data for analyzing the influencing factors of infection susceptibility probability (ISP) score of COVID-19. A confusion matrix is calculated for understanding the performance of classification based on the number of true-positives, true-negatives, false-positives, and false-negatives. These values further used to calculate the accuracy, precision, sensitivity, and F-scores.

Results From the classification results, it is noted that the random forest classifier provides a classification accuracy of 99.7%, precision of 99.8%, sensitivity of 98.8%, and F-score of 99.29% for the given medical dataset.

Conclusion Proposed machine learning approach can help the individuals to take additional precautions for protecting people from the COVID-19 infection, and clinicians and government officials can focus on the highly susceptible people for limiting the pandemic spread.

Keywords Machine learning · Prognostics · COVID-19 · Infection susceptibility · Mobile application tool · Random forests · Support vector regression

✉ S. Denis Ashok
denisashok@vit.ac.in

R. Srivatsan
r.srivatsan2017@vitstudent.ac.in

Prithviraj N. Indi
prithvirajn.indi2017@vitstudent.ac.in

Swapnil Agrahari
swapnil.agrahari2017@vitstudent.ac.in

Siddharth Menon
siddharth.menon2018@vitstudent.ac.in

Extended author information available on the last page of the article

Published online: 12 November 2020

Abbreviation

ISP	Infection susceptibility probability
SVR	Support vector regression
FT	Foreign trips
RF	Random forest
DT	Decision trees
RBF	Radial basis function
HO	Health officer

Introduction

The recent outbreak of coronavirus disease 2019 (COVID-19) has created a great challenge for the healthcare system (Hui et al. 2020). Considering the lethal nature of COVID-19 outbreak and its worldwide spread, the World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC) at different nations have provided provisional guidelines for protecting people from getting affected and preventing the further spread of COVID-19 virus from infected individuals. RT-PCR tests from deep nasotracheal samples and chest CT scan are commonly used for definitive diagnosis of COVID-19 (Repici et al. 2020). Due to the quick spread of COVID-19, physicians in the healthcare systems are facing extreme difficulty in the physical examination and analysis of subsequent para clinical healthcare data for the accurate diagnosis of COVID-19. Hence, it is necessary develop software tools for easier way for interpreting the large-scale health dataset which can help the government and healthcare officials for quicker decision-making during the COVID-19 pandemic situations.

With the capability of interpreting the hidden and complex patterns from huge, noisy, or complex data, artificial intelligence and machine learning techniques can play a major role in combating the COVID-19 pandemics. Few works have been reported by the researchers on use of machine learning techniques for the prediction and diagnosis of epidemics (Wynants et al. 2020). An artificial intelligence-based rapid diagnosis approach for COVID-19 patients was developed using the analysis of chest X-ray images (Mei et al. 2020). An artificial intelligence-based prediction model of the epidemics trend of COVID-19 is proposed by Yang et al. (2020a, b). Linear regression model is used for time series prediction of COVID-19 outbreak (Pandey et al. 2020). Mechanistic models have been reported to predict COVID-19 outbreak in real time (Liu 2020). K means algorithm is applied to categorize the countries based on the number of confirmed COVID-19 cases (Carrillo-Larco and Castillo-Cara 2020). XGBoost machine learning model is proposed to estimate the survival ratio of severely ill COVID-19 patients (Yan 2020). A classification using Fourier and Gabor methods is applied on dataset of COVID-19 (Al-Karawi 2020). Multilayered perception (MLP) and adaptive network-based fuzzy inference system are used for predicting (Metsky 2020). Support vector machine is applied to detect severely ill COVID patients from mild symptom COVID patients (Tang 2020). Convolutional neural network frameworks have been proposed to detect COVID-19 from chest X-ray images (Narin et al. 2020). A prediction model for the propagation analysis of the COVID-19 is proposed by Li et al. (2020). An interpretable mortality prediction model for COVID-19 patients is developed using the healthcare dataset (Yan et al. 2020).

It is found that many machine learning approaches have been successfully implemented for the prediction and diagnostic purposes of COVID-19 using the clinical and healthcare data. However, prognostic frameworks for early prediction of COVID-19 infection are found to be limited which can be helpful to take proactive measures to combat the virus spread. Random forest and support vector machine algorithms are found to be popular in achieving the satisfactory results for the different prediction applications. Hence, this paper presents random forest and support vector machine algorithms based on prognostic approach for predicting the infection susceptibility score for each individual using the healthcare data. The novelty of the proposed approach is the identification of the infection susceptibility prior to infection so that the regulative and preventive rules can be made for the individuals.

Methods

Recently, machine learning techniques have been applied for prognostic applications like prediction of disease symptoms, risks, survivability, and recurrence (Adnan Qayyum et al. 2020). In the present work, a machine learning based prognostic approach is presented for predicting the infection susceptibility probability (ISP) score of COVID-19 and categorizing the healthy individuals as low, medium, and high using open healthcare dataset. Among the various machine learning techniques, random forest and support vector regression have been considered for the present application due to their superior classification performance than other techniques such as linear regression and neural networks. Proposed prognostic approach is further developed as a strategic decision-making tool by involving a versatile database and a mobile application which enables the healthy individuals for taking precautionary measures and also government officials in prioritizing the resources in the hospital settings. Fig. 1 shows the major elements of the proposed approach, and it is described below:

- A healthcare survey dataset includes the demographic, epidemiological characteristics and underlying comorbidities of the individuals with the target classes of risk factors, namely, high risk (66–100%), medium risk

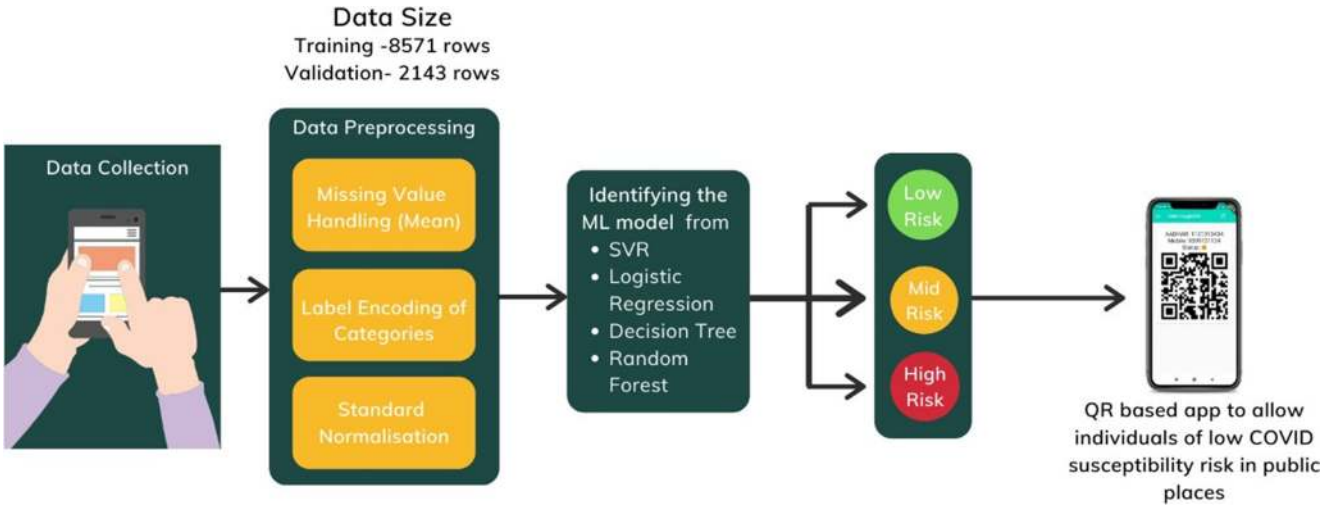

Fig. 1 Proposed prognostic approach for predicting the susceptibility score of individuals using healthcare data

(33–66%), and low risk (0–33%). As the collected healthcare data may contain the missing values during the data collection, an effective preprocessing is essential, and it is carried out before it is applied to the machine learning models for the classification applications.

- Supervised machine learning techniques consisting of random forest, support vector regression, linear regression, and neural networks for predicting infection susceptibility probability (ISP) score of COVID-19 using the labeled healthcare data
- A centralized data collection system and android mobile application which can be useful for sharing the multimodal healthcare data and predicting the infection susceptibility score to the individuals, healthcare professionals, and government administrative officials.

Description of healthcare dataset

In the present work, open healthcare dataset containing the demographic and epidemiological characteristics and underlying comorbidities of the individuals is used for demonstrating the proposed prognostic approach, and it is available in the online repository Kaggle (Singh 2020).The dataset contains 14,498 rows and 27 columns. Table 1 shows the typical fields of the healthcare dataset which include basic personal details such as age, gender, number of children in the household, and marital status along with medical data like coma score, pulmonary score, blood glucose level, and HDL cholesterol. Medical data chiefly includes comorbidity conditions such as severe acute respiratory infections (SARI), diabetes, and heart syndromes. Vitals such as heart rate have also been considered in the modeling of the predictor infection susceptibility score of COVID-19.

Data preprocessing and preparation

It is noted that the healthcare survey data is multimodal as it contains many numerical, categorical values (non-numerical), and many machine learning algorithms cannot handle data in this form. Also, there can be missing values in the relevant fields of the dataset which will reduce the accuracy of machine learning algorithms. In order to overcome these difficulties, data preprocessing and preparation is essential, and it is carried out on the healthcare dataset using label encoding to achieve accurate results using the machine algorithms.

Data normalization

Standardization of the dataset makes a very crucial role in the pipeline of the ML model since if the individual features do not reassemble standard normal distribution of data points, the model would become erratic in its predictions. This involves a technique of reducing mean value from each individual data point and performing a scaling operation to them in order to obtain unit variance per cell of the data.

$$x' = \frac{x-\mu}{\sigma} \tag{1}$$

where μ denotes data's mean value and σ denotes the standard deviation obtained as square root of variance.

Correlation analysis of healthcare data using heat map

With many fields involving demographic and epidemiological characteristics and underlying comorbidities of the individuals, a typical healthcare dataset formulates a high-dimensional feature space which has the strong and weak relevance to the infection susceptibility probability (ISP)

Table 1 The various fields of the medical and healthcare data (Singh 2020)

Variables	Description
people_ID	Unique ID for each person
Region	The area that the person belongs to
Gender	Gender of the person
Designation	Designation of the person
First_Name	Name of individual
Married	Marital status of individual
Children	Number of children in the family
Occupation	Sector of individual occupation
Mode_transport	Mode of transport that the individual frequently chooses to travel
cases/1M	Number of confirmed cases per 1 million population in that region
Deaths/1M	Number of death case per 1 million population in that region
comorbidity	Co-occurring medical condition
Age	Age of the person
Coma score	Neurological coma score
Pulmonary score	Pulmonary PaO2 (mmHg)/FiO2
Cardiological pressure	Cardiological mean systolic arterial pressure (mmHg)
Diuresis	Diuresis in mL/day
Platelets	Hematological platelets 10/L
HBB	Hepatic blood bilirubin (μmol/L)
d-dimer	d-dimer concentration in the blood (ng/mL)
Heart rate	Number of times a person's heart beats per minute
HDL cholesterol	High-density lipoprotein level (milligrams per deciliter)
Charlson index	index for a patient who may have any of the listed comorbid disease conditions
Blood glucose	strength of glucose present in the blood (millimoles per liter)
Insurance	Medical insurance spending cover (in Rs.)
Salary	Annual salary of the individual
FT/month	Average foreign trips taken by the individual per month, considering last 2 year data

score. As the highly correlated features add noise and inaccuracy to the machine learning model, it is necessary to analyze the correlation between the variables in the healthcare data. In this work, correlation analysis of the healthcare data is carried using the correlation coefficient and heat map with a color code which is used to visualize it. Based on the statistical analysis of healthcare data, correlation coefficient is calculated as the ratio of covariance, and the standard deviation between two feature sets a_i and b_i is given by the following formula:

$$C = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - b)}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \quad (2)$$

where n is the sample size, a_i and b_i are the i_{th} data values, and a and b are the mean values. The value of the coefficient (C) ranges between − 1 and + 1. The values of the correlation close to + 1 indicate the strong positive correlation, those close to − 1 show strong negative correlation, and those closest to 0 show no relation.

Machine learning techniques for predicting the infection susceptibility probability score of COVID-19

Based on the statistical analysis of the healthcare data with the relevant information on the severity of the COVID-19 infection, binary logistic regression model is fitted to the training sample, and the coefficients of the regression model are used for calculating the probability of infection (Menelaos Pavlou et al. 2015; Chen et al. 2017). In present work, supervised machine learning techniques such as random forest, support vector regression, linear regression, and neural networks have been applied for predicting and categorizing the susceptibility score of COVID-19 infection using the healthcare data. With the improved accuracy of classification, random forest and support vector regression are considered in the present work, and it is explained below:

Random forest classifier

With interpretable decision logic, random forest algorithm (RF) is found to be one of the most promising classifier which uses multiple decision trees (DT) to train and predict data samples. The general structure of random forest with multiple decision trees is shown in Fig. 2. The multiple ensemble DTs give rise to different classifications of infection susceptibility score. Here the value of n is chosen to be between 10 and 20 for optimum prediction. The majority scheme of vote is the terminal deciding factor of the model decision and throws the actual predicted class of the ISP of the individual as low, medium, and high.

In the random forests classification approach, the ensemble of decision trees (DT) is involved in calculating the GINI score as in Eq. (3).

$$\mathrm{ni_j} = \mathrm{w_j C_j} - \mathrm{w_{left(j)} C_{left(j)}} - \mathrm{w_{right(j)} C_{right(j)}} \tag{3}$$

Here, the notations of the parameters are as follows: $\mathrm{ni_j}$ refers to the significance of the node indexed j, $\mathrm{W_j}$ is indicating the weighted number of samples approaching the node indexed j, $\mathrm{C_j}$ indicates the impurity value of node indexed j, left(j) denotes the left child node from node indexed j, and right(j) shows the right child node from node indexed j.

The second step is to obtain the importance given by each feature of the DT. This significance parameter can be computed using Eq. (4):

$$fi_i = \frac{\sum j : \text{node } j \text{ splits on feature } i^{ni} j}{\sum_{k \in \text{all nodes}} ni_k} \tag{4}$$

where fi_i denotes the importance of feature indexed i and the ni_j refers to the importance of node indexed j.

$$\text{norm } fi_i = \frac{f\ i_i}{\sum_{j \in \text{all features}} fi_j} \tag{5}$$

These features fi_i are now normalized using the equation:

$$RFf\,i_i = \frac{\sum_{j \in all\ trees} fi_{ij}}{T} \tag{6}$$

Then we can obtain the final feature of importance as mean of those of all the DTs as given in Fig. 2, where $RF\ f_i$ refers to the importance of feature indexed i computed through all DTs in the RF and norm fi_{ij} refers to the normalized feature significance parameter for index i in the DT indexed j and T indicates the total number of DTs.

Support vector regression

In the present work, support vector regression is applied for predicting the susceptibility score of COVID-19 infection as a continuous variable. Due to very high nonlinearity in the PCs, a support vector regression based approach was chosen to be employed for obtaining probabilities of risk of infection using the medical dataset.

Assuming that the set of training medical data x_n is a multivariate set of N observations with observed response values y_n, a linear function is established as given below:

$$y = f(x) = < w, x > + b = \sum_{J=1}^{M} w_j x_j + b, y, b \in R, x, w \in R^M \tag{7}$$

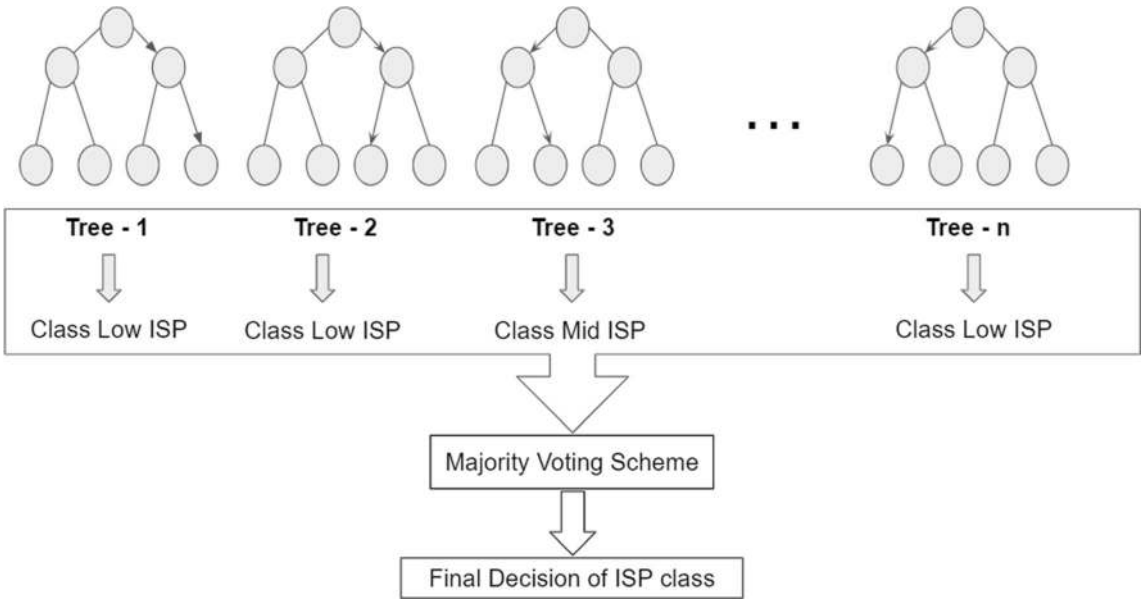

Fig. 2 Illustration of the random forest architecture

$$f(x) = [w\ b\ ]^r[x\ 1\ ] = w^r x + b\ x, w \in R^{M+1} \quad (8)$$

In the above equation x is a multidimensional input vector, with bias b and normal vector w. To ensure that it is as flat as possible, $f(x)$ with the minimal norm value, a convex optimization problem is formulated to minimize the following:

$$\min_w \frac{1}{2}\|w\|^2 \quad (9)$$

This shows that the normal vector should be approximated during the process. Magnitude of weights is usually interpreted as flatness to the function obtained in the computation.

$$f(x, w) = \sum_{i=1}^{M} w_i \quad x^i, x \in R, w \in R^M \quad (10)$$

To minimize the loss between the actual and predicted value which is a major constraint, SVR adopts epsilon-insensitive loss function. Although asymmetrical loss functions should be used to avoid underestimation and overestimation, the functions used are usually convex in nature.

Since most of COVID data is asymmetrical, linear methods would not provide accurate results. Nonlinear methods in support vector regression (SVR) are handled by mapping the features to higher dimensional space called kernels.

$$\min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \varepsilon_i + \varepsilon_i^* \quad (11)$$

To achieve higher accuracy, we replace all instances of x with K (xi, xj) from the earlier linear formula which leads to primal formulation shown in the above equation. The transformation of features to kernel space is shown in the above equation.

Prototype implementation of the proposed approach using a mobile application

In order to demonstrate the proposed approach for predicting the infection susceptibility probability (ISP) score of COVID-19 of healthy individual, a mobile application software tool with user interface is developed for collecting relevant healthcare data and categorizing the risk of infections as low, medium, and high using random forest classifier ML algorithm. Figure 3 shows the various elements of mobile application involving the central database, interfaces for various stakeholders like health officer, and filed workers along with predicted ISP score. The mobile application is interfaced with a versatile database to enable the government administrative and health officer to select individuals for further quarantine and infection elimination protocol execution.

The details of the mobile application with the functional modules are given below:

1. A centralized data collection system with a graphical user interface in every containment zone, to consolidate and upload data collected from field workers, doctors, and non-COVID lab tests in a single standardized format.
2. A live geo location tracking that integrates easily with existing open maps and APIs to deliver low power intensive tracking of smartphones using GPS geocoding and decoding techniques.
3. Entry Point Infection Check (EPIC) allows real-time risk assessment through infection susceptibility probability (ISP) score by the proposed machine learning model which produces the QR codes. This can be verified at the public entry points by the concerned health officers (HO).

Based on the risk factor score and geo location tracking system in the mobile application, proposed mobile application software tool can be useful contact tracing, tracking of individuals in public places. Also, it can be used for entry point checking tool for screening the individuals based on the risk score of the individuals. The source code for the proposed android mobile application is available in Github digital repository.

Performance metrics of machine learning techniques

In order to analyze the classification performance of the different machine learning techniques, accuracy, precision, sensitivity, and F-scores are calculated using number of true-positives, true-negatives, false-positives, and false-negatives of the classification.

Accuracy The most common metric for performance assessment measured as a ratio of the number of correctly predicted data to the total number of data provided for testing as calculated by Eq. 12:

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

Precision It is a quantitative measure to describe the number of correct instances of prediction as compared to the total number of instances provided to the model during testing. This is calculated by Eq. 13.

$$\frac{TP}{TP + FP} \quad (13)$$

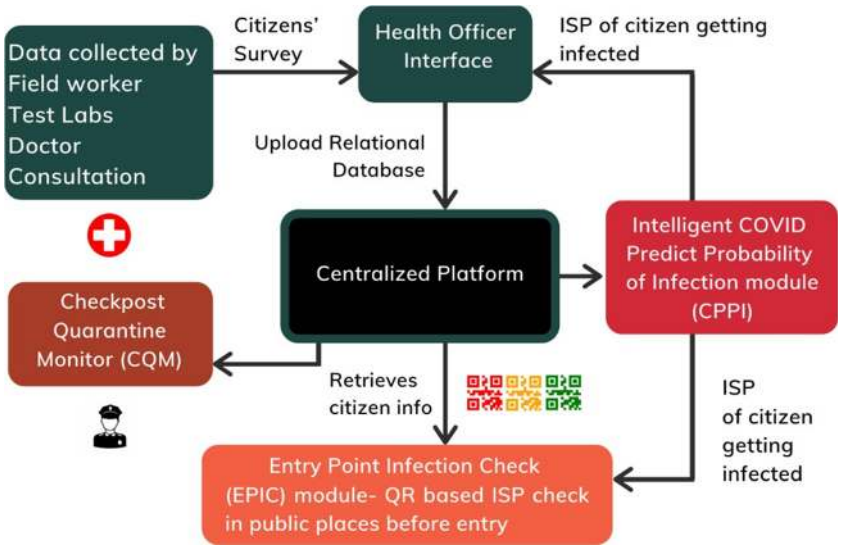

Fig. 3 Modules of mobile application platform and different interfaces

Sensitivity It is measure of ability of the machine learning model to predict the positive labels to all the given labels that should have been predicted positive, and it is calculated by Eq. 14.

$$\frac{TP}{TP + FN} \tag{14}$$

F-score This is measured by weighted average between precision and recall as calculated by Eq. 15.

$$\frac{TP}{TP + 0.5*(FP + FN)} \tag{15}$$

Here TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

Results

In the present work, proposed approach is implemented in Python computing and programming environment using the major computing libraries and mathematical functions such as Numpy, Pandas, and Scikit learn in Jupyter Notebook environment involving various machine learning algorithms such as random forest, logistic regression, support vector regression, linear regression, and neural networks. An open healthcare dataset with the infection susceptibility probability score available in the online repository Kaggle is used for training, validation, and testing purposes (Singh 2020).The sample healthcare survey dataset is shown in Fig. 4. It can be seen that the dataset contains categorical data and numerical data.

The hardware specifications of training, testing, and inference include a processor of Intel Xeon CPU at 2.20 GHz and a memory of 0.88 GB. The given healthcare dataset is preprocessed for overcoming the errors due to the missing data. It was found that random forest model was best among the machine learning models for the supervised classification of target classes of infection probability, and the inference timing for 10,000 data was found to be 2.85704 s.

Influencing factors of infection susceptibility probability score of COVID-19

As the healthcare dataset contains many fields involving epidemiological characteristics, and underlying comorbidities of the individuals, a heat map with a color and magnitude variation is developed to quickly check correlations and visualizing the correlation matrix as shown in Fig. 5.

This heat map shows the correlation coefficient which varies from +1 to −1 for providing an understanding of the vital and detrimental correlated factors of healthcare data on the infection probability. From the correlation matrix, four main factors influencing infection probability (ISP) are found to be "No of children" (0.42), "Cases per Million" (0.21), "Deaths per Million" (0.24), and "Platelet Count" (0.12). These results indicate that demographic fields are critical in influencing the infection susceptibility score of the individual. It can be seen that the gender and married fields also have an influence on the results which can be quantitatively analyzed from the correlation matrix. Unnecessary fields such as Name, Insurance, Salary, and People_ID are found to be non-contributing to our analysis and prediction; thus, they are removed during data preparation. We can reduce the correlation matrix of the dataset in

Fig. 4 Sample healthcare dataset (Singh 2020)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
people_ID	Region	Gender	Designation	Name	Married	Children	Occupation	Mode_transport	Cases/1M	Deaths/1M	Comorbidity	Age	Coma Score	Pulmonary Score
1	Bhuban	Female	Mrs.	mansi	YES	1	Farmer	Public	2	0	Hypertension	68	8	<400
2	Bhuban	Female	Mrs.	riya masi	YES	2	Farmer	Walk	2	0	Diabeties	64	15	<100
3	Bhuban	Female	Mrs.	sunita	NO	1	Cleaner	Public	2	0	None	19	13	<300
4	Bhuban	Female	Mrs.	anjali	YES	1	Driver	Car	2	0	Coronary Hear	33	9	<200
5	Bhuban	Female	Mrs.	champa k	NO	2	Manufacturer	Car	2	0	Diabeties	23	7	<400

	15	16	17	18	19	20	21	22	23	24	25	26
people_ID	Cardio Logica	Diuresis	Platelets	HBB	d-dimer	Heart Rate	HDL Choles	Charlson	Blood Glucose	Insurance	salary	FT/month
1	Normal	441	154	96	233	82	58	27	7	3600000	1E + 06	2
2	Stage-0		121	56	328	89	68	5	6	1600000	400000	1
3	Elevate	416	124	137	213	77	43	40	6	3400000	900000	1
4	Stage-0	410	98	167	275	64	60	27	7	700000	2E + 06	1
5	Normal	390	21	153	331	71	64	32	7	3200000	1E + 06	1

which 4 features have negative correlation with the output label.

Classification performance of RF classifier

In the random forest model, the decision trees developed over 100–200 nodes indicating the complex relationships mapped between the dataset and the target classes such as low, medium, and high infection susceptibility probability (ISP) score of COVID-19 infection. A confusion matrix is calculated for understanding the performance of classification based on the number of true-positives, true-negatives, false-positives, and false-negatives. These values further used to calculate the accuracy, precision, sensitivity, and F-scores. Inferring from the confusion matrix as given in Fig. 6, proposed model classifies the different groups of people based on their susceptibility with high levels of precision. Here 1 denotes that all (100%) the data points belonging to low and high susceptibility were classified correctly, and 0.96 denotes that 96% of the data points were correctly classified under medium susceptibility along with 0.36 denoting 3.6% being misclassified under low susceptibility.

The performance metrics such as accuracy, precision, sensitivity, and F-scores are calculated for different machine learning models, and it is compared with the RF model as shown in Fig. 7. It is found that the random forest approach has an overall classification accuracy of 99.7% for the validation dataset of the healthcare dataset. The proposed random forest classifier gives a precision of 99.8%, sensitivity of 98.8%, and F-score of 99.29%.

The higher sensitivity of random forest model shows that a high proportion of actual positives are classified correctly. Also the higher specificity highlights that a good percentage of the "safe" populations is identified as not susceptible to the infection. The precision indicates that there may be some cases of individuals who may not have a high susceptibility, but they may be declared as a considerably risky individual.

Discussions

In order to quantitatively analyze the proposed RF model, train, and test accuracies over different sizes of samples, scalability of the model and performance of the model are presented and the results are analyzed.

Effect of size of the dataset

In order to study the aspects of over fitting of the proposed random forest model, it is applied to different numbers of training and testing datasets. This provides an indication of the over fitting of the model and the ability of generalization of classification for the given dataset. Figure 8 shows convergence of test and train

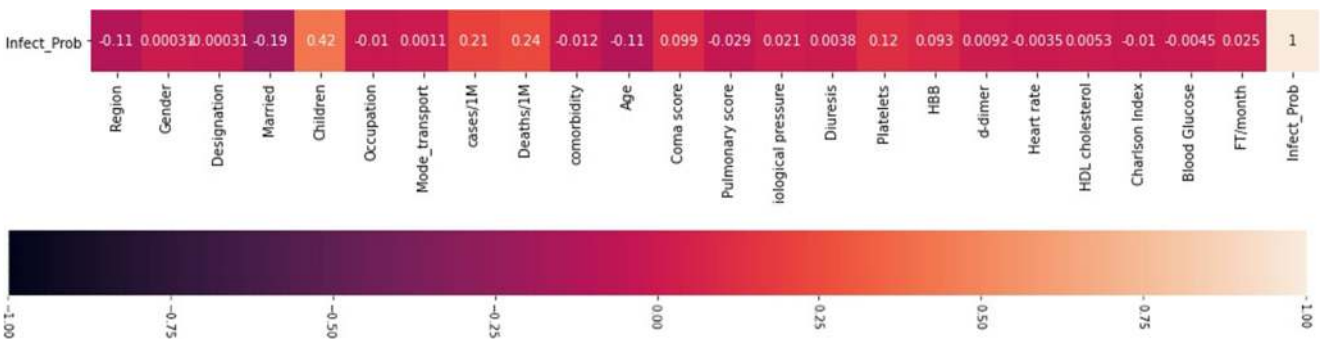

Fig. 5 Heat map of healthcare data

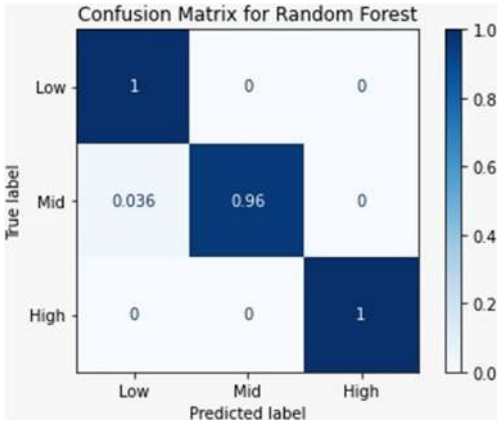

Fig. 6 Confusion matrix for random forest model for the healthcare data

accuracies, and it highlights that the model is not over fitting.

The model can produce highly accurate predictions even with a lower number of training examples as the accuracy plot starts at over 99% even for a smaller batch of dataset. This shows the agility of the model to adapt to different dataset sizes without over fitting.

Scalability curves

Further, a study has been carried out to understand the scalability of the proposed random forest model and the effect of number of training examples of healthcare dataset on the accuracy of classification and training time. The results are shown in Fig. 8 and Fig. 9.

It shows a linear trend for the training time when the size of the dataset is changed, and it indicates the

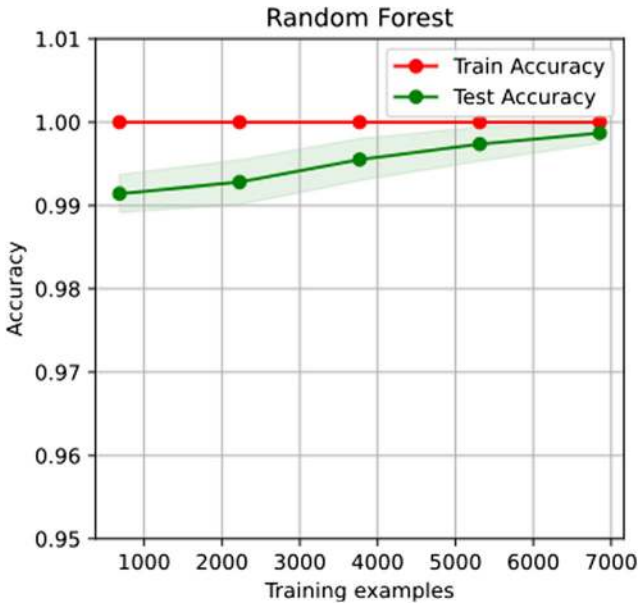

Fig. 8 Graph of accuracy of the random forest model versus the number of training examples

robustness of the model with a higher number of data points.

Training time

Figure 10 depicts the effect of different numbers of training examples on the training time, and it shows the linear trend in accuracy of classification when the size of the dataset is changed.

It can be seen that there is a smooth increase in accuracy which highlights the robustness of the model with a higher number of training examples. These results indicate that the

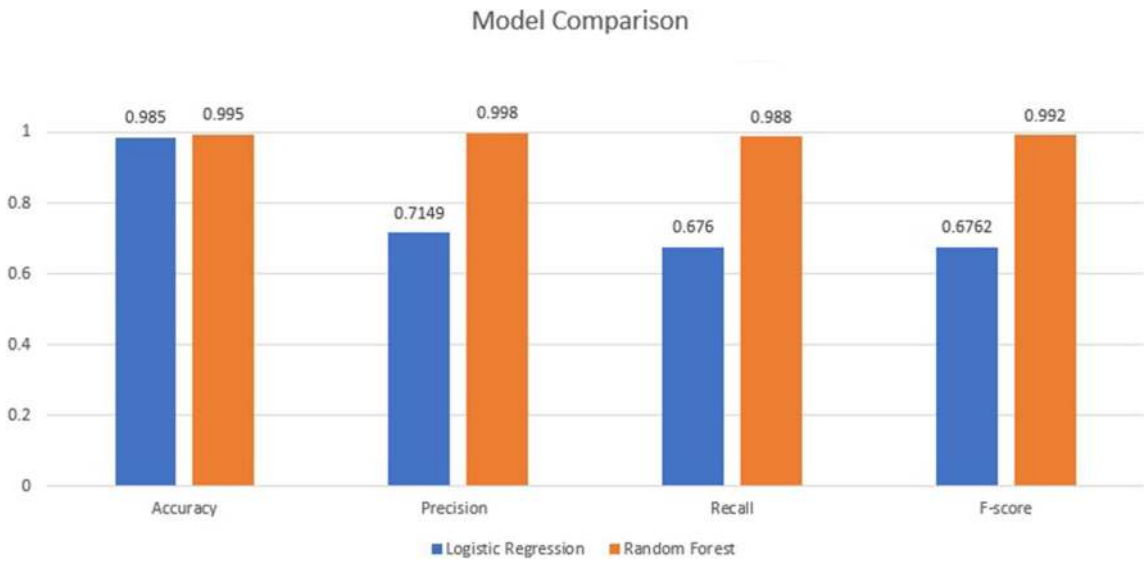

Fig. 7 Comparison of performance measure of random forest classifier

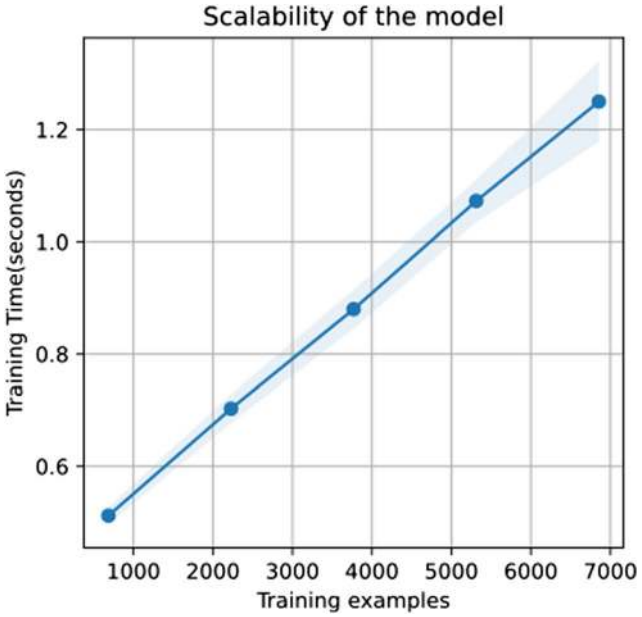

Fig. 9 Graph of scalability of the random forest model versus the number of training examples

proposed random forest model performs well with high accuracy despite smaller dataset sizes and requires shorter training time despite large dataset sizes.

Regression performance of support vector regression

Support vector regression is applied to the healthcare dataset for predicting the infection susceptibility probability score of COVIG-19. From Fig. 11a, it can be seen that there is a lesser deviation between the

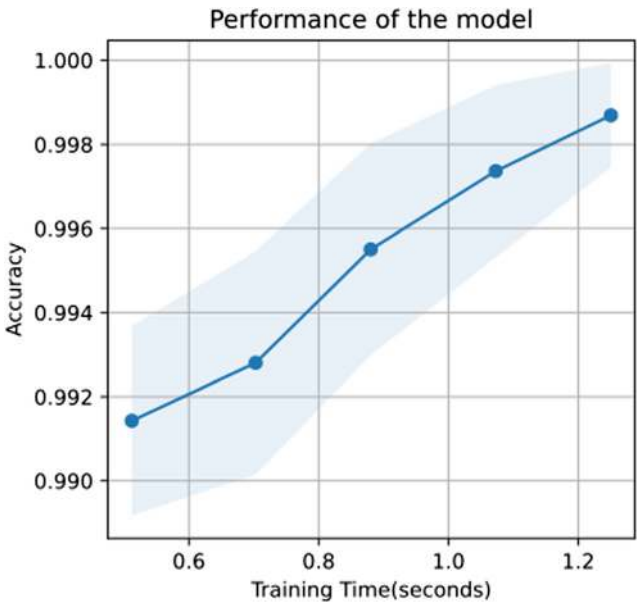

Fig. 10 Effect of training time on accuracy of prediction

predicted values and actual values using the RBF kernel function in support vector regression.

Figure 11 b shows the function fit by the linear kernel of SVR displaying more deviation from the actual values of infection susceptibility probability (ISPs). These results highlight the improved prediction of support vector regression using RBF kernel as compared to the linear kernel.

Conclusions

This paper presented a machine learning based prognostic model for the categorical classification of COVID-19 infection susceptibility as low, medium, and high based on the healthcare data of an individual. A medical dataset available in the online repository is used for training, testing, and validation of the proposed approach. A mobile application with user interface is developed for collecting relevant healthcare data and categorizing the risk of infections as low, medium, and high using random forest classifier ML algorithm. Correlation analysis of healthcare data is carried out to study the influencing factors on influencing the infection susceptibility probability (ISP) score using the heat map and correlation. Demographic fields are found to be critical in influencing the COVID-19 infection such as "No of children"(0.42), "Cases per Million" (0.21), "Deaths per Million" (0.24), and "Platelet Count" (0.12). Based on the comparative study of performance measures for the classification of infection score, it is found that the random forest classifier has the better overall classification accuracy of 99.7%, precision of 99.8%, sensitivity of 98.8%, and F-score of 99.29% as compared to the other machine learning algorithms such as logistic regression, support vector regression, and neural network. Studies on proposed random forest model for different number of datasets highlight robustness and scalability of the proposed approach, and it highlights the high accuracy despite smaller dataset sizes and a shorter training time despite large dataset sizes. It is found that support vector regression using RBF kernel function is found to be superior to the linear kernel function in prediction of infection susceptibility of individual for COVID-19.

These results highlighted the application of machine learning approaches for analyzing the healthcare data in understanding the infection severity of individual for COVID-19. From the larger public healthcare perspective, proposed approach will be helpful in identification of individuals who are highly susceptible for the COVID-19 infection in a containment zone which can give a decisive role to physicians and government officials for

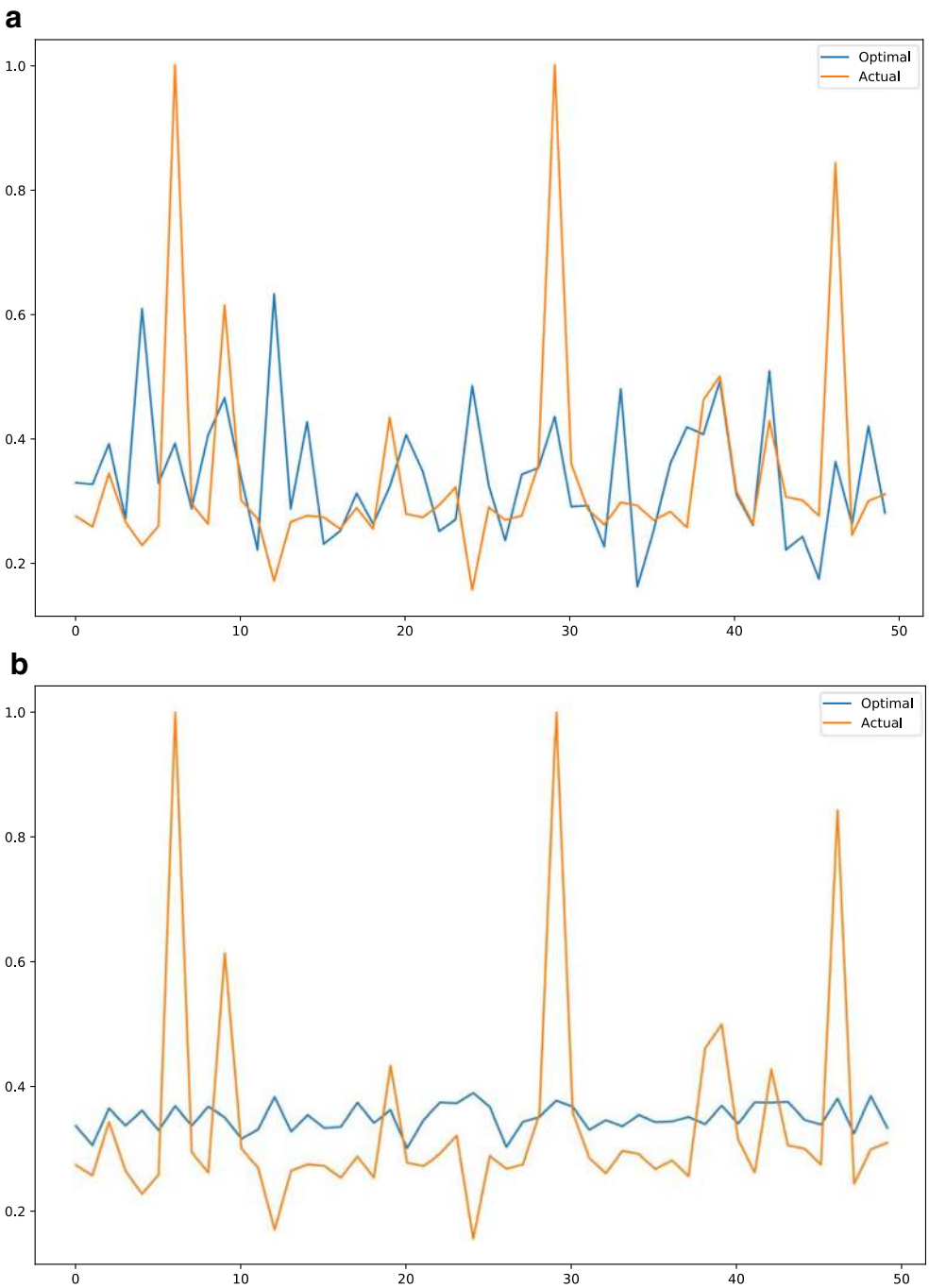

Fig. 11 Comparison of kernel functions used in support vector regression. (**a**) Optimal vs actual result plot of SVR using RBF kernel. (**b**) Optimal vs actual result plot of SVR using linear kernel

planning the more aggressive treatment and a better chance of survival. Also the early detection can also help hospitals prioritize intensive care resources.

Acknowledgments The authors thank the management of Vellore Institute of Technology, Vellore, for providing the necessary facilities to carry out this research work.

Data availability The datasets presented in this study can be found online in open source repositories.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interests.

Code availability The opensource reproducible code is available at the code repository: https://github.com/srivatsanrr/autonom_covid. The following open source libraries are used for the implementation of our work: Keras, https://keras.io; Sklearn, https://scikit-learn.org/stable/; and statsmodels, https://www.statsmodels.org/stable/index.html.

References

Al-Karawi D. Machine learning analysis of chest CT scan images as a complementary digital test of coronavirus (COVID-19) patients. medRxiv. 2020.

Carrillo-Larco RM, Castillo-Cara M. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: an unsupervised machine learning approach. Wellcome Open Res. 2020;5(56):56.

Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access. 2017;5:8869–79. https://doi.org/10.1109/ACCESS.2017.2694446.

Hui DS, Azhar EI, Memish ZA, Zumla A. Human coronavirus infections—severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and SARS-CoV-2, vol. 2. 2nd ed: Elsevier Inc.; 2020. https://doi.org/10.1016/b978-0-12-801238-3.11634-4.

Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, et al. Propagation analysis and prediction of the COVID-19. Infect Dis Model. 2020;5:282–92. https://doi.org/10.1016/j.idm.2020.03.002.

Liu D. A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. arXiv preprint arXiv:2004.04019. 2020.

Mei X, Lee H, Diao K. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat Med. 2020;26:1224–8. https://doi.org/10.1038/s41591-020-0931-3.

Metsky HC. CRISPR-based COVID-19 surveillance using a genomically-comprehensive machine learning approach. bioRxiv. 2020.

Narin, Ali, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv: 2003.10849. 2020.

Pandey G. SEIR and regression model based COVID-19 outbreak predictions in India. arXiv preprint arXiv:2004.00958. 2020.

Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ 2016. 2015;353:i3235. https://doi.org/10.1136/bmj.i3235.

Qayyum A, Qadir J, Bilal M, Al Fuqaha A. Secure and robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng. 2020:1. https://doi.org/10.1109/rbme.2020.3013489.

Repici A, Maselli R, Colombo M, Gabbiadini R, Spadaccini M, Anderloni A, et al. Coronavirus (COVID-19) outbreak: what the department of endoscopy should know. Gastrointest Endosc. 2020;92:1–6. https://doi.org/10.1016/j.gie.2020.03.019.

Singh S. Flipr Hiring Challenge, 1. 2020. Retrieved May 2020 from https://www.kaggle.com/srijansingh53/flipr-hiring-challenge/version/1. Accessed 30 May 2020.

Tang Z. Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. arXiv preprint arXiv:2003.11988. 2020.

Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ. 2020;369. https://doi.org/10.1136/bmj.m1328.

Yan L. Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan. medRxiv. 2020.

Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell. 2020;2:283–8. https://doi.org/10.1038/s42256-020-0180-7.

Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis. 2020a;12:165–74. https://doi.org/10.21037/jtd.2020.02.64.

Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis. 2020b;12:165–74. https://doi.org/10.21037/jtd.2020.02.64.

Affiliations

R. Srivatsan [1] · Prithviraj N. Indi [1,2] · Swapnil Agrahari [3] · Siddharth Menon [1] · S. Denis Ashok [4]

[1] School of Electronics Engineering, VIT University, Vellore, India

[2] School of Electrical Engineering, VIT University, Vellore, India

[3] School of Mechanical Engineering, VIT University, Vellore, India

[4] Department of Design and Automation, School of Mechanical Engineering, VIT University, Vellore, Tamil Nadu 632014, India