COVID 19 Prediction using Regression, Time Series and SIR Model

Sushila Ratre[1], Rohan Sharma[2], Mudra Kevadia[2], Venkatsaigautam Bathina[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Amity University Mumbai, India
[2]UG Scholar, Department of Computer Science and Engineering, Amity University Mumbai, India

---

Abstract *- The paper is focused primarily upon the prediction of the ongoing COVID-19 outbreak in India using the Machine Learning approach. The outspread of COVID-19 within the whole world has put the humanity at risk. The resources of some of the most important economies are wired because of the big infectivity and transmissibility of this disease. Due to the growing magnitude of number of cases and its subsequent stress on the administration and health professionals, some prediction strategies would be needed to predict the quantity of cases in future. During this paper, data-driven estimation methods have been used like Regression Analysis, SIR Modelling and Time Series Analysis. The model is prognosticating the number of confirmed, recovered, and death cases based on the data acquired from July 12, 2020, to August 20, 2021.*

Keyword: *India; COVID 19; SIR Model; Linear Regression; Reproduction Number; Time Series Analysis.*

1. INTRODUCTION

The Coronavirus (COVID-19) began spreading widely around December 2019 in the Wuhan region of China. The outbreak of the virus could be associated with exposure in a seafood market in Wuhan. According to the analysis of various researchers, bats may have been the original host of the virus. Novel coronavirus sickness (2019-nCoV or COVID-19) reported from the urban center, which has affected Asian nations, Japan, North Korea, and the United States, has been confirmed to be a new coronavirus variant[1]. As of now, worldwide in total of 217 million cases have been confirmed, out of which 212.5 million have recovered, and 4.5 million have succumbed to the coronavirus. The coronavirus has mutated in multitude of ways and can even spread through aerial/aerosol medium. It's the deadliest plague outbreak the modern century has ever witnessed. Coronavirus poses a threat to the health and safety of people all over the world. Older people and those with pre-existing medical conditions are at a higher risk. It has spread at an alarming rate, bringing economic activity to near-standstill and causing unemployment in various sectors.

In recent analysis it is evidenced that unlike coronavirus, in nearly every century, the world is ruined and plagued by a virulent sickness. Recent studies have shown that in every hundred years the world has witnessed virulent disease of such high magnitude; take for instance the 1720 plague, infectious disease occurrence in 1820, in 1920 Spanish grippe, and now the 2020 Coronavirus [2].

India reported its first COVID-19 case on January 30 in Thrissur, Kerala where the patient had a travel history from Wuhan. In a recent report it was confirmed that India has now become the second worst-affected country by Coronavirus. However, India's recovery rate is found to be better than that of USA, which is ranked first in the number of cases registered. On an average for every 1 million cases, roughly 91 per cent people have recovered from the virus in India. The states with highest percentage of Confirmed cases are Maharashtra, Kerala, and Karnataka. Maharashtra is the worst affected state in India with more than 6.4 million confirmed cases and the highest death toll in the country.

Rules and regulations are being imposed by the Government of India to contain the virus and decrease the spread of infection. It is pivotal to evaluate and predict how the virus is advancing among the population, which would in turn help in estimating the healthcare requirements, allocation of resources and protective measures to be taken by every individual. Hence, with this view of helping the Government as well as other healthcare professionals, we have attempted to create a prediction model of the COVID-19 for India using Machine Learning approach, with the intent to determine its magnitude and thereby take precautionary measures.

2. LITERATURE SURVEY

In the paper Saud Shaikh et al., [3] tried to determine the ideal regression model for an in-depth diagnosis of the unprecedented coronavirus in India. It was implemented using two regression models namely linear and polynomial based on the data available from March 12 to October 31, 2020. Furthermore, the time series forecasting model was being employed to vaticinate the total count of confirmed cases in the future. The results depict that the effects of the virus in India was expected to be high between the August 2020 and September 2020 and controlled towards the end of October 2020. The number of predicted confirmed cases was expected to reach above 73 Lakh, and thus the total deaths to be above 1 Lakh by October 31, 2020. The predicted number of cases supports the data available between March 12 and October 31, 2020, by employing polynomial Degree 5 model.

In the paper M Rohini et al., [4] presented the prediction and analysis of COVID-19 using various supervised machine learning algorithms. The algorithms used in these models categorize the COVID patients based on several subsets of features and forecast their likeliness to get affected to this disease. This model was tested with 20 metrics comprising of the patient's geographical location, travel history, health record statistics, etc., to predict the severity of the case and the conceivable consequence. The experimental results showed that the model developed using KNN algorithm provided best performance in terms of accuracy with 98.34% when analogized to other models developed with SVM, DT and RF.

In the paper [5], Vakula Rani J and Aishwarya Jakka focussed on evaluating the numbers of COVID-19 confirmed cases in the country and their implications in the future, using different models such as sigmoid modelling, ARIMA, SEIR model and LSTM. The information-driven forecasting method was used to approximate the average number of confirmed cases in coming months. The LSTM model gave very promising results than other tested models. It was predicted that it will cross 2 Cr confirmed case by the end of August 2020. By July 14, 2020, there were around 9.36 Lakh confirmed cases in India, and at the end of August 2020, it was guesstimated to be more than 2.0 Cr cases.

In [6], Deep Learning models were utilized for foreseeing the coronavirus positive cases in India by Debanjan Parbat and Monisha Chakraborty. RNN based LSTM variations such as Deep LSTM, Convolutional LSTM, and Bidirectional LSTM was connected on Indian datasets to foresee the no. of positive cases. The proposed work employees support vector regression model to vaticinate the total number of cases. The data was collected of approximately 60 days starting from 1st March 2020. The total number of cases as on 30th April was estimated to be 35,043 confirmed cases with total of 1,147 deaths and numerical value of recovered patients to be about 8,889. The model was developed using Python programming language, providing an accuracy of roughly above 97% in predicting cases and 87% accuracy in predicting daily new cases. The methodology turned very efficacious and generated better accuracy than the other applied and tested models.

In [7], Anuradha Tomar and Neeraj Gupta utilized data-driven estimation strategies like long short-term memory (LSTM) and curve fitting for the expectation of the range of COVID-19 cases in India 30 days ahead in advance and the impact of preventive measures like social separation and lockdown have on the spread of the virus. The forecast of different parameters (no. of positive cases, no. of recovered cases, etc.) gotten by the proposed strategy was precise.

3. PROPOSED MODEL

The process to determine the new cases in the near future is proposed as shown in the figure:

The aim is to determine the optimal regression model for an in-depth analysis of the novel coronavirus in India. This is implemented using three distinct Methodologies namely Regression Algorithm, Time Series and SIR Model. The COVID-19 dataset for India is being utilized for serving the research of this paper [8].

After splitting the dataset into training and testing in the ratio of 70 is to 30, OLS regression is then carried out following which a equation of a line is obtained in the form of y=mx+c. Here y denotes the dependent variable which basically stands the number of cases on the following day whereas x denotes the independent variable which stands for number of days from the start of the pandemic.

In the SIR Model one of the most crucial parameters for evaluation is R0, also known as the Reproduction Number which fundamentally denotes that one person can potentially transmit the virus to how many other people.

In time series analysis when feeding the data to our Model and considering various parameters the model first performs the modelling part by splitting the data into training and testing, following which it plots a graph with three main boundaries the upper boundary, the lower boundary and the one with prediction. The upper and the lower boundary indicate the maximum and minimum cases that can be detected on that particular day. For forecasting the future trend of these cases, we are utilizing the time series forecasting approach.

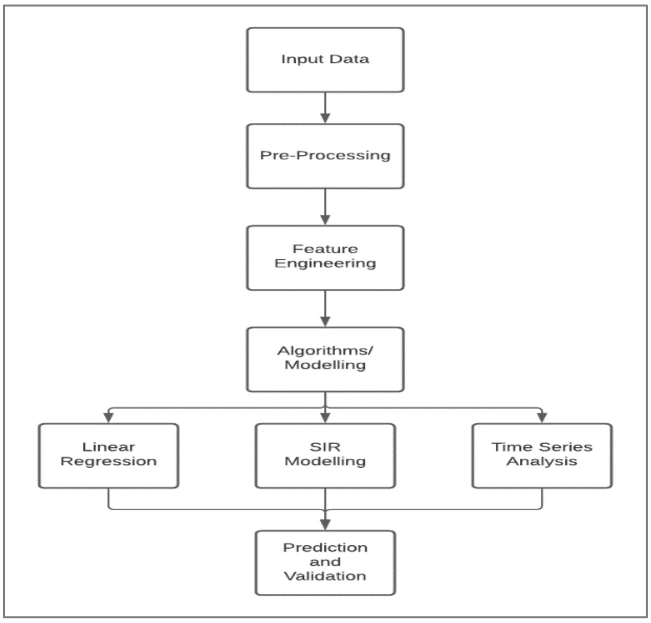

Figure 1. Flowchart for Proposed Model

4. IMPLEMENTATION

A] DATA ACQUISITION

The data used for designing the model is acquired from open source data accessible on covid19.india.org[8]. It consists of 4 attributes namely date, number of positive, recovered and deceased figures.The data has been further splitted into train and test sets, in the ratio 70 is to 30 respectively, which makes it easier to work with and evaluate the real time scenario.

The below figure demonstrates the data utilized as input for the model which starts noting the registered cases from 12 July 2020 [9].

Date	Confirmed	Recovered	Deceased
12 July 2020	29106	18198	500
13 July 2020	28178	17683	541
14 July 2020	29917	20977	587
15 July 2020	32607	20646	614
16 July 2020	35468	22867	680
17 July 2020	34820	17476	676
18 July 2020	37411	23583	543

Table I. Cases Per Day in India

B] MODELLING

B.1 Linear Regression

Linear Regression, a statistical model used for predictive analysis, is used to obtain the dependent variable value, the number of cases based on the independent variable, number of positive cases, recovered cases and deceased, one at a time. It measures the association between the two variables and the regression line thus formed provides the best fit line in accordance with the model [10].

B.2 Prophet

Time series analysis is performed, using the Facebook prophet library, to understand time-based trends of the data points which are ultimately critical in any project. The basic objective of this is to determine a model that describes the pattern of the time series and could be used for forecasting. Unlike the classical time series forecasting techniques, prophet enables intuitive parameters which are easy to tune. The library is considered for making predictions owing to its promising results.

B.3 SIR

A typical SIR model specifies that at a certain time t, the population (with size N) can be classified as people who are susceptible S(t), infected I(t), and recovered R(t) according to the following series of differential equations [3]:

Susceptible Rate Equation:

$$\frac{dS}{dt} = b\ s(t)I(t) \tag{1}$$

Infected Rate Equation:

$$\frac{di}{dt} = \ b\ s(t)i(t) - \ k\ i(t) \tag{2}$$

Recovered Rate Equation:

$$\frac{dr}{dt} = k\ i(t) \tag{3}$$

In the SIR model, there is one important parameter that has to be considered, which is the R0 value (also known as reproduction rate). It signifies how much infection would be passed on by one person, potentially coming in contact with several others. Considering the current reproduction rate of India, standing out to be 1.71, it indicates that an individual in India can potentially transmit the virus 1.7 times to other people. It is specially applied by Mathematicians and used for predictions whenever the world is hit by an Epidemic.

Assuming that a time series of COVID-19 incidence was observed for cases up to a time t, the goal was to make predictions of incidence cases for the next two to three weeks.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

The validation part is achieved by checking the rmse value and adjusted r square value for Linear Regression. We have attained the rmse value as 0.5967 whereas the adjusted r square value stands out to be 0.869. The accuracy part for SIR Model and Time Series Analysis can be directly calculated from Fig.2 and Fig.3 by checking the predicted value for that particular day with respect to the actual value of cases on that day.

Results are listed in Table II which is according to the infected cases recorded in India. Also note that if the reproduction rate increases, there will be a quick increase in the transmission rate which will result in increase in average contact between infected and susceptible person. If that increases, it states that social distancing norms are not being followed properly.

Date	Actual	LR	SIR	Time Series
23 August 2021	24794	28905	35780	18378
24 August 2021	37739	33145	38907	23464
25 August 2021	46129	37134	40960	28956
26 August 2021	44550	42150	43112	34120

27 August 2021	46806	45120	45763	39671
28 August 2021	45064	45925	46923	43996
29 August 2021	43374	44987	47455	47799

Table II. Comparison Of Actual Results with The Predicted Ones with The Help of Sir Model and Machine Learning Model

It is evident from Fig.2 that the cases are expected to rise exponentially, from the period of September 2021 up till November 2022, which has also been verified by several scientific professionals. It thereby provides clear evidence for the approaching third wave in India. Naturally, the confirmed cases were seen to be high during the weekends and beginning of the week due to increased movement of masses during that period.

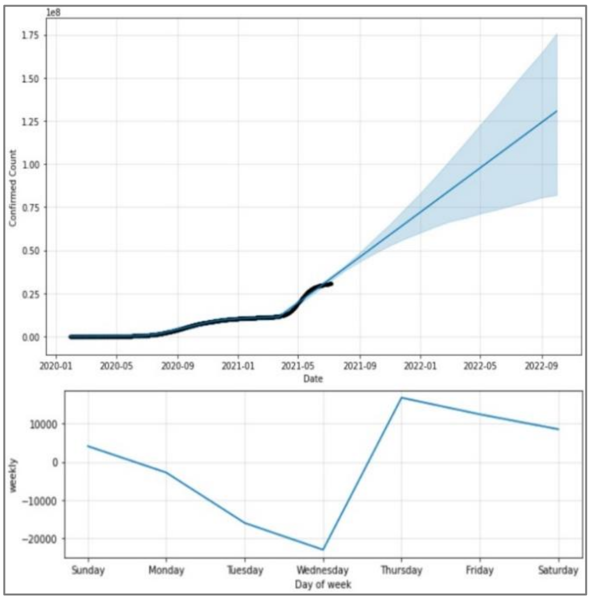

Figure 2. Prediction for Time Series Analysis

Using the Linear Regression model, the results indicate that the positive cases would rise exponentially to attain a peak, after which a plateau would be observed. The observation shows that for first 100 days from the start of pandemic, cases were increasing linearly.

From the illustrated Fig.3, it can be noted that if all the parameters continue to remain constant, considering the current reproduction rate of India, R0 = 1.71, the cases are expected to reach 4-5 Lakh per day from mid October 2021 to end of November 2021, which has been predicted by the proposed SIR model.

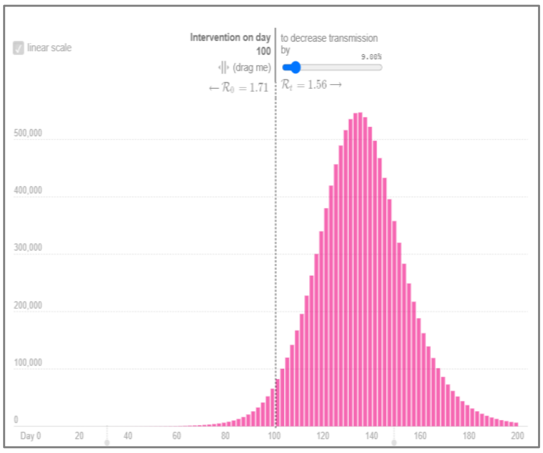

Figure 3. SIR Model Predicting Positive Cases

It is to be kept in mind that the introduction of vaccine schemes affects these figures greatly. Moreover, on 27th August 2021 and 31st August 2021, India attained a figure of more than 1 crore vaccinations. It would act as a barrier to the upcoming third wave, which in turn would hamper the predictions in the next few weeks. At the same time, however, lack of stringent rules on social distancing continues to contribute to rising cases, especially in metropolitan areas.

However, one drawback of the models was that the outcome was accurate up till a certain period of time, generally a few weeks. The reason being that Covid cases were ultimately bound to change, depending either on the restrictions imposed in specific cities or the influx of vaccination in India. As the time series analysis cannot be termed reliable for longer intervals, the experimental results are considered valid only for a short amount of time. Due to the drawbacks of our model type and insufficient data for various parameters, the model can predict accurately only for short intervals. The imposition of lockdowns, social distancing, and interactions with infected person are not persistent. Due to these daily fluctuating conditions, the validity of the model is short-lived and may fail in longer run as the accuracy would start decreasing.

The attributes such as immunity of an infected person, age of the patient, pre-existing medical conditions, can help us improve the accuracy of the model. It is essential to handle this situation by following the rules and regulations imposed by WHO and the Government of India, maintain social distancing and proper hygiene. In the coming time, with the knowledge and the experience gained from working on this model, we hope to work on larger areas using better methods and gaining more accurate predictions.

6. CONCLUSION

The Machine Learning model's accuracy was found out to be roughly 87%, followed by the SIR model which stands out to be around 82% followed by Time Series Analysis which stands out to be 73%. But the accuracy of all the 3 models is expected to fall in the long run, due to significant variations in parameters. Moreover, time series analysis was done to obtain automated forecasts. After critical analysis of the models, it was observed that the machine learning – Regression model predictions were relatively more accurate and satisfactory than SIR Model and Time Series Analysis.

The maturation of the virus can be prevented provided that the public abides by the healthcare measures and rules imposed by the respective state and the central government. Needless to say, the primary step is wearing masks in public places and sanitizing properly. Taking an approved vaccine is another footstep for avoiding getting infected. In the end, only when the citizens and the administration work hand in hand, we will be able to eradicate the SARS-CoV2 infection.

REFERENCES

[1] Jernigan, D. B., & CDC COVID-19 Response Team (2020). Update: Public Health Response to the Coronavirus Disease 2019 Outbreak - United States, February 24, 2020. *MMWR. Morbidity and mortality weekly report*, *69*(8), 216–219. https://doi.org/10.15585/mmwr.mm6908e1.

[2] Rafiq, D., Suhail, S. A., & Bazaz, M. A. (2020). Evaluation and prediction of COVID-19 in India: A case study of worst hit states. Chaos, solitons, and fractals, 139, 110014. https://doi.org/10.1016/j.chaos.2020.110014

[3] S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara and M. Roja Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 989-995, DOI: 10.1109/Confluence51648.2021.9377137.

[4] M. Rohini, K. R. Naveena, G. Jothipriya, S. Kameshwaran and M. Jagadeeswari, "A Comparative Approach to Predict Corona Virus Using Machine Learning," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 331-337, DOI: 10.1109/ICAIS50930.2021.9395827.

[5] V. R. J and A. Jakka, "Forecasting COVID-19 cases in India Using Machine Learning Models," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 466-471, DOI: 10.1109/ICSTCEE49637.2020.9276852.

[6] Parbat, D., & Chakraborty, M. (2020). A python-based support vector regression model for prediction of COVID19 cases in India. Chaos, solitons, and fractals, 138, 109942. https://doi.org/10.1016/j.chaos.2020.109942

[7] Tomar A, Gupta N., Prediction for the spread of COVID-19 in India and effectiveness of preventive measures, Sci Total Environ. 2020 Aug 1;728:138762, DOI: 10.1016/j.scitotenv.2020.138762.

[8] COVID-19 Tracker, India. https://www.covid19india.org/. Accessed April 21, 2020.

[9] COVID19 Government Measures Dataset | ACAPS. https://www.acaps.org/covid19-government-measures-dataset. Accessed April 9, 2020.

[10] Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010 Nov;107(44):776-82. doi: 10.3238/arztebl.2010.0776. Epub 2010 Nov 5. PMID: 21116397; PMCID: PMC2992018.

BIOGRAPHIES

Mrs. Sushila Ratre
Assistant Professor,
Department of Computer Science and Engineering,
Amity University Mumbai, India

Rohan Sharma
UG Scholar,
Department of Computer Science and Engineering,
Amity University Mumbai, India

Mudra Kevadia
UG Scholar,
Department of Computer Science and Engineering,
Amity University Mumbai, India

Venkatsaigautam Bathina
UG Scholar,
Department of Computer Science and Engineering,
Amity University Mumbai, India