Covid-19 Prediction with Vaccinations as a factor: the case of India and USA

Vishruth Veerendranath*, Vibha Masti*, Harshith Mohan Kumar*
*Department of Computer Science and Engineering, PES University,
Bengaluru, India
{vishruth, vibha, harshithmohankumar}@pesu.pes.edu

***Abstract*—This paper is a study on methods to predict COVID-19 cases in India and USA while accounting for vaccination as a factor. The study explores various mathematical and predictive models to predict if and when a third wave of infection might occur in India and the rate at which vaccination needs to occur in order to prevent a severe wave in the future. The paper also contrasts the scenarios of USA and India to find out the effect of vaccination in curbing transmission.**

***Index Terms*—SIR model, SEIR model, COVID-19, vaccine, pandemic, prediction, forecasting**

I. Introduction

The COVID-19 pandemic situation India was deadly during its second wave in May 2021. In order to prevent a third wave of similar magnitude, the Indian population must be vaccinated rapidly. In the case of USA, there was a surge in cases in August 2021, which was mainly due to the unvaccinated population [1] . Since USA had a rapid vaccination response while India had a relatively slower one, we hope to explore the effect that vaccination had on case load, by comparing a similar timeline for both India and USA.

The objective of this paper is to use data on COVID-19 confirmed cases and vaccine doses administered (first and second doses) to predict if and when a third-wave of infection might occur in India and USA. We also hope to explore the different methods/models to achieve accurate predictions. We will also use this data to determine the rate of vaccination required to prevent the occurrence of a third wave in both countries, and the effect of vaccination in reducing transmission.

II. Related Work

A. Mathematical Models for Transmission of Infection

The first objective of this paper is to come up with an accurate mathematical model for transmission prediction. The earliest prediction model for the 2019-nCov virus is described by Zhuo et al. [2]. The authors have used the Susceptible-Exposed-Infected-Recovered (SEIR) compartment model, which is an improvement over the SIR model for epidemics first proposed by Kermack and McKendrick in [3], to come up with a simple *Basic Reproduction Number* $\mathrm{R_t}$ for 2019-nCov. When $\mathrm{R_t} < 1$, the epidemic can be considered under-control.

The probabilistic analysis of the SEIR compartment model lead to the below equations for each of the compartments int the SEIR model:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\frac{\beta S(t) I(t)}{N} \tag{1}$$

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} = \frac{\beta S(t) I(t)}{N} - \gamma_1 E(t) \tag{2}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \gamma_1 E(t) - \gamma_2 I(t) \tag{3}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma_2 I(t) \tag{4}$$

Here N is the total individuals in the the system i.e sum of all the compartments. β, γ_1 and γ_2 are the probabilities of a person being exposed, infected and recovered respectively.

Based on the above constants defined int the SEIR model, the authors estimated the *basic reproduction number* R_t by taking $\lambda = ln\frac{Y(t)}{t}$ which gives the *exponential growth* and *Y(t) is the number of infected people with symptoms*. $T_E = 1/\gamma_1$ and $T_I = 1/\gamma_2$ give the time for which a person was exposed and infected respectively. Taking generation time $T_g = T_E + T_I$ and $\rho = T_E/T_g$, R_0 was finally derived as

$$R_0 = 1 + \lambda T_g + \rho(1-\rho)(\lambda T_g)^2 \tag{5}$$

The authors then estimated R_0 based on the Covid-19 data available at the time, which was only limited to Wuhan and some other parts of China. The assumptions for estimation were made based on the limited and possibly unreliable data, which might've lead to an *inaccurate* R_0. The authors concluded that the 2019-nCov had a higher transmissibility than SARS, but not an exceptionally high as other viruses such as Zika and MERS had a higher transmissibility.

The inaccuracies and pitfalls of SIR models have been described in Moein et al. [4]. The authors applied the simple SIR model and fit it to data aggregated in Ishfan, to obtain R_0 in 3 different scenarios namely - good, feasible and bad based on stringency of restrictions. The SIR model was unable to predict the second peak, but was able to predict the controlled epidemics i.e the first peak. The reasons for the inability of SIR models to predict the second peak was attributed to - *oversimplification* of the epidemic and ignorance to other socio-economic and behavioral changes. This presents the need for a more sophisticated model.

Work in still ongoing in coming up with a modified SIR model that takes vaccinations into account, although authors haven't settled on a particular model. Some of the new modified SIR models are described below. In the paper by Webb [5], the author uses a model to predict the effectiveness of vaccination in controlling the spread of the virus in the United States of America (USA). The study improves upon previous models by taking into account the level of relaxation of social distancing measures, the estimation of the number of unreported cases, number of symptomatic and asymptomatic cases and the level of vaccination achieved.

The author utilises a weekly rolling average of daily reported cases to smooth out the curve. The author also makes the assumption that 1 in 4 COVID-19 infections were reported in the USA (taken from the Centers for Disease Control and Prevention - CDC).

The model used is a modification of the SIR model, known as the SUIR compartment model. The model is again a system of ordinary differential equations like the SEIR model, where $S(t)$ refers to susceptible, $I(t)$ refers to asymptomatic infectious, $R(t)$ refers to the number of reported symptomatic infected and $U(t)$ refers to the number of unreported symptomatic. The parameter τ represents the level of relaxation of social distancing measures as a result of vaccination.

The model predicts that with slow vaccination rates (as is observed due to vaccine hesitancy) and high levels of social distancing restorations, the daily reported cases first increase slowly (starting from June 2021) and then decrease to relatively low levels in January 2022. While this model does not perfectly predict the surge in cases observed in USA in August 2021, it shows that accounting for social distancing measures in the model shows an improvement in the prediction.

Two other recent works [6], [7] have also included Vaccinated, Deceased and Quarantined compartments into the SIR and SEIR compartments models to propose namely - the *SIRDV* compartment model in Userwood et al [6] and the *Extended SEIR* model in Ghoustine et al [7]. The more comprehensive model of the 2 - the *Extended SEIR model* described the epidemic system as per the below 7 differential equations for each compartment

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \Lambda - \beta S(t)I(t) - \alpha S(t) - \mu S(t) \tag{6}$$

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} = \beta S(t)I(t) - \gamma E(t) + \sigma\beta V(t)I(t) - \mu E(t) \tag{7}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \gamma E(t) - \delta I(t) - \mu I(t) \tag{8}$$

$$\frac{\mathrm{d}Q(t)}{\mathrm{d}t} = \delta I(t) - (1-\kappa)\lambda Q(t) - \kappa\rho Q(t) - \mu Q(t) \tag{9}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = (1-\kappa)\lambda Q(t) - \mu R(t) \tag{10}$$

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = \kappa\rho Q(t) \tag{11}$$

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = \alpha S(t) - \sigma\beta V(t)I(t) - \mu V(t) \tag{12}$$

The important constant among all of these is σ which signifies the *vaccine inefficacy*. $\sigma \in (0, 1)$. Therefore $(1 - \sigma)$ gives us the *vaccine efficacy/protection*. Besides the Extended SEIR model [7] also proposes a *Data Assimilation* Technique - *The Ensemble Kanman Filter (EnKF)*. Data assimilation is used to sequentially update the model as and when data becomes available to keep the model on the right track. This helps mitigate uncertainties.

Apart from the *SIRDV compartment model*, Usherwood et al [6] have also proposed to incorporate *behavioral aspects* to the transmission of COVID-19. The proposed equation for transmission rate β is

$$\beta. = \beta_0 f_1 f_v \tag{13}$$

where

$$f_1 = e^{-d_1 I} \tag{14}$$

$$f_v = \frac{1}{f_I} + \left(1 - \frac{1}{f_t}\right) e^{-d_v V} \tag{15}$$

d_1 is the *caution factor* will quantify governmental measures, precautionary measures by people to reduce transmission and d_v is the *sense of safety factor* which will quantify lockdown fatigue. Together this will lead to changing behavioral patterns during a pandemic to be incorporated into a predictive model, especially with sense of safety seeming to increase in most nations with a fast vaccination rollout.

B. Predictive Model Implementations

The second major objective of this paper was to explore different prediction techniques and choose the best one in our case. There is existing literature that have tried to compare different predictive models, from *Regression* to *Time Series Analysis* and *Deep Learning Techniques*. Tomar et al. [8] have described two of the predictive models they chose namely - *LSTM* and *Classical Curve Fitting* and their results on the specific case of the COVID-19 scenario in India in 2020. They chose LSTM over clasical RNN due to the *vanishing/exploding gradient problem*, even though RNNs perform well at time-series predictions. The input, forget, update/control, output gates in the LSTM DNN are given by the popularly known equations

Classical Curve Fitting was done as per *Power Law* ($f(x) = ax^b$). The LSTM captured smaller variations while the Curve Fitting model ignored them and assumed an ideal exponential curve. The pitfall of the paper is that it doesn't provide a performance metric to compare the models, instead only presents the results of each model separately.

This is where some of the more recent works improve. The paper by Kim [9], presents two *statistical predictors* for time-series data, namely - *Auto-Regressive Integrated Moving Average (ARIMA)* and *Generalized Auto-Regressive Conditional Heteroschedasticity (GARCH)* and compares them with a variation of LSTM DNN.
In the Auto-Regressive models, the daily cases is denoted as a process X_t.

X_t is ARIMA(p,q) if

$$X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t + \sum_{j=1}^{q} \beta_j \epsilon_{t-j} \tag{16}$$

X_t is GARCH(p,q) if $X_t = \sigma_t \epsilon_t$ and

$$\sigma_t^2 = \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \beta_0 + \sum_{j=1}^{q} \beta_j X_{t-j}^2 \tag{17}$$

The pesudo-codes to then predict X_t in ARIMA and GARCH is also given. The Deep Learning (DL) model that ARIMA and GARCH are compared with is the *Stacked LSTM DNN*, which is very similar to the LST DNN described above, but instead of only 1 LSTM cell, multiple LSTM cells are stacked/chained together in the network to improve performance. The authors have compared the three models using 2 performance measures - *Mean Absolute Error (MAE)* and *Root Mean Square Error (RMSE)*. The results conclude that the Stacked LSTM DNN performed upto 90% better than the 2 statistical models. The authors have also suggested that Generative Adversarial Networks (GANs) could be explored for prediction.

III. Dataset

We have used two data sources and a total of four datasets in our experiment. A summary of these datasets is given in Table I. These datasets contain time series values from January 22nd 2020 to September 12th 2021.

The first and second datasets correspond to the daily state-wise COVID-19 cases for India and United States of America (USA) accordingly. Both of these datasets has been taken from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins [10]. The dataset for USA roughly contains twice as many attributes as that of India but both contain a significant amount of missing and duplicate values. In order to reduce redundancy, certain repeated attributes have been merged together.

The third dataset is the daily state-wise vaccinations for India. This dataset has been taken from the COVID-19 India API .

The fourth dataset contains the daily state-wise vaccinations for USA. This dataset has been taken from the Our World in Data (OWID) COVID-19 repository [11].

IV. EDA and Visualizations

The data used for the study has undergone several steps of sourcing, cleaning, preprocessing and visualisation.

TABLE I
Dataset used in the Experiments.

Dataset	m	n	Source
Cases for India	18214	22	Johns Hopkins
Cases for USA	31860	23	Johns Hopkins
Vaccinations for India	9990	24	COVID-19 India API
Vaccinations for USA	17501	14	OWID

A. Data Sourcing

The data were sourced from the four sources in Table I. The data for state-wise daily cases in India and USA from the *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University* [10] had to be merged from multiple files (each file for one date) from 01-Jan-2020 to 12-Oct-2021. Additionally, the data for India had to be filtered out from the daily world cases containing data for all non-USA countries, as there was no separate repository for Indian daily cases.

The data for daily state-wise vaccinations in India were obtained from COVID19-India API [12]. The data from this source were already mostly cleaned, apart from multiple NaN values from future dates of the current month (13-Oct-2021 to 31-Oct-2021). These rows were simply dropped.

The data for daily state-wise vaccinations in USA were obtained from *Data on COVID-19 (coronavirus) vaccinations by Our World in Data* [11].

B. Data Cleaning

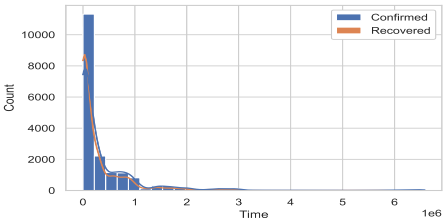
The data were cleaned by following the steps below:

- Summary statistics calculated used to detect any outliers.
- The fraction missing values in each column is calculated ($null\_ratio$). If the $null\_ratio \approx 1$, the column is dropped.
- Columns with different names representing the same data are identified and merged. These columns are typically named inconsistently but represent the same attribute.
- Missing values are either ignored or interpolated, depending on the importance of the attribute.

C. Data Preprocessing

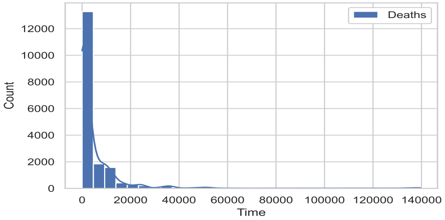
The cleaned daily state-wise data for both USA and India are then preprocessed in two steps to make the discrete data smoother and continuous:

1) The rolling weekly (7-day) averages of daily `Confirmed`, `Recovered` and `Deaths` as well as daily vaccinations were calculated to smooth out the jagged curve.
2) The values are then interpolated using a continuum cubic spline curve, as done by Webb [5], which is a twice-differentiable continuous function.



,,

Fig. 1. Histogram of Confirmed and Recovered cases in India



,,

Fig. 2. Histogram of Deaths in India

D. Data Visualisation

The following visualisations were plotted for the sourced and cleaned data:

1) Histograms comparing the daily `Confirmed` cases and the daily `Recovered` cases are plotted with a bin size of 30. The plot for India is shown in Fig. 1.
2) Histograms showing the daily `Deaths` are plotted with a bin size of 30. The plot for India is shown in Fig. 2.
3) Bar charts showing the `Confirmed`, `Recovered` and `Deaths`.
4) Bar charts displaying the number of first doses and the number of second doses administered.
5) Line plots showing the cumulative `Confirmed`, `Recovered` and `Deaths` numbers (Fig. 3).
6) Line plots showing the daily `Confirmed`, `Recovered` and `Deaths` numbers (Fig. 4).

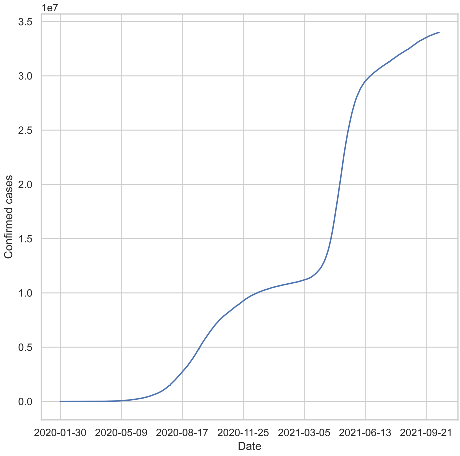
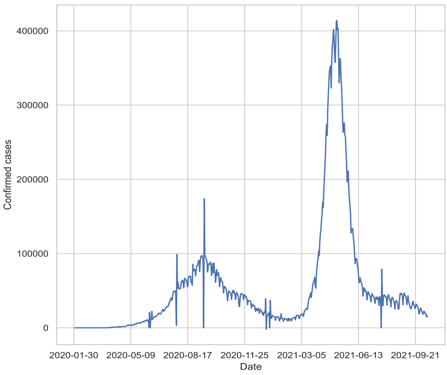
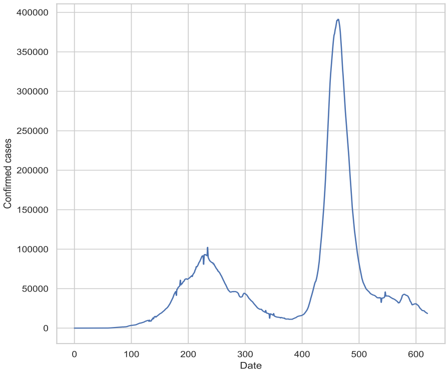


Fig. 3. Cumulative confirmed cases in India



,,

Fig. 4. Daily confirmed cases in India



,,

Fig. 5. Rolling weekly average of daily confirmed cases in India

7) Line plots showing the daily weekly rolling average of `Confirmed`, `Recovered` and `Deaths` numbers (Fig. 5).
8) Line plots showing the first derivative of the daily weekly rolling average of `Confirmed`, `Recovered` and `Deaths` numbers (using cubic spline curve interpolation or `CS` interpolation). The plot for the first derivative of the `CS` curve for daily confirmed cases in India is shown in Fig. 6.
9) Correlation matrix of attributes of daily time-series case data for India and USA. The matrix for India is shown in Fig. 7.
10) Correlation matrix of attributes of daily time-series vaccination data for India and USA. The matrix for India is shown in Fig. 8.

V. Inferences and Conclusions

The following observations were made upon cleaning and visualising the data:

1) From the data sources, we observe that multiple columns are either duplicated or named inconsistently, which is solved by merging the columns.
2) We observe that multiple columns have a $null\_ratio$ close to 1, meaning that the attributes add very little value to the dataset and can be dropped.

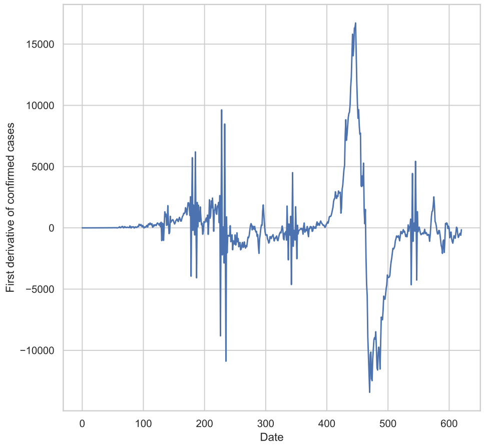


Fig. 6. First derivative of CS curve of weekly average of daily confirmed cases in India

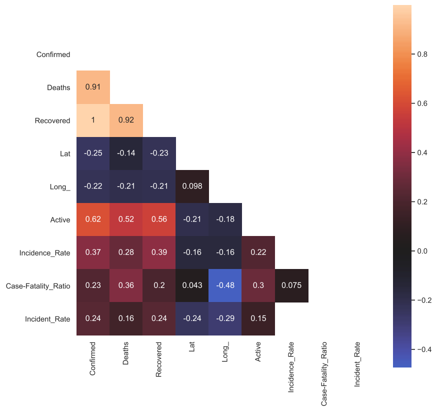


Fig. 7. Correlation matrix for case data in India

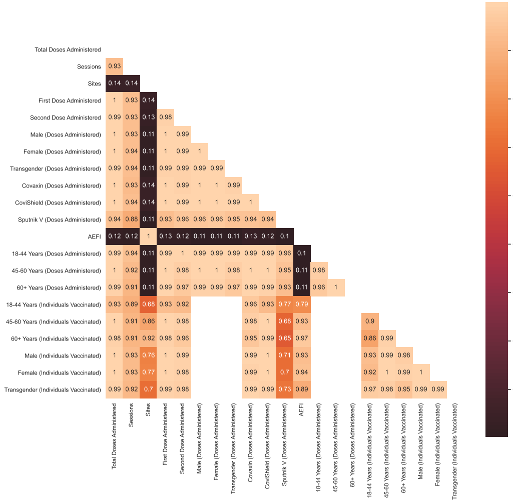


Fig. 8. Correlation matrix for vaccine data in India

3) From the line plots showing the seven-day rolling average, we see that the rough spikes of daily cases are smoothed out.
4) The correlation matrix for case data in India shows that the attributes `Confirmed`, `Recovered` and `Deaths` are strongly positively correlated, which is as expected. There is almost no correlation between any other attributes.
5) The correlation matrix for vaccine data in India shows that almost all the attributes are strongly positively correlated, which is again as expected, as most of the attributes are describing subsets of the same metric (number of vaccine doses administered in a day)
6) Since the data being studied is time-series data, using PCA to visualise the data does not assist in a significant way and was hence not performed.

In order to best proceed with the study, we must use a combination of the various mathematical and predictive models to deliver an accurate prediction of the pandemic situation post-vaccination. We plan to use the Extended SEIR model and compare LSTM, Time-Series methods (ARIMA) and GANs for prediction.

References

[1] F. P. Havers *et al.*, “Covid-19-associated hospitalizations among vaccinated and unvaccinated adults >=18 years – covid-net, 13 states, january 1 – july 24, 2021,” *medRxiv*, 2021. [Online]. Available: https://www.medrxiv.org/content/early/2021/08/29/2021.08.27.21262356

[2] T. Zhou, Q. Liu, Z. Yang, J. Liao, K. Yang, W. Bai, X. Lu, and W. Zhang, “Preliminary prediction of the basic reproduction number of the wuhan novel coronavirus 2019-ncov,” *Journal of Evidence-Based Medicine*, vol. 13, no. 1, pp. 3–7, 2020.

[3] W. O. Kermack and A. G. McKendrick, “A contribution to the mathematical theory of epidemics,” *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, vol. 115, no. 772, pp. 700–721, 1927.

[4] S. Moein, N. Nickaeen, A. Roointan, N. Borhani, Z. Heidary, S. H. Javanmard, J. Ghaisari, and Y. Gheisari, “Inefficiency of sir models in forecasting covid-19 epidemic: a case study of isfahan,” *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021.

[5] G. Webb, “A covid-19 epidemic model predicting the effectiveness of vaccination in the us,” *Infectious Disease Reports*, vol. 13, no. 3, pp. 654–667, 2021. [Online]. Available: https://www.mdpi.com/2036-7449/13/3/62

[6] T. Usherwood, Z. LaJoie, and V. Srivastava, “A model and predictions for covid-19 considering population behavior and vaccination,” *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[7] R. Ghostine, M. Gharamti, S. Hassrouny, and I. Hoteit, “An extended seir model with vaccination for forecasting the covid-19 pandemic in saudi arabia using an ensemble kalman filter,” *Mathematics*, vol. 9, no. 6, p. 636, 2021.

[8] A. Tomar and N. Gupta, “Prediction for the spread of covid-19 in india and effectiveness of preventive measures,” *Science of The Total Environment*, vol. 728, p. 138762, 2020.

[9] M. Kim, “Prediction of covid-19 confirmed cases after vaccination: Based on statistical and deep learning models,” *SciMedicine Journal*, vol. 3, no. 2, pp. 153–165, 2021.

[10] E. Dong, H. Du, and L. Gardner, “An interactive web-based dashboard to track covid-19 in real time,” *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[11] E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, and L. Rodés-Guirao, “A global database of covid-19 vaccinations,” *Nature human behaviour*, pp. 1–7, 2021.

[12] B. H, K. M, A. Shukla, and J. Babu, “Covid19-india api,” https://github.com/covid19india/api, 2020.