Computer Modeling in Engineering & Sciences
DOI:10.32604/cmes.2020.012467
Article

Tech Science Press

Real-Time Analysis of COVID-19 Pandemic on Most Populated Countries Worldwide

Meenu Gupta[1], Rachna Jain[2], Akash Gupta[2,*] and Kunal Jain[2]

[1]Department of Computer Science and Engineering, Chandigarh University, Punjab, 140301, India
[2]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, 110063, India
*Corresponding Author: Akash Gupta. Email: akashgupta752000@gmail.com
Received: 01 July 2020; Accepted: 14 September 2020

Abstract: The spread of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has already taken on pandemic extents, influencing even more than 200 nations in a couple of months. Although, regulation measures in China have decreased new cases by over 98%, this decrease is not the situation everywhere, and most of the countries still have been affected by it. The objective of this research work is to make a comparative analysis of the top 5 most populated countries namely United States, India, China, Pakistan and Indonesia, from 1st January 2020 to 31st July 2020. This research work also targets to predict an increase in the number of deaths and total infected cases in these five countries. In our research, the performance of the proposed framework is determined by using three Machine Learning (ML) regression algorithms namely Linear Regression (LR), Support Vector Regression (SVR), and Random Forest (RF) Regression. The proposed model is also validated upon the infected and death cases of further dates. The performance of these three algorithms is compared using the Root Mean Square Error (RMSE) metrics. Random Forest algorithm shows best performance as compared to other proposed algorithms, with the lowest RMSE value in the prediction of total infected and total deaths cases for all the top five most populated countries.

Keywords: COVID-19; SARS COV-2; country-wise analysis; most populated countries; pandemic; mortality rate

1 Introduction

Human wealth plays an essential role in the growth of every country, and every country makes some policy for the security of their people's health and wealth. But, medical-related health issues are the more severe concern of every country where people lost their lives in huge amount, and reasons on these types of health issues are their food culture. Nature gives us fruits, vegetables and food which help people to live happily. But it is a bitter truth that peoples are in habit to eat living (i.e., seafood) things for their survival. Because of these types of food habit, they suffer from different kinds of diseases such as flu, pneumonia etc. As disease came into the picture,

our scientist (of every country) starts preparing the vaccine for that disease but every time it's not happened.

The same case arises in the present time where every country suffers from one virus disease that is called COVID-19 (or novel coronavirus). Even they suffered from this kind of disease form past many years, and our scientist was cable of preparing the vaccine for those types of disease. But, in the present time, they are helpless for providing vaccines for this disease, and every day thousands of peoples lost their life worldwide. All the global pandemics in the history of the world are discussed in Fig. 1.

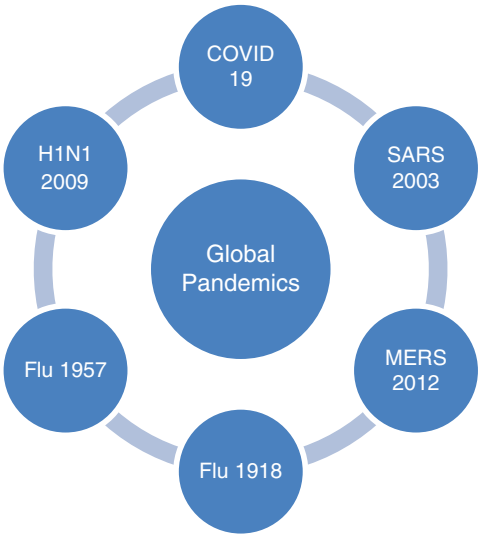

Figure 1: Global history of pandemics

COVID-19 is a deathly disease originated by the virus SARS-CoV-2 [1], and it spreads from human interaction. The Wuhan city of China witnessed the first case of coronavirus at the end of November 2019 and has since spread globally [2,3]. As of the 31st July 2020, around 17.3 million cases have been reported worldwide so far. At present, around 213 countries and territories suffer from this deathly disease, resulting in more than 673,279 deaths [4]. The COVID-19 outbreak was declared a worldwide pandemic on 11th March, 2020 by World Health Organization (WHO) [5]. Initially, this virus spreads among people through interactions and close contacts i.e., by the transmission of little droplets produced by sneezing, talking and coughing [6]. Fever, dry cough and tiredness are the major symptoms of this virus [7]. The symptoms can be felt typically around five days after coming into contact with the virus. However, it may take 10–14 days based on the immune system of the person [8,9]. According to WHO, there are no available vaccines to deal with this deadly disease [10].

Here is the comparison of COVID-19 with other pandemics and viruses. Tab. 1 shows the comparison based on R nought (or R0) value and average mortality rate.

R0 value shows how contagious an infectious disease is; it is a numerical equation that shows what number of individuals will get influenced by each contaminated individual.

Table 1: Comparison of global pandemics and viruses based on R0 value and average mortality rate [11]

Kinds of virus disease	R0 value	Average mortality rate
COVID-19	5.7	3.64%
2009 H1N1	1.5	0.02%
1957 flu	1.7	0.60%
1918 flu	1.8	2.50%
2003 SARS	3.5	10%
MERS	0.5	35%

- If R0 value is less than 1, then each infectious person can infect at most 1 person. In this case, the disease will eventually die out.
- If R0 value is equal to 1, then each infectious person can transmit the disease to 1 other person. In this case, the disease will stay for a long time but won't converted into an outbreak.
- If R0 value is greater than 1, then each infectious person can infect more than 1 person. In this case, the disease becomes an outbreak.

R0 value is calculated as shown in Eq. (1).

$$R0 = \frac{\beta}{\gamma} \tag{1}$$

where, $\beta =$ Transmission Rate or Contact Rate, $\gamma = 1/$average infectious period.

The mortality rate is determined as the total number of deaths divided by the total number of infected patients. Both the R0 and mortality rate varies from country to country, and both depend on many factors including geography, quality of health care, lifestyle, age of population etc. This paper seeks to carry out the comparative analysis about the prediction of an increase in the deaths and total cases in five most populated countries near future. This work uses three Machine learning regression algorithms; LR, SVR, and RF. The data is taken from ourworldindata.org website. Root Mean Square Error (RMSE) metrics is used for comparing the performance of the proposed models.

The rest of the paper is sorted out as follows: Section 2 gives the detailed review related to the various researches taking place in the world related to COVID-19 as well as work is done in healthcare industry using ML models. Section 3 discusses the methodology, which includes data processing and the different algorithms used. Section 4 reports the results, simulations and comparative analysis. Section 5 discusses the limitations of the proposed model and threats to validity of the results. Finally, Sections 6 concludes the comparative study.

2 Literature Review

A COVID-19 is a virus disease which frequently spreads through human interaction and in just a couple of months almost every country gets affected by this. Scientists from every country are engaged day and night in preparing the vaccine to induce coronavirus, but not a single country got success yet. One thing must be understood that the disease does not differentiate between the poor and rich persons because it does not follow any religion. The daily routine of peoples and the economy of a country also affected due to this novel coronavirus. This (viruses

like COVID-19) is the reason where many countries having the best medical facilities or despite being developed countries suffer a lot to handle the situation arises due to COVID-19. Now here are some researches that take place in the field of medicine by using various machine learning algorithms and various researches taking place in the world related to coronavirus.

From years, there has been an expanding enthusiasm in the development of regression-based techniques for organ transplantation in healthcare industry. Since the tissues and organs are arranged in a particular manner in the human body, the human body not tends out of the ordinary voxels, because of their contextual data, can anticipate the encompassing life systems. For example, if the area around the voxel shows the presence of an ordinary heart tissue, other than the situation of the heart, it can provide a gauge of position of the nearby lungs.

Zhou et al. [12] presented a methodology dependent upon boosting ridge regression technique to identify and limit the left ventricle in cardiovascular ultrasound two-dimensional pictures. There, the proposed function tries to predicts the orientation, scale and relative position of the LV based upon Haar-like features processed on the two-dimensional pictures. Remarkable outcomes are exhibited on echocardiogram arrangements. Zheng et al. [13] presented a methodology called negligible space learning in order to recognize and restrict the heart chambers in 3D cardiovascular CT scan. To separate the intricacy of adapting legitimately in the full 3-dimensional closeness change space, the researchers show that readiness of classifier on projections of main space adequately diminishes the search space. Utilizing this thought, they manufacture a course of classifiers subject to the Probabilistic Boosting Tree to first anticipate the position, then position-orientation and finally entire three-dimensional pose. Researchers derive the thought further to flexible minor space getting the hang of utilizing factual shape models [14]. Even though these methodologies have demonstrated excellent execution on CT scans, but developing this type of classifiers is a very advanced learning strategy and it needs huge dataset for training.

Criminisi et al. [15] applied random forest regression algorithm in their research to study about the limitations of organs in three dimensional CT scans. The authors demonstrated that their strategy accomplishes excellent presentation over map book enrolment and this while profiting of quick preparing and testing. The authors could rely upon supreme radio density esteems by CT, here, we oversee Magnetic Resonance (MR) pictures which give simply relative characteristics and experience the ill effects of field inhomogeneities. To handle this difficult issue, the author adopts the regression forest framework by presenting 3D LBP descriptors. Also, the author executes an arbitrary ferns regression approach and compares it with forest regression. Both regression methods are compared and accessed with a map book-based registration approach. Chloroquine phosphate (old medicine commonly used in malaria treatment) is considered to give evident viability. The medicine is considered to be associated with the accompanying type of guidelines for the diagnosis, treatment and prevention of pneumonia caused by COVID-19 provided by the NHC of the People's Republic of China [16,17].

Research is done on the comparison of the reproduction number (R0) of COVID-19 and SARS coronavirus. Shan et al. discovered the average R0 as around 3.28 and R0 estimates for SARS in the range of 2 and 5 (i.e., same as a mean range of COVID-19). This is expected because of the similarities in both pathogen and area of introduction. In next, the COVID-19 is as of now broader than SARS, demonstrating it might be progressively transmissible [18,19]. SARS-CoV-2 contaminates host cells via ACE2 receptors, prompting coronavirus disease, while additionally making intense damage or injury to cardiovascular network that may resulted in cardiac arrest [20].

The research has done on people of China in order to measure the various factors associated and psychological reactions to Corona virus during its starting phase [21]. The conclusion that can be drawn from this study is that during the beginning period of COVID-19, over half of the occupants evaluated their mental effect as moderate-to-serious, and about 33% announced moderate-to-extreme nervousness [22], Especially, Females and understudies experienced more significant levels of pressure, tension, and sadness [23]. Around one-fourth individuals in the United Kingdom are assigned high risk by a coronavirus that mainly includes all grown-ups matured more than 70 or have some underlying wellbeing conditions like respiratory and cardiovascular ailment [24]. Strict limitations are presently set up for everybody, with the exception for essential labourers. These standard actions will be set up for a considerable amount of time [25]. These harsh but necessary limitations will possibly result in loss of physical and mental functions as well as loneliness among old aged people.

A lot of researches have been done across the world related to COVID-19. Tab. 2 provides the summary of some of those researches, by incorporating various methods or algorithms used, along with the results obtained from the researches.

Table 2: Related work on COVID-19

Author	Title	Methods used	Outcome
Poon et al. [26]	Early diagnosis of SARS Coronavirus infection by real-time RT-PCR	• RNA extraction and reverse transcription. • Conventional PCR technique. • Real-time PCR technique.	By increasing the RNA sample for RNA extraction and optimizing the RT-PCR, the detection of the Corona Virus ca be greatly enhanced.
Shan et al. [27]	Lung infection quantification of COVID-19 in CT images with deep learning	• DL based 'VB Net' Neural networks • Human in the loop strategy	The quantitative evaluation showed high accuracy for automatic infection region delineation, POI metrics
Davies et al. [28]	Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK	Stochastic age-structured transmission model that includes school closure, self-isolation, social distancing and shielding of older people.	Extreme measures are required to bring the epidemic under control
Ceylan [29]	Estimation of a COVID-19 prevalence in Italy, Spain, and France	• ARIMA (0,2,1) • ARIMA (1,2,0)	ARIMA (0,2,1) comes out as the best model for France and Italy. and ARIMA (1,2,0) for Spain.
Yan et al. [30]	Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan.	XGBoost Classifier machine learning algorithm.	The three indices-based prognostic prediction model can present a clinical route to the recognition of critical cases from severe cases.
Ardabili et al. [31]	COVID-19 Outbreak Prediction with Machine Learning	• Multi-Layer Perceptron (MLP) • An Adaptive Neuro-Fuzzy Inference System (ANFIS)	The results of two ML models reported a high generalization ability for long-term prediction.
Shoeibi et al. [32]	Automated Detection and Forecasting of COVID-19 using Deep Learning Techniques	• AlexNet • VGGNet • GoogleNet • DenseNet • XceptionNet • SqueezeNet • GAN etc.	With more public databases, better DL models can be developed to detect and predict the COVID-19 accurately.

After seeing different researchers view, the work is focused on the analysis of the total number of death and infected case happened due to COVID-19 around the top five populated countries worldwide. The main focus of this research is to the point that precaution always better than cure and how we can secure our self from a hidden enemy (i.e., COVID-19). Different classification algorithms have been applied for the analysis. This work considers the case of top five populated countries and focuses on how precaution saves human life and support country growth in an adamant time where every country suffers a lot. This work is in consideration of people health and to make them aware of how precaution can save their life where the enemy is in front of you but hidden.

3 Methodology

For analyzing the impact of COVID-19 on top 5 most populated countries (i.e., United States, India, China, Pakistan and Indonesia), we collect the data set from the web and apply the different classification algorithms namely LR, SVR and RF, which are discussed below.

3.1 Data Used

The data for this research is incorporated from ourworldindata.org website [33]. Two CSV files containing the total infected and total death cases worldwide are used in this research. The dataset is pre-processed to limit our research for the analysis of the five most populated countries throughout the world by eliminating the data of the other countries. The NaN values are dealt with by the Imputer class of imputer module of sklearn library. The final datasets provide the total infected and the total death cases due to the COVID-19 in the top five most populated countries namely China, India, United States, Indonesia and Pakistan, from 1st of January to 31st of July.

3.2 Techniques and Algorithms Used

In this work, three regression models, namely LR, SVR, and the RF Regression Model, are used for the analysis of COVID-19 impacts on the top 5 most populated countries. By using these algorithms, we are trying to predict the behaviour of the rise in the infected and death cases in the top five most populated countries on future dates as well. The motivation behind the selection of these three algorithms for this research work is discussed below.

- The main motivation behind the selection of Linear Regression is its simplest estimation process and it is also observed that many countries had recorded a linear growth in COVID-19 cases.
- The main motivation behind the selection of Support Vector Regression is that its computational abilities are independent of the input dimensions. Hence, proves to be very beneficial.
- The main motivation behind the selection of Random Forest Regression is its ability to randomize the data and its focus upon the general pattern among the data values rather than any particular linear or exponential curve, which is the actual requirement in this pandemic since there are a lot of countries which are showing different patterns in the rising of COVID-19 cases other than linear or exponential.

3.2.1 Linear Regression (LR)

LR algorithm of ML is based on the concept of supervised learning. The algorithm is used to perform the regression. Regression models focus on the prediction of a target vector based on different independent feature vectors. It is commonly utilised for determining the association

among the variables and estimating factor. Various regression models vary from one other based upon the sort of connection between the dependent and independent vectors and the number of independent factors being considered.

Steps to Building an LR Model:

A. Write a hypothesis function for Linear Regression by calculating the values of θ1 and θ2, as shown in Eq. (2) [34].

$$y_i = \theta_1 + \theta_2 * x_i + \varepsilon_i \tag{2}$$

where, x: Input data for training, y: Labels to data, θ_1: Intercept, θ_2: x coefficient, ε: error term in ith observation.

B. Calculate the Cost Function (J) for the model. To achieve the best-fitted regression line, the linear regression model targets to predict the y values in such a manner that the error difference between the predicted and the true one is minimum. The cost function is calculated by using Eqs. (3) and (4) [35].

$$Minimise \frac{1}{n}\sum_{i=1}^{n}(pred\,(i) - y(i))^2 \tag{3}$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred\,(i) - y(i))^2 \tag{4}$$

where, J represents the Cost Function, $pred(i)$ denotes the predicted value at ith iteration, $y(i)$ represents an actual value and n is the total number of iterations.

The θ1 and θ2 values are then updated to decrease the value of the Cost function and determining the best-fitted line, this linear regression model uses the concept of Gradient Descent. The basic idea is, to begin with, the arbitrary estimations of θ1 and θ2 and afterwards recursively updating its values, until least cost is calculated.

Pseudocode:

1. Begin
2. For i = 1 to n:
   Read x_i and y_i
3. Initialize hypothesis function
4. Calculate θ1 and θ2
5. Calculate J
6. While J is not minimum:
   Update θ1 and θ2
   Calculate J
7. End

3.2.2 Support Vector Regression (SVR)

SVR algorithm supports both linear as well as non-linear regression. As it seems in Fig. 2, the goal is to fit as many instances as possible between the lines while constraining the margin violations. The violation idea in this model is shown as epsilon (ε) [36].

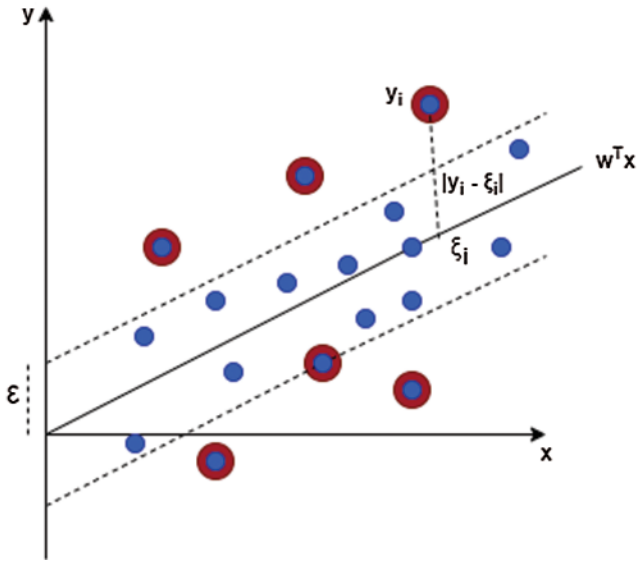

Figure 2: Support vector machine regression [37]

Steps to Building SVR Model:

A. Collect the training set {x, y}.

B. Select a kernel and parameter and regularization if needed.

C. Construct the correlation matrix $K_{i,j}$, as shown in Eq. (5).

$$K_{i,j} = \exp\left(\Sigma_k \theta_k \left|x_k^i - x_k^j\right|^2\right) + \in \delta_{i,j} \tag{5}$$

D. Train your model, to get the contraction coefficient ($\vec{\alpha}$) by utilizing the main part of the algorithm.

$$\vec{K}\vec{\alpha} = \vec{y} \tag{6}$$

where, $\vec{K}$ = Correlation Matrix, $\vec{\alpha}$ = Unknown Values Set, $\vec{y}$ = Vector corresponding to a training set

E. Now, use the coefficient from Eq. (6) and correlation vector ($\vec{C}$) from Eq. (7) to calculate the value of estimator (y^*).

$$\vec{C}_i = \exp\left(\Sigma_k \theta_k \left|x_k^i - x_k^*\right|^2\right. \tag{7}$$

$$y^* = \vec{\alpha}.\vec{C} \tag{8}$$

where, y^* = estimator value, $\vec{C}$ = Correlation Vector

It differs from Linear Regression in the sense that the target in linear regression is to reduce the error between the estimated and the actual data provided while in SVR, the target is to make sure that the errors will not exceed the threshold [38].

Pseudocode:

1. Begin
2. For i = 1 to n:
   Read x_i and y_i
3. Calculate Correlation Matrix values ($K_{i,j}$)
4. Calculate $\vec{\alpha}$
5. Calculate $\vec{C}$
6. Calculate estimator (y^*)
7. If error < estimator:
   Go to Step 8
   Else:
   Go to Step 4
8. End

3.2.3 *Random Forest (RF) Regression*

RF techniques are capable of handling both the regression and classification tasks by the utilisation of various decision trees and a procedure known as Bootstrap Aggregation, generally known as bagging [39]. The fundamental idea behind this algorithm is to combine the various decision trees to decide the final output rather than depending upon individual decision trees [40]. A simple model of the random forest algorithm is represented in Fig. 3.

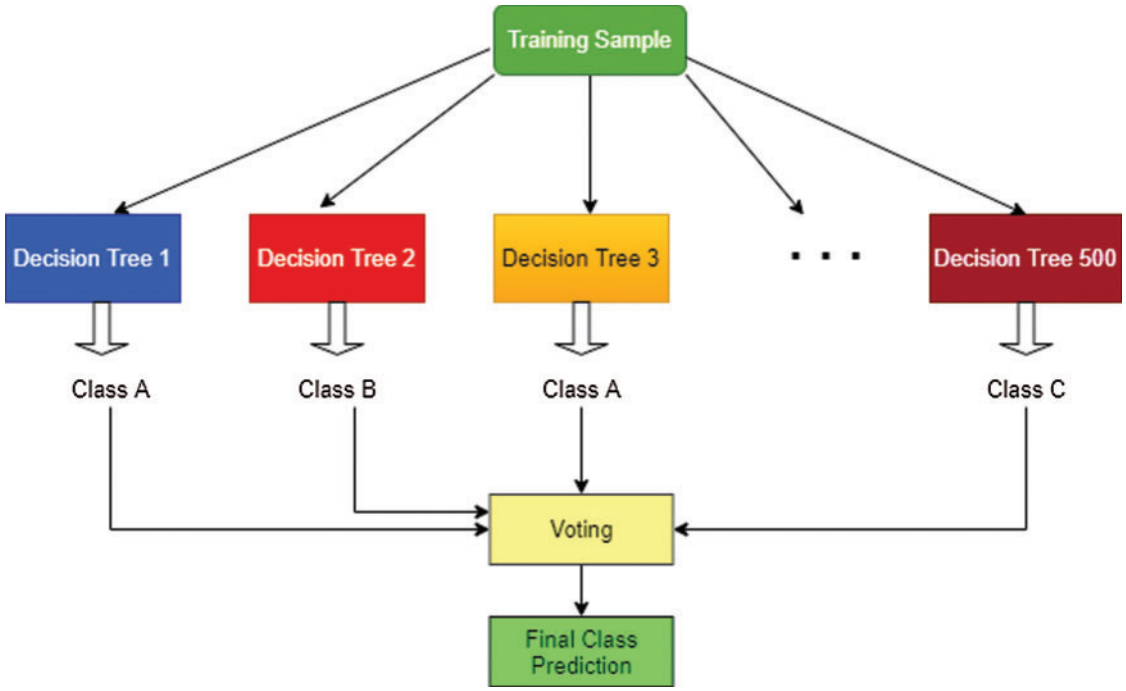

Figure 3: Random forest regression [41]

This model is one of the additive models which calculates its estimations by joining results from a grouping of base models. All the more officially, we can compose this class of models, as shown in Eq. (9).

$$h(x) = f_0(x) + f_1(x) + f_2(x) + f_3(x) + \cdots + f_n(x) \tag{9}$$

where, this final model h is defined as the summation of these base models fi where i is ranging from 0 to n. Every base classifier is nothing but a simple decision tree which is known as the model ensemble [42]. In the RF algorithm, all the base models are built separately utilizing another subsample of the same information.

3.3 Proposed Methodology

In this proposed research model, initially, the data related to total infected cases and deaths of the top five most populated countries is provided as the input to this model. In which 2/3rd of the data is used for the training purposes and 1/3rd is used for testing it. The model comprises of three algorithms namely Linear Regression, Support Vector Regression and Random forest Regression as shown in Fig. 4 which helps to make a prediction of the infected and death cases on further dates as well, assuming the circumstances remains same.

4 Experimental Results and Analysis

The entire world faces a widespread of the coronavirus (COVID-19) disease. Every day there is a large increase in the total number of infections and deaths caused by this disease. In this work, we considered five countries based on their population. This work gives analyse about the COVID-19 outbreaks in terms of total death cases and infected cases in these countries. Different classification algorithms are applied in this work for further analyses which is further discussed in subsections with different metrics used.

4.1 Evaluation Metrics

In this research work, since we have to deal with the larger error values, the Root Mean Squared Error (RMSE) metrics proves to be very beneficial metric as compared to other regression metrics like MSE and MAE. RMSE metrics is discussed in detail below.

4.1.1 Root Mean Square Error (RMSE)

RMSE is characterised as the standard deviation of the errors [43]. It commonly occurs when estimation is required to be performed on the dataset [44]. It is similar to MSE (Mean Squared Error) with the only difference that the root of the value is considered for determining the model's accuracy. It is calculated using Eq. (10).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Actual - Predicted)^2} \tag{10}$$

where n represents the frequency or total observations considered, the sigma symbol represents the difference between the real and the estimated values taken on each ith value ranging from 1 to n. It is implemented using mean_squared_error method of sklearn library. It is used to calculate the root mean square values which are discussed as under.

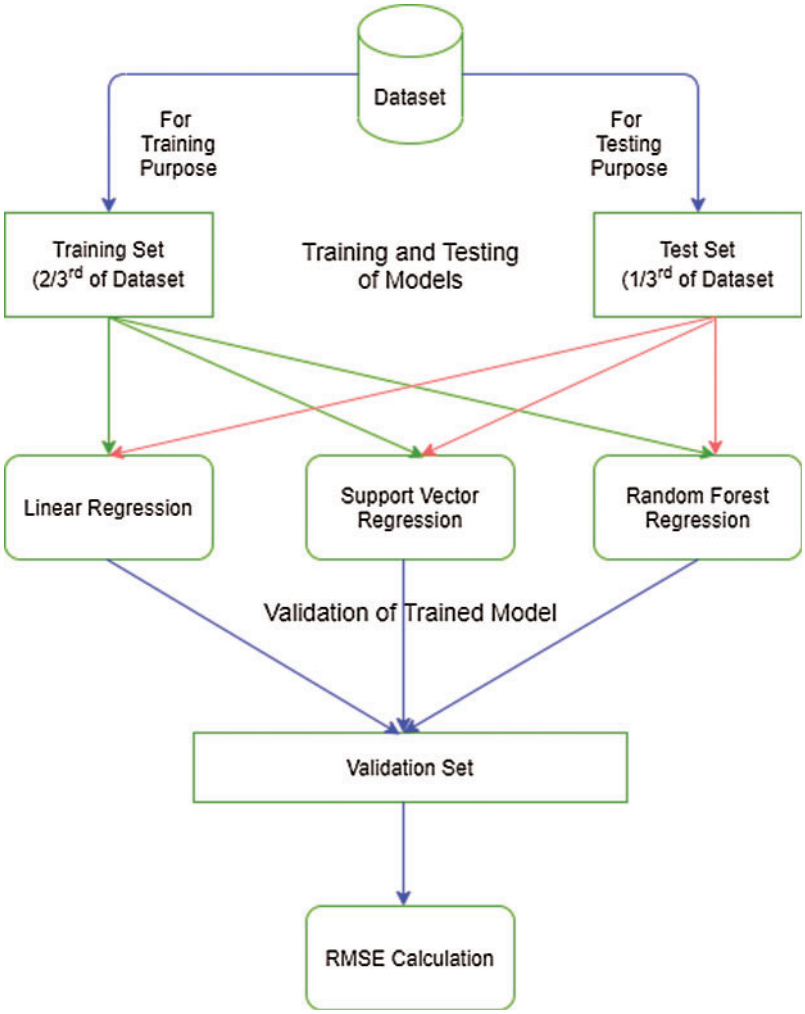

Figure 4: A proposed model using various regression algorithms

4.2 Country Wise Analysis

The population-wise analysis of deaths and infected cases due to COVID-19 is discussed below. Here, the analysis is performed on the total deaths and infected cases of the top five populated countries from 1st January 2020 to 31st July 2020. The x-axis represents the number of days counted from 1st January onwards. Hence, 1 represents the 1st of January, 2 represents the 2nd of January and so on, till 31st August.

4.2.1 United States

As can be analysed from the graph shown in Fig. 5, there were around negligible cases in the initial two months of this pandemic. After that, the curve takes an exponential growth and turns into devastation. The curve shows that there are approximate 4.5 million cases were recorded in the United States till 31st July 2020.

Fig. 6 shows that this pandemic leads to approximate 152,000 deaths in the USA in around 200 days.

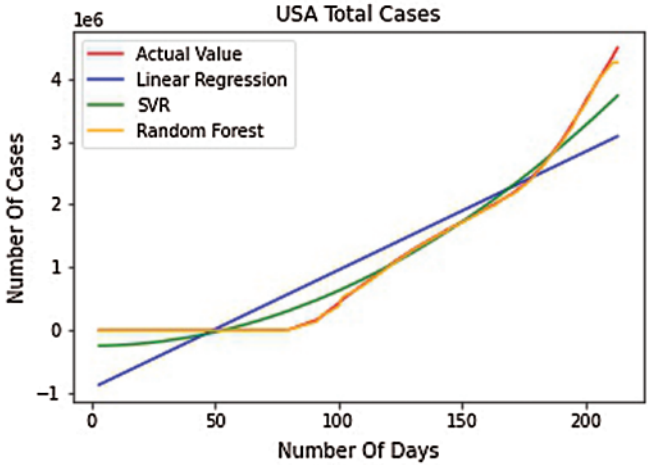

Figure 5: Total infected cases in the USA

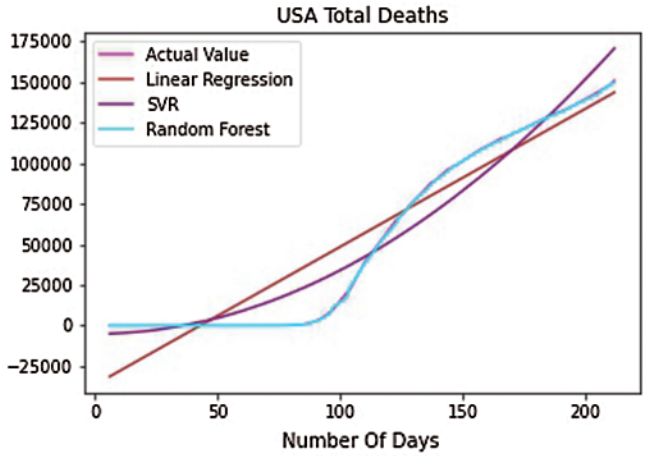

Figure 6: Total death cases in the USA

The US is the 3rd most populated country with a population of 33 million people and have the highest number of cases in the world [45]. The primary reason for having such a high number of cases is the delay showed off in the testing. Also, the United States government initially refused to relax the regulations, which does not allow the health departments and the states to develop their testing kits based on WHO guidelines. All the COVID-19 infected samples were being sent to the headquarters of the CDC (Centers for Disease Control and Prevention) in Atlanta. After that, the delay increases because of the faulty test kit sent by CTC. These primary factors are responsible for the considerable increase in COVID-19 cases in the United States.

Figs. 5 and 6 represents the prediction of infected and death cases in the US by Linear Regression, SVR and Random Forest and it can be concluded from graphs that the Random Forest shows the best performance among them.

4.2.2 India

Fig. 7 gives an analysis that like the US, the spread of the coronavirus among India starts around March 2020 and increase gradually till the end of May, but after that, it takes an exponential rise in cases, and there are approximate 1.6 million cases recorded in India till 31st of July.

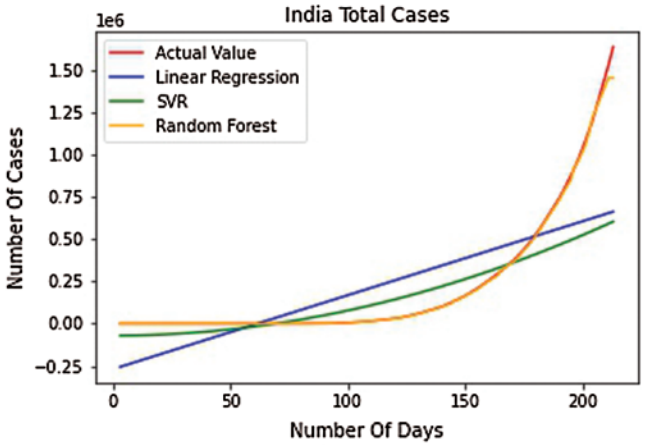

Figure 7: Total infected cases in India

Fig. 8 shows that the COVID-19 takes approximate 35,000 lives till the 31st of July in India.

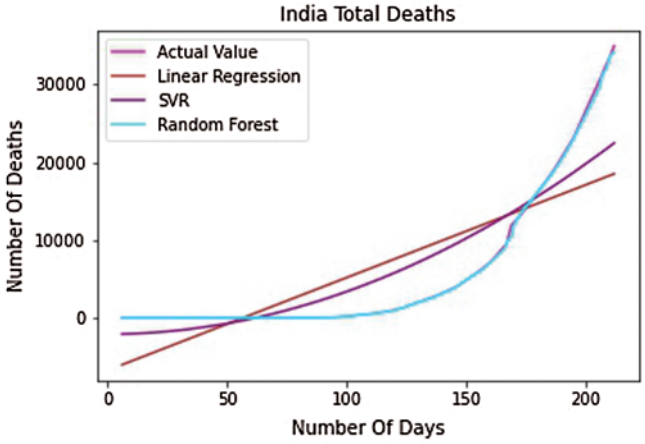

Figure 8: Total death cases in India

India is the 2nd most populated country with a population of 1.35 billion people and rank 3rd at risk of importing Coronavirus cases [46]. In India, this disease was initially identified in a Kerala student who returned from Wuhan on January 30, 2020. But it starts spreading among a large number of people from mid-march. Till now, four phases of lockdown had been employed in India. Lockdown 1.0 and Lockdown 2.0 are somewhat successful, and most of the Indian

states showed a positive response by reducing the number of infected cases in those states. But after April 30, when the Lockdown 3.0 starts, most of the states in India shows exponential growth in the infected cases. Since, because of different population density, culture, and diversity of the Indian states, it is difficult to analyze the impact of COVID-19 on the whole of India simultaneously.

Figs. 7 and 8 show the prediction of infected and death cases in India by Linear Regression, SVR and Random Forest and it can be concluded from graphs that the Random Forest shows the best performance among them.

4.2.3 China

By looking into the graph in Fig. 9, we can analyse that corona cases spread highly in the beginning. There were approximate 80,000 corona patients in just the first 30 days of this global pandemic. But after that, there is no exponential increase in the number of cases.

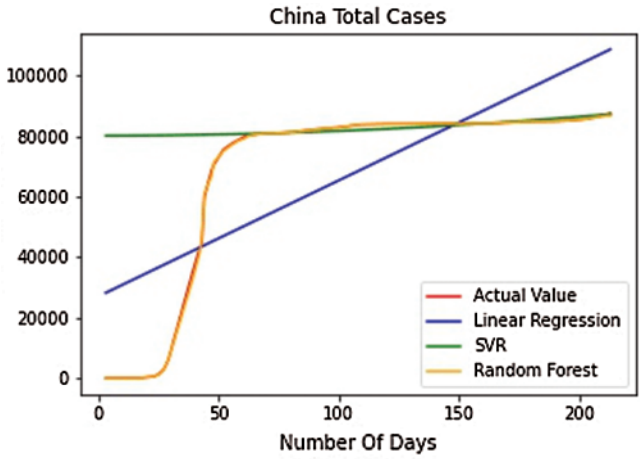

Figure 9: Total infected cases in China

Fig. 10 shows that COVID-19 takes about 4500 lives till the 31st of July in China.

China is the most populated country with a population of about 1.4 billion people. And Corona Virus started spreading from China Itself [47]. As on 31st July, 87,489 people are affected by this deadly disease in China, in which 4659 lost their lives and about 79 thousand people got recovered from it. Early detection and timely isolation help them to get rid of this deadly disease. China's government had taken proper actions to fight against this virus. They timely restricted all the international borders, suspended all the flights and trains. The people of Wuhan district confined in their houses around two months and properly followed all the guidelines provided by the government to get rid of this virus. Finally, the things get back normal in Wuhan, and China's case proves that with the combined efforts of the local public and the government authorities, it became possible to fought against this virus.

Figs. 9 and 10 shows the prediction of infected and death cases in China by Linear Regression, SVR and Random Forest and it can be concluded from graphs that the Random Forest shows the best performance among them.

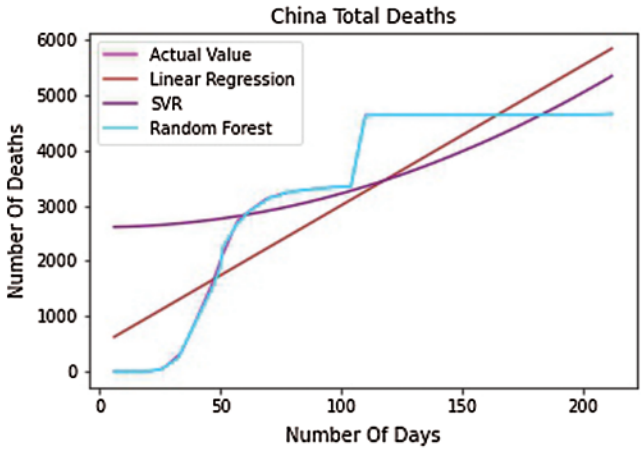

Figure 10: Total death cases in China

4.2.4 Pakistan

Pakistan has a population of approximate 22 million people. With this, it is the 5th most populated country [48]. As can be analysed from the graph shown in Fig. 11, there were zero corona cases in first 50 days. In the next 50 days, there is a gradual increase in the number of corona cases. But after that, we can see the exponential growth in the corona cases in Pakistan. Now, Pakistan is suffering from COVID-19 with approximate 278,000 corona cases.

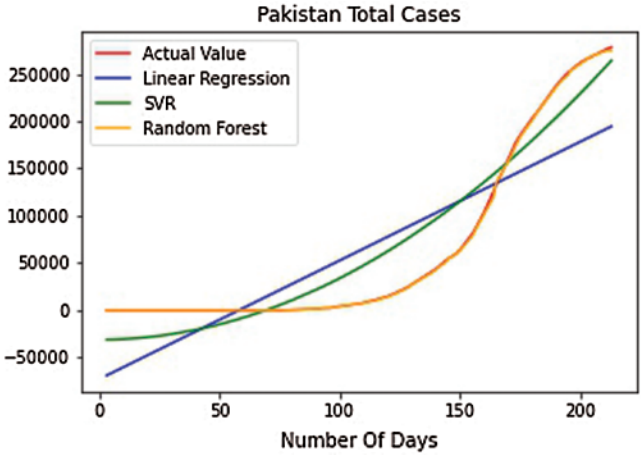

Figure 11: Total infected cases in Pakistan

Fig. 12 shows that COVID-19 takes about 6000 lives on the 31st of July in Pakistan.

The virus was first observed in a Karachi student who returned from Wuhan on February 26, 2020, and spread among four provinces by March 18, 2020. Pakistan has 4th highest number of Covid cases in Asia, and 2nd highest in South Asia. According to a report, 27% of COVID-19 infected cases were increased due to religious gatherings that took place in Lahore. Till now, the maximum number of cases was recorded in Sindh state, and the highest number of deaths were observed in Punjab. Pakistan government employed various phases of lockdown April 1 to May 9,

after which the lockdown was somewhat eased in the different provinces. Not following social distancing rules leads to an increasing number of COVID-19 cases as compared to the earlier number of cases.

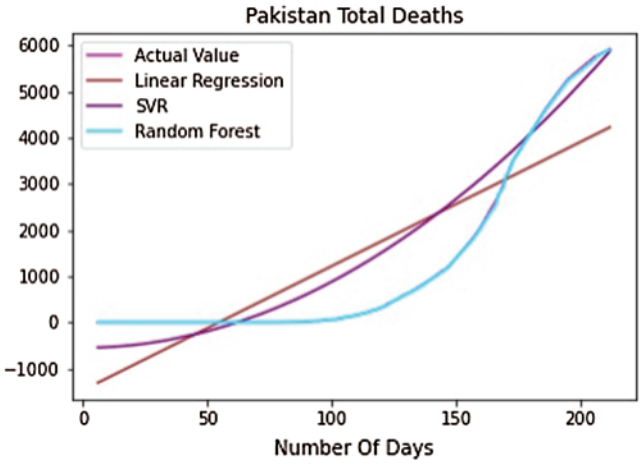

Figure 12: Total death cases in Pakistan

Figs. 11 and 12 shows the prediction of infected and death cases in Pakistan by Linear Regression, SVR and Random Forest and it can be concluded from graphs that the Random Forest shows the best performance among them.

4.2.5 Indonesia

As can be analysed from the graph shown in Fig. 13, there were around negligible cases in the initial two months of this pandemic. After that, there was a steep increase in corona cases in Indonesia. At this point in time, Indonesia has approximated 106,000 corona patients [49].

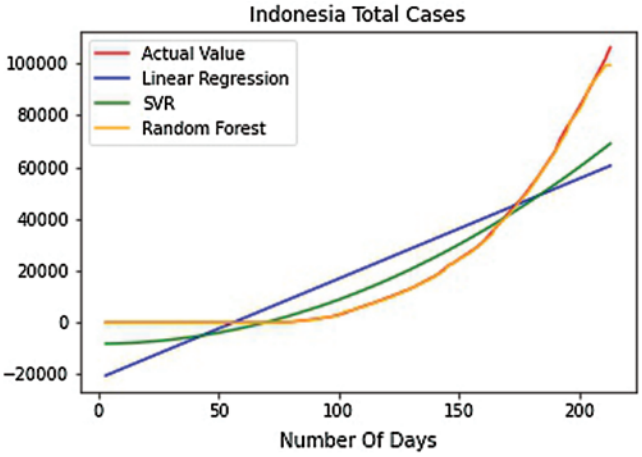

Figure 13: Total infected cases in Indonesia

Fig. 14 shows that COVID-19 takes about 5 thousand lives on the 31st of July in Indonesia.

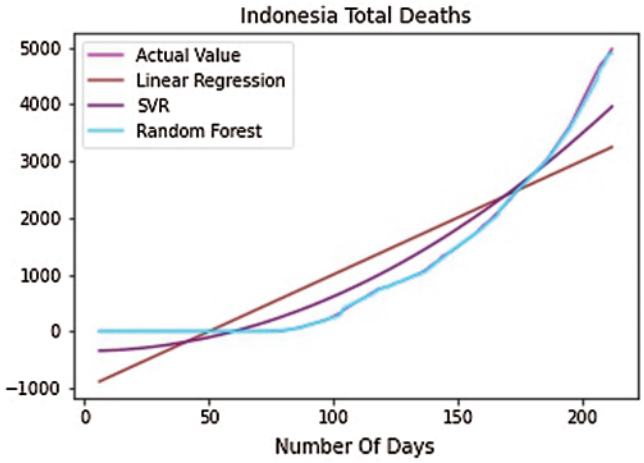

Figure 14: Total death cases in Indonesia

Indonesia is the fourth most populated country across the globe, with a population of about 27 million people. Currently, Indonesia has 106,336 cases of COVID-19, and 5,058 people lost their lives due to this deathly disease in Indonesia. The first cases of COVID-19 observed in the dance instructor and her mother, who was infected by a Japanese person. Indonesia ranks 5th in Asia in terms of death numbers due to COVID-19, and second highest in South-East Asia concerning many cases. Indonesia has conducted 1,757,425 tests so far, making it one of the lowest testing rates in the world. WHO has urged the nation to perform more tests. The government had imposed large scale interaction only in the cities or areas which are highly affected by COVID-19. And all these are the reasons for still increasing in several cases.

Figs. 13 and 14 shows the prediction of infected and death cases in Indonesia by Linear Regression, SVR and Random Forest and it can be concluded from graphs that the Random Forest shows the best performance among them.

4.3 Overall Analysis

The overall analysis is provided based on a change in mortality rate in 5 selected countries (Fig. 15). The mortality rate is a very important aspect to consider for analysis as it tells us how many people may die among 100 infected people. So, the country with more cases but with very less mortality rate (say 1% or 2%) need not worry so much about the disease than the country with a smaller number of cases but high mortality rate (of 5% or 6%). Tab. 3 shows the total infected and death cases of these top 5 most populated countries along with their population.

As we analysed, China is the most populated country and had the most corona cases in the beginning, and the curve grows exponentially day by day. But almost after one month, corona cases did not increase that much. Only 4000 cases increase in the last 5 months and their recovery rate is also very high. China ranks 32 worldwide in corona cases despite being the most populated country. This proves that the population is not a factor for more corona cases in any country. China's government had taken proper actions to fight against this virus. They timely restricted all the international borders, suspended all the flights and trains. The people of Wuhan district confined in their houses around two months and properly followed all the guidelines provided by the government to get rid of this virus. Although China has very few numbers of active cases, the mortality rate is very high. (see Tab. 3). The mortality rate of China is highest among these 5 countries and it is a matter of extreme concern.

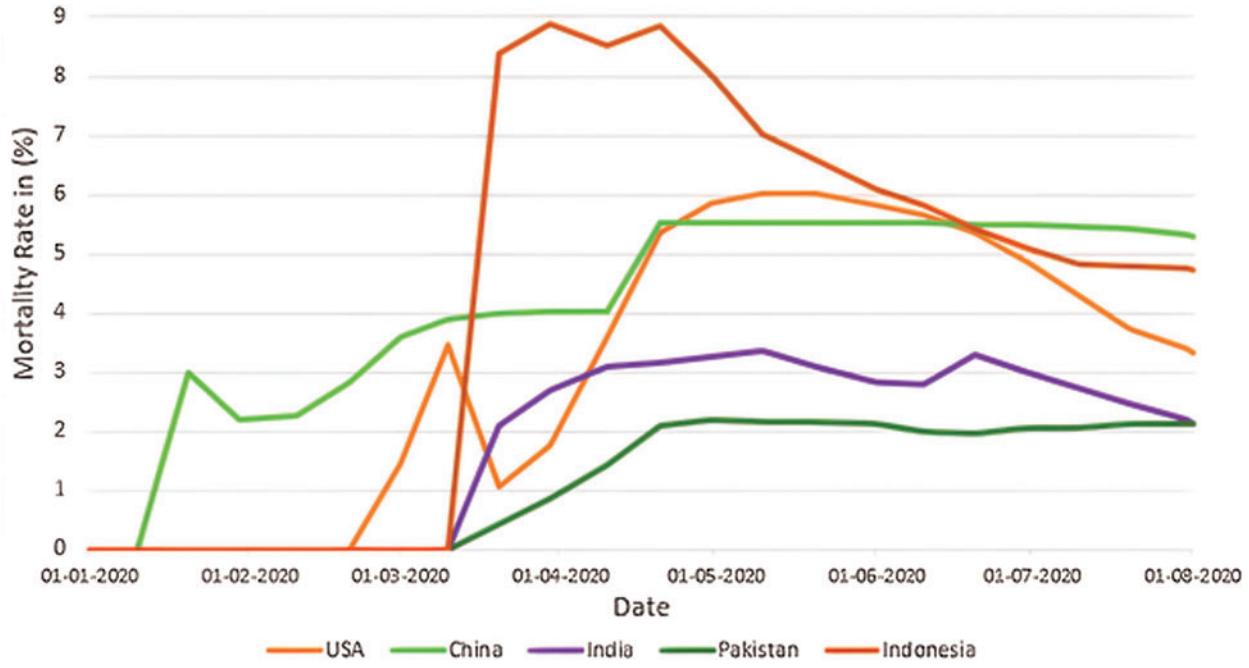

Figure 15: Change in mortality rate (%) of 5 selected countries

Table 3: Total infected and death cases of 5 most populated countries as on 31st July 2020

Countries	Population	Total cases	Total deaths	Mortality rate (%)
US	330769230	4495014	152070	3.38
India	1378364845	1638870	35747	2.18
China	1438656995	87489	4659	5.32
Pakistan	220367900	278305	5951	2.14
Indonesia	273173742	106336	5058	4.76

India is 2nd most populated country and rank 3rd worldwide in corona cases. Till 1st June, there is a very slow increase in corona cases in India. It is because of the four phases of lockdown imposed by govt and because of the health-care workers and doctors who sacrifice their sleep for treating the corona patients. But the cases increase rapidly after when some restrictions are removed from lockdown (or since the unlock down starts). But the mortality rate of India is very less than in China or the US. And from Fig. 15, we can see the continuous decrease in the mortality rate in last month.

The US is 3rd most populated country in the world with the geographical area of 9.834 million km^2 [50], and have the highest number of coronavirus cases in the world. But the primary reason for having such a high number of cases is the delay showed off in the testing, not the population. Also, the United States government initially refused to relax the regulations, which does not allow the health departments and the states to develop their testing kits based on WHO guidelines. Although the US has much more corona cases than China or Indonesia. But, the mortality rate of the US is less than these countries. From Fig. 15, we can analyse that, at first

there is a linear increase in mortality rate till 1st March. Then the mortality rate starts decreasing for the next month. After that, the curve grows exponentially for 2 next months. And from 1st June, there is a continuous decrease in mortality rate till now.

Indonesia is 4th most populated country but rank 23rd across the globe in several corona cases. But it ranks 5th in Asia in terms of death numbers due to COVID-19. The mortality rate of Indonesia is quite high (4.76%) that comes after China among the 5 most populated countries. There was the least number of tests conducted so far in Indonesia concerning its population. The government had imposed large scale interaction only in the cities or areas which are highly affected by COVID-19. And all these are the reasons for such a scenario.

Pakistan with a population of 22 million people makes it 5th most populated country. It has 4th highest corona cases in Asia and 2nd highest in South-Asia. But it ranks 14th worldwide and also it has very fewer active cases remaining. The mortality rate of Pakistan is least among these 5 countries. But the matter of concern is that the mortality rate goes on increasing every day (see Fig. 15). The govt of Pakistan must follow some strict measures to bring down the number of cases and also the death rate of people.

After analysing top five most populated countries individually, we can conclude that the large population is not a cause of the increase in the COVID-19 cases in any country. As, all the countries having different rules, regulations, norms, culture, and medical facilities, so on these factors the COVID-19 cases are counted.

5 Discussion

The RMSE values of all the three algorithms, namely Linear Regression, SVR and Random Forest Regression are given in Tabs. 4 and 5.

Table 4: RMSE values of regression algorithms of total cases prediction

Countries	LR	SVR	RF
US	520918	236731	41047
India	275203	267763	29296
China	20483	32620	449
Indonesia	16189	11516	1167
Pakistan	50786	30199	1389

Table 5: RMSE values of regression algorithms of total deaths prediction

Countries	LR	SVR	RF
US	19284	14046	751
India	5927	4302	248
China	771	1178	26
Indonesia	678	358	26
Pakistan	1053	754	21

From Tabs. 4 and 5, we concluded that the Random Forest Algorithm shows the minimum RMSE value for each considered country, so Random Forest is the best algorithm for prediction of total infected and total death cases.

Although the Random Forest algorithm shows minimum RMSE values, still the prediction model having a lot of limitations. The proposed model is only able to predict the total infected and death cases for a similar environment. i.e., the circumstances assumed to be same. It does not consider the following scenarios:

(a) If any country will able to develop COVID-19 vaccine in future.
(b) If any country will terminate the lockdowns completely and people will not follow social distancing any more.
(c) If any country will suddenly increase the COVID-19 tests daily.

All of these scenarios can change the complete picture of the country. Our proposed model is not considering these scenarios. This model is utilized to get insights about the infected and death cases in any country if the proper measures were not taken. Also, the main focus of this paper is to analyse the impact of COVID-19 on the top 5 most populated country for around 7 months. The model is proposed to analyze the same for future dates.

6 Conclusion

This research aims to develop the best machine learning model to predict the rise or fall in death rates and total cases due to COVID-19 in most five populated countries and analysis various reasons for such situations. This paper concludes that the Random Forest Regression model gives the minimum error (RMSE) as compared to Linear Regression and Support Vector Regression model in both cases, i.e., in the prediction of total cases and total deaths as well (see Tabs. 4 and 5). And after analyzing each of the most five populated countries, we can conclude that the population is not the factor for higher mortality rate and rise in the number of cases of COVID-19 in any country. Instead, it depends upon various other factors like population density, rules and regulations, norms, culture, diversity, and the medical facilities of that country. Along with government initiatives, the attitude of the public towards the government actions is also matters.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F. et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet, 395(10223),* 507–513. DOI 10.1016/S0140-6736(20)30211-7.
2. Hui, D. S., Azhar, E. I., Madani, T. A., Ntoumi, F., Kock, R. et al. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases, 91,* 264–266. DOI 10.1016/j.ijid.2020.01.009.
3. Wang, C., Horby, P. W., Hayden, F. G., Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. *Lancet, 395(10223),* 470–473. DOI 10.1016/S0140-6736(20)30185-9.

4. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet, 395(10223),* 497–506. DOI 10.1016/S0140-6736(20)30183-5.
5. Jung, S. M., Akhmetzhanov, A. R., Hayashi, K., Linton, N. M., Yang, Y. et al. (2020). Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. *Journal of Clinical Medicine, 9(2),* 523. DOI 10.3390/jcm9020523.
6. Rossignol, J. F. (2016). Nitazoxanide, a new drug candidate for the treatment of middle east respiratory syndrome coronavirus. *Journal of Infection and Public Health, 9(3),* 227–230. DOI 10.1016/j.jiph.2016.04.001.
7. Paules, C. I., Marston, H. D., Fauci, A. S. (2020). Coronavirus infections—More than just the common cold. *Journal of the American Medical Association, 323(8),* 707–708. DOI 10.1001/jama.2020.0757.
8. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B. et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet, 395(10224),* 565–574. DOI 10.1016/S0140-6736(20)30251-8.
9. Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H. et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *Lancet, 395(10223),* 514–523. DOI 10.1016/S0140-6736(20)30154-9.
10. Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L. et al. (2020). Real estimates of mortality following COVID-19 infection. *Lancet Infectious Diseases, 20(7),* 773. DOI 10.1016/S1473-3099(20)30195-X.
11. Lovelace Jr., B. (2020). The coronavirus may de deadlier than 1918 flu: Here's how it stacks up to other pandemics. https://www.cnbc.com/2020/03/26.
12. Zhou, S. K., Zhou, J., Comaniciu, D. (2007). A boosting regression approach to medical anatomy detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. DOI 10.1109/CVPR.2007.383139.
13. Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D. (2008). Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features. *Institute of Electrical and Electronics Engineer Transactions on Medical Imaging, 27(11),* 1668–1681.
14. Zheng, Y., Georgescu, B., Comaniciu, D. (2009). Marginal space learning for efficient detection of 2D/3D anatomical structures in medical images. *International Conference on Information Processing in Medical Imaging*, pp. 411–422. Berlin, Heidelberg: Springer.
15. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E. (2010). Regression forests for efficient anatomy detection and localization in CT studies. *International MICCAI Workshop on Medical Computer Vision*, pp. 106–117. Berlin, Heidelberg: Springer.
16. Savarino, A., Boelaert, J. R., Cassone, A., Majori, G., Cauda, R. (2003). Effects of chloroquine on viral infections: An old drug against today's diseases. *Lancet Infectious Diseases, 3(11),* 722–727. DOI 10.1016/S1473-3099(03)00806-5.
17. Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J. et al. (2020). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) *in vitro*. *Cell Research, 30(3),* 269–271. DOI 10.1038/s41422-020-0282-0.
18. Wu, J. T., Leung, K., Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *Lancet, 395(10225),* 689–697. DOI 10.1016/S0140-6736(20)30260-9.
19. Shen, M., Peng, Z., Xiao, Y., Zhang, L. (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China. bioRxiv, 271, 2223. DOI 10.1101/2020.01.23.916726.
20. Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X. et al. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA, 323(11),* 1061–1069. DOI 10.1001/jama.2020.1585.

21. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L. et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *New England Journal of Medicine, 382(2020),* 1199–1207. DOI 10.1056/NEJMoa2001316.
22. Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G. et al. (2020). Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *New England Journal of Medicine, 382(10),* 970–971. DOI 10.1056/NEJMc2001468.
23. Ryu, S., Chun, B. C., Epidemiology, K. S. (2020). An interim review of the epidemiological characteristics of 2019 novel coronavirus. *Epidemiology and Health, 42,* 1–7. DOI 10.4178/epih.e2020006.
24. Public Health England (2019). *Seasonal influenza vaccine uptake in GP patients: Winter season 2018 to 2019*, pp. 1–35. PHE Publications.
25. GOV. UK (2020). New rules on staying at home and away from others. Embassy of The Republic of Indonesia in London: The United Kingdom Accredited to The Republic of Ireland and Imo. https://kemlu.go.id/london/en/news/5831/new-rules-on-staying-at-home-and-away-from-others.
26. Poon, L. L. M., Chan, K. H., Wong, O. K., Yam, W. C., Yuen, K. Y. et al. (2003). Early diagnosis of SARS coronavirus infection by real time RT-PCR. *Journal of Clinical Virology, 28(3),* 233–238. DOI 10.1016/j.jcv.2003.08.004.
27. Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N. et al. (2020). Lung infection quantification of COVID-19 in CT images with deep learning. arXiv preprint arXiv: 2003.04655, pp. 1–23.
28. Davies, N. G., Kucharski, A. J., Eggo, R. M., Gimma, A., Edmunds, W. J. et al. (2020). Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: A modelling study. *Lancet Public Health, 5(7),* e375–e385. DOI 10.1016/S2468-2667(20)30133-X.
29. Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. *Science of the Total Environment, 729,* 138817. DOI 10.1016/j.scitotenv.2020.138817.
30. Yan, L., Zhang, H. T., Xiao, Y., Wang, M., Sun, C. et al. (2020). Prediction of criticality in patients with severe COVID-19 infection using three clinical features: A machine learning-based prognostic model with clinical data in Wuhan. DOI 10.1101/2020.02.27.20028027.
31. Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R. et al. (2020). COVID-19 outbreak prediction with machine learning.
32. Shoeibi, A., Khodatars, M., Alizadehsani, R., Ghassemi, N., Jafari, M. et al. (2020). Automated detection and forecasting of COVID-19 using deep learning techniques: A review. arXiv preprint arXiv: 2007.10785, pp. 1–20.
33. Ritchie, H. (2020). Coronavirus Source Data. Our World in Data. https://ourworldindata.org/coronavirus-source-data.
34. Narula, S. C., Wellington, J. F. (1982). The minimum sum of absolute errors regression: A state of the art survey. *International Statistical Review/Revue Internationale de Statistique, 50(3),* 317–326. DOI 10.2307/1402501.
35. Barzilai, J., Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis, 8(1),* 141–148. DOI 10.1093/imanum/8.1.141.
36. Smola, A. J., Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14(3),* 199–222. DOI 10.1023/B:STCO.0000035301.49549.88.
37. Rosenbaum, L., Dörr, A., Bauer, M. R., Boeckler, F. M., Zell, A. (2013). Inferring multi-target QSAR models with taxonomy-based multi-task learning. *Journal of Cheminformatics, 5(1),* 33. DOI 10.1186/1758-2946-5-33.
38. Chapelle, O., Vapnik, V. (2000). Model selection for support vector machines. In: *Advances in Neural Information Processing Systems,* pp. 230–236.
39. Lin, Y., Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association, 101(474),* 578–590. DOI 10.1198/016214505000001230.
40. Breiman, L. (2004). *Consistency for a simple model of random forests*. Statistical Department. University of California at Berkeley.

41. Yang, J., Gong, J., Tang, W., Shen, Y., Liu, C. et al. (2019). Delineation of urban growth boundaries using a patch-based cellular automata model under multiple spatial and socio-economic scenarios. *Sustainability, 11(21),* 6159. DOI 10.3390/su11216159.
42. Lehmann, E. L., Casella, G. (2006). *Theory of point estimation*, pp. 1–493. Springer Science & Business Media.
43. Willmott, C. J., Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30(1),* 79–82. DOI 10.3354/cr030079.
44. JHUoMCR, C. (2020). COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://coronavirus.jhu.edu/map.html.
45. Greene, J. P., Pole, J. R. (2008). *A companion to the American revolution*, vol. 17, pp. 1–796. John Wiley & Sons.
46. Dyson, T. (2018). *A population history of India: From the first modern people to the present day*, pp. 1–303. Oxford, UK: Oxford University Press.
47. Zhang, W., Wong, C. M. (2003). Evaluation of the 1992–1999 world bank schistosomiasis control project in China. *Acta Tropica, 85(3),* 303–313. DOI 10.1016/S0001-706X(02)00263-2.
48. Akhtar, S., Dhanani, M. R. (2012). Rural population distribution in Sindh, Pakistan—A quantitative analysis. *Sindh University Research Journal (Science Series), 44(3),* 411–416.
49. Liu, Y., Yamauchi, F. (2014). Population density, migration, and the returns to human capital and land: Insights from Indonesia. *Food Policy, 48,* 182–193. DOI 10.1016/j.foodpol.2014.05.003.
50. US Census Bureau, United States. Bureau of the Census (2009). *New York, 2000: 2000 census of population and housing,* vol. 23, pp. 1–21. US Department of Commerce, Economics and Statistics Administration, US Census Bureau: Summary Social, Economic, and Housing Characteristics.