

Predicting reduction of COVID-19 cases in India Using Machine Learning Algorithm

¹Rashmi Welekar, ²Sharvari Tapase, ³Shubhi Bajaj, ⁴Isha Pande,
⁵Abhishek Verma, ⁶Ashutosh Katpatal, ⁷Vaibhav Mishra

¹Shri Ramdeobaba College of Engineering Katol Rd, Gittikhadan,
Nagpur 440010, Maharashtra, India.

^{2,3,4,5,6,7}Department of Computer Science and Engineering, Shri Ramdeobaba College
of Engineering Katol Rd, Gittikhadan, Nagpur 440010, Maharashtra, India.

ABSTRACT

The coronavirus disease (later renamed as COVID-19) originated in Wuhan, China and now has spread throughout the world. Many methods have been implemented so as to bring the condition under control, however it has been a difficult task to predict when the number of cases will start decreasing. Hence, we are trying to predict, using the Susceptible-Exposed-Infectious-Removed (SEIR) model, when the number of cases in India will start decreasing. The implementation will start on district/city level, then to the state level and then the country level. We are going to develop a WebApp, which will maintain a dashboard of the active cases, cured cases, and the death cases daily of India. We are trying to train the model on the data from a city/country where there was a significant decrease in the cases, to check our model's accuracy. We are expecting to get a probable period of time when the cases will decrease on an average all across India.

KEY WORDS: SEIR MODEL, COVID-19, PREDICTION, REDUCTION, COVID-19 CASES IN INDIA, MACHINE LEARNING ALGORITHM

INTRODUCTION

In the month of December 2019, we found the sudden outbreak of the Coronavirus, later renamed as COVID-19, in the city of Wuhan, China. This virus then rapidly started to spread to other parts of China, as well as to other many countries of the world. The condition in India is also at par with the world. Currently, as of mid-August 2020, there have been about 22,63,951 confirmed cases, out of which 6,38,368 are currently active and 15,80,231

are recovered, with the death toll at 45,352. Because of the asymptotic nature of the virus, it is difficult to predict its behaviour, and thus, it is difficult to predict its spread. However, it is essential for every citizen to know when the number of cases will start reducing.

The problem we are trying to solve is to predict when the number of cases will start reducing. For that, we have decided to use the SEIR model. Initially, we will try to predict the reduction in some particular cities/districts, where the number of cases has significantly decreased, so that we can test our model's accuracy on historical data. If the model is accurate, we will try to expand our scope from districts to states, and then to the entire country. The reason why the decrease in the number of cases is important to the entire population is because the daily lives of each individual has come to a standstill. This model's results will be beneficial to the government, so that it can concentrate on the areas where the reduction

ARTICLE INFORMATION

*Corresponding Author: welekarr@rknc.edu

Received 11th Oct 2020 Accepted after revision 27nd Dec 2020

Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal



NAAS Journal Score 2020 (4.31)

A Society of Science and Nature Publication,
Bhopal India 2020. All rights reserved.

Online Contents Available at: <http://www.bbrc.in/>

Doi: <http://dx.doi.org/10.21786/bbrc/13.14/45>

will take place slowly, and apply the resources efficiently, rather than using them in places where the reduction in cases is going to be rapid. Similarly, the companies which have started producing sanitizers, masks, and other such sanitary products will know when to stop the production so as to save themselves from heavy losses. It will be easier for the ministries to decide when to uplift a lockdown, or when to start the schools, colleges, shops, etc. in a particular area. In short, it will be easier to predict when we can have the condition a little closer to normal.

Literature Review: Lots of research papers indicate that the research for the prediction, reduction and detection of COVID-19 is going on all over the world. The excess mortality in England [Peckham, R., 2020] was analyzed, to see the difference between the number of deaths in the presence of COVID-19 and its absence. We can see from the paper on “Outbreak and Analysis of COVID-19” [Hamzah, F.B., 2020] that a corona tracker is being developed, so that the cases around China can be tracked, and they are also working on the sentiment analysis of people to know the effect of this virus on the mental health of the people. They are using the SEIR model for prediction. Whereas from the “Journal of Thoracic Disease” we can see that [Yang, Z., 2020] they have integrated population migration data before and after January 23 and most updated COVID-19 epidemiological data into the Susceptible-Exposed-Infectious-Removed (SEIR) model to derive the epidemic curve. They have also used an artificial intelligence (AI) approach, trained on the 2003 SARS data, to predict the epidemic. A machine learning forecasting model for COVID-19 pandemic in India [Sujath, R., 2020] was created, to detect the number of cases in India. They have used the Linear Regression model with FANN for the prediction.

Why the prediction of COVID-19 cases so difficult, is addressed in the [Roda, W.C., 2020] paper. They have taken into account that the SIR model works better, and have implied that using more complex models will not guarantee reliability, as compared to simpler models. One more study [Gupta, R., 2020] on the outbreak situation in India can be seen. They have suggested that only the essential services and goods should be made available and a national lockdown should be observed for at least a month. Their study will be helpful to the Government of India and the states in India, and also the doctors, health workers, scientists and researchers. A study in Barcelona [Baena-Díez, J.M., 2020] was made where they had two parameters: age of the patient and the socio-economic status. They have observed that the impact of the virus has hit the hardest in the lower income section of the society. Lower the income, more the impact. They have concluded that not only the people who already have ailments should be prioritized, but also a healthcare strategy should be implemented for the whole population, especially the economically deprived regions.

We can also find similarities between the coronavirus and SARS virus that was identified in the year 2003, and [Raju, N.V., 2020] we can find the Indian publications on

SARS-CoV-2 in the WHO database on COVID-19. There has been a rise in the publications related to COVID-19 in the past few months, with authors belonging to the Indian institutes like AIIMS and ICMR. In a similar paper [Di Girolamo, N., 2020], we can find the study of characteristics of the articles or reports published during the first 3 months of the coronavirus pandemic. Their search strategy was that they searched articles having “COVID” or “COVID-19” as the keywords, and found a lot of articles during the initial 3 months of the pandemic.

In the paper [Ali, I., 2020] the highlights were that it contains detailed information on the coronavirus, advanced methods of management and prevention of the spread, treatment strategies and this article can be helped to create awareness amongst the public. In one more paper related to study in India [Anirban Ghataka, 2020], they have implemented the compartment based epidemiological model, which incorporates social distancing and lockdown strategies for the top 15 affected states in India, upto May 2020. They have also implemented how to have post lockdown relaxation measures to prevent further spread of the disease using the eSIR model.

This study [Remuzzi, A., 2020] came out when there was a surge in the number of cases in Italy, after the surge in the number of cases in China. Also, the exponential growth in the number in Italy could not be properly predicted and managed due to the difference in the social distancing measures and the capacity to arrange facilities, as compared to that of China. Coming to the part, as to how coronavirus affects people already suffering from an ailment, we can see that [Zheng, Y.Y., 2020] SARS-CoV-2 infects the host cells through the ACE2 receptors, leading to the COVID-19 related pneumonia and causing a chronic damage to the patient's cardiovascular system. Also, whether the patients suffering from diabetes and certain cardiovascular diseases are at an increased rate of infection [Fang, L., 2020] is studied. Summarizing this [Fang, L., 2020] paper, we can see that the understanding of the clinical chest CT features of the COVID patients will help to detect the traces of the virus in the early stages and access the treatment earlier. They have concluded that the chest CT scans show the high possibility of presence of COVID-19 pneumonia if the person has come in contact with a COVID affected person.

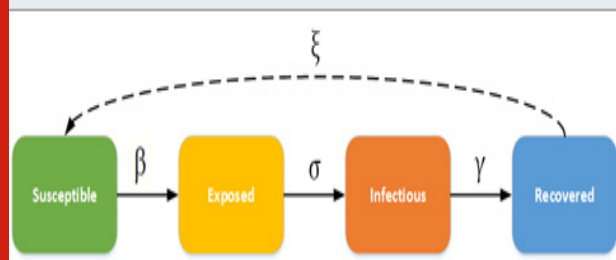
In [Jakhar, M., 2020] the paper, we can observe that the prediction on increase in cases of COVID-19 in different states of India is done using the SIR model. They came up with the results that the most affected state, having the maximum number of positive cases, was recorded in the state of Maharashtra. We can see from [Hoque, M.M., 2020] this paper, that the misconceptions about the effect of weather on the virus should be cleared up, especially in largely affected regions like the USA. Using the prediction model on the US, similar conclusions were drawn regarding India too. The relationship between weather conditions and the virus will help to identify the high-risk zones in India.

Using SIR model [Cooper, I., 2020], the effect of the coronavirus on different communities like South Korea, USA, India, etc. has been observed. Also, the data is from January to June 2020, where the difference in the number of cases is observed before and after the implementation of strict control measures. Our project is going to detect the reduction in the number of cases in India, using the SEIR model. We will try to predict the cases first on a district level, and then increase the range to the entire country.

Principles of the Proposed Method: SEIR stands for Susceptible, Exposed, Infectious, and Removed or Recovered, respectively. SIR model forms the base for the SEIR model, but SEIR adds the Exposed part as an added variable. The SEIR model then may become an SEIRS (Susceptible - Exposed - Infectious - Recovered - Susceptible) model, where already recovered people may become susceptible again.

The rate of infection, β controls the rate of spread which represents the probability of transmitting disease from a susceptible person to an infectious individual. The incubation rate, denoted by σ , is the rate of latent individuals becoming infectious (where the average duration of incubation is $1/\sigma$). Recovery rate, $= 1/D$, is determined by the average duration D of infection. For the SEIRS model, is the rate at which recovered individuals return to the susceptible status due to reasons like loss of immunity, etc.

Figure 1: SEIR Model Flowchart



The base SEIR model is as follow:

$$\begin{aligned}\frac{dS(t)}{dt} &= -\frac{\beta S(t)I(t)}{N} \\ \frac{dE(t)}{dt} &= \frac{\beta S(t)I(t)}{N} - \sigma E(t) \\ \frac{dI(t)}{dt} &= \sigma E(t) - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t)\end{aligned}$$

Figure 2: The calculations for the SEIR model

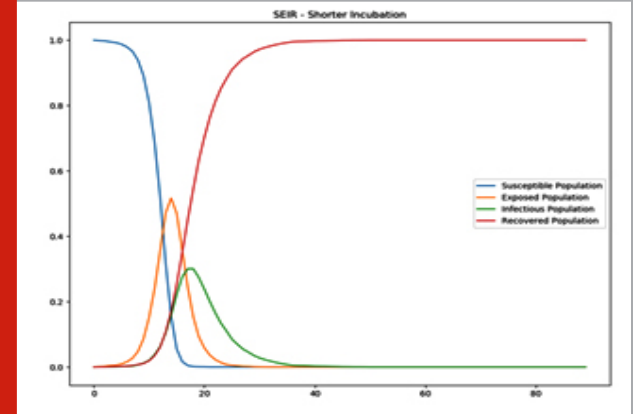
where

$$N = S + E + I + R$$

Figure 3: Total population equation

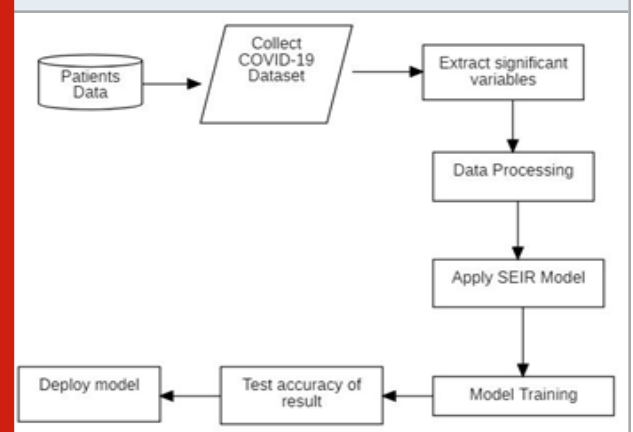
is the total population. Here, it is assumed that the latent population $[E]$ is asymptomatic but infectious, and $[I]$ refers to the symptomatic as well as the infectious population. The incubation rate, σ is defined as the rate by which the exposed individual develops symptoms.

Figure 4: The graphs plotted between different populations number of people vs infected



Research can be further done on: The high risk zones classified according to population density, age group, comorbidity, metabolic conditions patients, residential infrastructure, availability of medical aid; The variation in spread of disease if a person travels from high-risk zone to low risk zone and vice versa; To monitor if at a given time a person has gone from high-risk zone to low risk zone and vice versa; and group monitoring

Figure 5: Flow Diagram of the project



METHODOLOGY

Step1: Collect the dataset from sources. We are planning to use the dataset provided by John Hopkins University on COVID-19 and the dataset provided by Kaggle.

Step 2: Classify the data, select the data to be uploaded on the dashboard which shows the number of cases tested positive, death cases, recovered cases, etc.

Step 3: We will then classify the data required for the actual prediction and the train the data using the SEIR model.

Step 4: We will test the accuracy of the model using data from places where there has already been a significant reduction in the number of cases.

Step 5: Based on the accuracy of the model, we will decide whether to continue using the same model, and also the accuracy given by other models.

Step 6: The prediction will be carried out on a district level, first, then on the state level and then on the entire country.

Step 7: We will repeat the process of prediction every week to see the accuracy of the model as well as the change in the time period as to when the cases will actually start to decrease.

Proposed Result: We are aiming for 80-90% accuracy in prediction of the reduction in number of cases using the SEIR model. Also, we are hoping to find a near probable date, rather than a time period as to when there will be a reduction in the number of cases.

REFERENCES

- Ali, I. and Alharbi, O.M., 2020. "COVID-19: Disease, management, treatment, and social impact. Science of the Total Environment", p.138861.
- Anirban Ghataka,1,1, Ranraj Singhba Formerly of Indian Institute of Management Visakhapatnam Indian Institute of Technology Dhanbad. "COVID-19 in India: Policy Suggestions using Epidemiological Modeling"
- Baena-Díez, J.M., Barroso, M., Cordeiro-Coelho, S.I., Díaz, J.L. and Grau, M., 2020. "Impact of COVID-19 outbreak by income: hitting hardest the most deprived." *Journal of Public Health*, 42(4), pp.698-703.
- Cooper, I., Mondal, A. and Antonopoulos, C.G., 2020. "A SIR model assumption for the spread of COVID-19 in different communities." *Chaos, Solitons & Fractals*, 139, p.110057.
- Di Girolamo, N. and Reynders, R.M., 2020. "Characteristics of scientific articles on COVID-19 published during the initial 3 months of the pandemic." *Scientometrics*, 125(1), pp.795-812.
- Fang, L., Karakiulakis, G. and Roth, M., 2020. "Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?." *The Lancet. Respiratory Medicine*, 8(4), p.e21.
- Gupta, R., Pal, S.K. and Pandey, G., 2020. "A Comprehensive Analysis of COVID-19 Outbreak situation in India." *medRxiv*.
- Hamzah, F.B., Lau, C., Nazri, H., Ligot, D.V., Lee, G. and Tan, C.L., 2020. "CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction." *Bull World Health Organ*, 1(32).
- Hoque, M.M., Saima, U. and Shoshi, S.S., 2020. "Correlation of Climate Factors with the COVID-19 Pandemic in USA." *Biomedical Statistics and Informatics*, 5(3), p.65.
- Jakhar, M., Ahluwalia, P.K. and Kumar, A., 2020. "COVID-19 Epidemic Forecast in Different States of India using SIR Model." *medRxiv*.
- Peckham, R., 2020. "COVID-19 and the anti-lessons of history." *The Lancet*, 395(10227), pp.850-852.
- Raju, N.V., 2020. "Indian Publications on SARS-CoV-2: A Bibliometric Study of WHO COVID-19 Database." *medRxiv*.
- Remuzzi, A. and Remuzzi, G., 2020. "COVID-19 and Italy: what next?." *The Lancet*.
- Roda, W.C., Varughese, M.B., Han, D. and Li, M.Y., 2020. "Why is it difficult to accurately predict the COVID-19 epidemic?." *Infectious Disease Modelling*.
- Sujath, R., Chatterjee, J.M. and Hassanien, A.E., 2020. "A machine learning forecasting model for COVID-19 pandemic in India." *Stochastic Environmental Research and Risk Assessment*, p.1.
- Yang, Z., Zeng, Z., Wang, K., Wong, S.S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z. and Liang, J., 2020. "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions." *Journal of Thoracic Disease*, 12(3), p.165.
- Zheng, Y.Y., Ma, Y.T., Zhang, J.Y. and Xie, X., 2020. "COVID-19 and the cardiovascular system." *Nature Reviews Cardiology*, 17(5), pp.259-260.