2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) | 978-1-7281-8337-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362817

Analysis and Prediction of COVID-19 Pandemic in India

Narayana Darapaneni
Director - AIML
Great Learning/Northwestern University
Illinois, USA
darapaneni@gmail.com

Praphul Jain
Student - AIML
Great Learning
Gurgoan, India
praphuljain.iitd@gmail.com

Rohit Khattar
Student - AIML
Great Learning
Gurgoan, India
rkhattar@gmail.com

Manish Chawla
Student - AIML
Great Learning
Gurgoan, India
manish.k.chawla@gmail.com

Rijy Vaish
Student - AIML
Great Learning
Gurgoan, India
rijuvaish@gmail.com

Anwesh Reddy Paduri
Research Assistant - AIML
Great Learning
Gurgoan, India
anwesh@greatlearning.in

Abstract: **In this paper, we have analysed the COVID-19 progression in India and the three most affected Indian states (viz. Maharashtra, Tamil Nadu and Andhra Pradesh) as of 29-Aug-20 and developed a prediction model to forecast the behaviour of COVID-19 spread in the future months. We used time series data for India and applied the Susceptible-Infective-Removed (SIR) model and the FbProphet model to predict the peak infectives and peak infective date for India and the three most affected states. In this paper, we further performed the comparative analysis of the prediction results from SIR and FbProphet models. From this study, we concluded that with the assumption that a total 5% of India's population might be infected by the pandemic, the countrywide spread is forecasted to reach its peak by the end of Nov-20. And till the time there is no vaccination, for the states that have already reached their peak and with festivals around the corner, there are high chances of resurgence in the number of cases if the social distancing and other control measures are not followed diligently in the coming months.**

Keywords — COVID-19, SARS-CoV-2, Coronavirus, SIR model, FbProphet model, India

I. INTRODUCTION

The Coronavirus disease 2019 (COVID-19) was declared as a global pandemic by World Health Organization (WHO) on 11-Mar-20. [1] The disease has severely impacted globally across 188 countries and territories, with the confirmed cases count reaching 24.7 million. [2] The disease was first detected in India on 30-Jan-20 and currently it has third highest number of confirmed cases globally after the United States and Brazil. As of 29-Aug-20, the total (cumulative) count of confirmed infected population is 3,463,972 across India. [3] A total of 62,550 deaths have been reported while 2,648,998 people have recovered. Active cases stand at 752,424.

In the SIR model, R_0 is the basic reproduction number. It is the expected count of secondary cases produced by a single infection in an entirely susceptible population. R_0 helps in determining if an emerging disease can spread in a population. Bhoomika and Vishesh [4] built the SIR model to predict the endpoint of the pandemic in India and three states - Maharashtra, Gujarat and Delhi. The results show that the endpoint in India will be reached on 12-Sep-20, while endpoints in Delhi, Gujarat and Maharashtra will be reached on 20-Aug-20, 30-Jul-20 and 9-Sep-20 respectively. Liu et al. [5] performed studies and calculated the R_0 of COVID-19 in China and came up with a R_0 value of 4.2 from studies using mathematical modelling, and a R_0 value of 2.67 from studies using statistical methods. The values from both these studies were higher than the one predicted by the WHO (1.95). Zhao et al. [6] estimated the R_0 at 2.1 for the Diamond Princess cruise ship using the Poisson process approach. Tindale et al. [7] estimated 1.97 as the value for R_0 for Singapore and 1.87 for Tianjin, China.

Pandey et al. [8] used the SEIR and regression models for studying the COVID-19 growth in India over a two weeks' period. They observed that the results for the regression model were higher by 15.7% as compared to the SEIR model. Dhanwant and Ramanathan [9] and Sardar et al. [10] predicted the progress of COVID-19 using the SIR model during the first 21-day lockdown, and concluded that only social distancing will not help to contain the spread of COVID-19 if the lockdown were to be lifted after 21 days. Poonia and Azad [11] analyzed the state wise growth of COVID-19 in India and identified the regions where relaxation of the restrictions might be possible post-April-20. Peipei et al. [12] applied the Logistic model to predict the epidemic trend, and then used the predicted values into FbProphet model to derive the epidemic curve. As per their estimations, the disease will peak in Oct-20 with an infected population of 14.12 million.

II. MATERIALS AND METHODS

We have used two different models

A. The Susceptible-Infective-Removed (SIR) Model

The SIR model divides the total population(N) into three compartments: Susceptibles (S): The potential population having the disease, Infectives(I): The population currently having the disease and could infect the others, and Removed(R): The people who have already caught the disease and either have recovered from it or have died.

Like all mathematical models, here we have to make various assumptions to simplify the real-world phenomenon. First, the epidemic is sufficiently short and does not last that long and the total population(N) remains constant. The births and deaths (excluding the COVID-19 deaths) can be ignored. Second, the rate of increase in the Infectives(I) is proportional to the contact between Susceptible(S) and Infectives(I). And third, the rate of increase in the Removed (R) is proportional to the number of infectives(I)

Equations which govern the model:

New infections arise due to the contact between susceptible and infective population. As per the second assumption, βSI is the rate of occurrence of new infections for some positive constant β. With the occurrence of new infection, the infected individual moves from the susceptible compartment to infective compartment. Therefore,

$$dS/dt = -\beta SI \quad (1)$$

The other scenario is when infective individuals enter the removed compartment. The assumption is that γI is the rate of change of Infectives (I) for some positive constant γ which gives us other differential equations.

$$dI/dt = \beta SI - \gamma I \quad (2)$$
$$dR/dt = \gamma I \quad (3)$$

We need some initial data to solve these sets of differential equations. We assume $S = S_0$ and $I = I_0$ and $R = R_0$ initially.
As per the first assumption $dN/dt = 0$. However, $N = S + I + R$, therefore, $d(S + I + R)/dt = 0$.

Knowing the maximum number of infectives i.e. the peak of active cases will be very helpful in planning how to distribute health resources. Now dividing the equation (2) by equation (1) and rearranging, we get:

$$dI/dS = \gamma/\beta S - 1 = (1/qS) - 1 \quad (6)$$

Solving this differential equation, we get:

$$I + S - (1/q)lnS = I_0 + S_0 - (1/q)lnS_0 \quad (7)$$

Infectives(I) will be maximum when $dI/dS = 0$ i.e. when $S = 1/q$. Substituting $S = 1/q$ and rearranging,

$$I_{max} = I_0 + S_0 - (1/q)(1 + ln(qS_0)) \quad (8)$$

The parameters β and γ are estimated by minimizing the difference between actual and predicted values of C. $C \in (S, I, R)$

$$\min||C_t - C_t^*(\beta, \gamma, So)||^2 \quad (9)$$

where $C_t = (C_1, C_2 \dots \dots C_n)$ represents the actual values and $C_t^* = C_1^*, C_2^* \dots \dots \dots C_n^*$ represents the predictions calculated by the model at time $t_i, i \epsilon 1 \dots \dots n$ respectively.

B. FbProphet (Facebook Prophet) Model

FbProphet is for time-series forecasting based on an additive model which was made open source by Facebook in 2017. [13], [14] The non-linear trends of the FbProphet are fitted with yearly, weekly, and daily seasonality, plus holiday effects. The ideal FbProphet model not only does future prediction, but also helps in filling the missing values and detecting the anomalies. The FbProphet model can be described by the below equation,

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (10)$$

Here, *g*(*t*) is a function used for the analysis of non-periodic time-series changes. s(t) reflects the periodic change, such as the weekly or a yearly periodicity. h(t) refers to the impact of an occasional day(s), such as a holiday. єt refers to the error term due to any unobserved factors not being captured by the model. In this paper, we have only considered the time-series changes which are non-periodic in nature.

Our Analysis

We have considered the time duration from 01-Jun-20 to 30-Jul-20 for training the models. We have done our analysis on the data gathered from the official website of the Ministry of Health and Family Welfare, Government of India. Also, for the sake of our analysis, considering the huge population of our country, we have assumed that not the whole population is likely to be infected with this disease. Therefore, we have considered three scenarios that only - (i) 2%, (ii) 5% and (iii) 10% of the total population is susceptible to this disease.

The SIR model estimates the optimal values of parameters β and γ. It uses minimizing the RMSE (Root Mean Square Error) metric for actual and predicted values of S, I and R as the criterion for optimal values of β and γ. Further, the MAPE (Mean Absolute Percentage Error) metric has been used as a measure of the performance of the model.
The β and γ were estimated for different periods of time. Minimum RMSE (S, I, R) was considered to derive the

Hence, $S + I + R = S_0 + I_0 + R_0$.

SIR model tries to answer some critical questions related to the epidemic.

Question 1: Will the disease spread?

We have an initial number of infective people I_0. But we want to know whether this number will grow. Now since, dS/dt is negative, we expect S to decrease with time. Therefore

$$S \leq S_0$$
$$dI/dt < I(\beta S_0 - \gamma) \quad (4)$$

Now, the disease will spread if,

$$dI/dt > 0 \; ie \; if \; , S0 > \gamma/\beta = 1/q \quad (5)$$

q is called the contact ratio: the fraction of the population that comes in a contact with an infective individual during the period when they are infectious. As introduced earlier in this paper, we define here the parameter $R = \beta S_0/\gamma$. R_0 is called basic reproductive ratio and it represents the number of secondary infections in the population caused by one initial primary infection. So, the disease will spread if $R > 1$.

Question 2: What is the maximum number of infectives at any day i.e. the peak?

optimum values of β and γ. MAPE was targeted to be kept within 30% for the acceptable β and γ values.

In FbProphet forecasting, we started with the creation of a Prophet instance. Its fit and predict functions were then called for training and prediction respectively. The input t o FbProphet model is always a time-series data with two features: date *ds* and value *y*. Here in our study, *ds* is the date of day, and *y* is the accumulated cases for the Confirmed, Deaths and Cured cases in India, Maharashtra, Tamil Nadu and Andhra Pradesh.

III. RESULTS

In this section, we have presented the results from the SIR and FbProphet model.

TABLE 1 : POPULATION CONSIDERED

Country/States	Population
India	1,380,004,385
Maharashtra	128,466,921
Tamil Nadu	83,704,074
Andhra Pradesh	90,959,737

Both the models are trained on time-series data from 01-Jun-20 to 30-Jul-20 and the prediction starts from *Day 0* i.e. 31-Jul-20

A. SIR Model Prediction Results

TABLE 2: FORECASTING WITH THE SIR MODEL – INDIA

Scenario	2%	5%	10%
N	27,600,088	69,000,219	138,000,439
S_0	27,565,115	68,809,684	137,809,903
R_0	1.51	1.48	1.48
β	0.0924	0.0926	0.0914
γ	0.061	0.0625	0.0616
RMSE	28,200	25,576	24,287
MAPE(S)	0.04%	0.02%	0.01%
MAPE(I)	2.88%	2.52%	2.65%
MAPE(R)	2.51%	2.92%	2.47%
Peak Infectives	1,830,144	4,136,872	8,342,276
Peak Date	21-Oct-20	24-Nov20	20-Dec-20

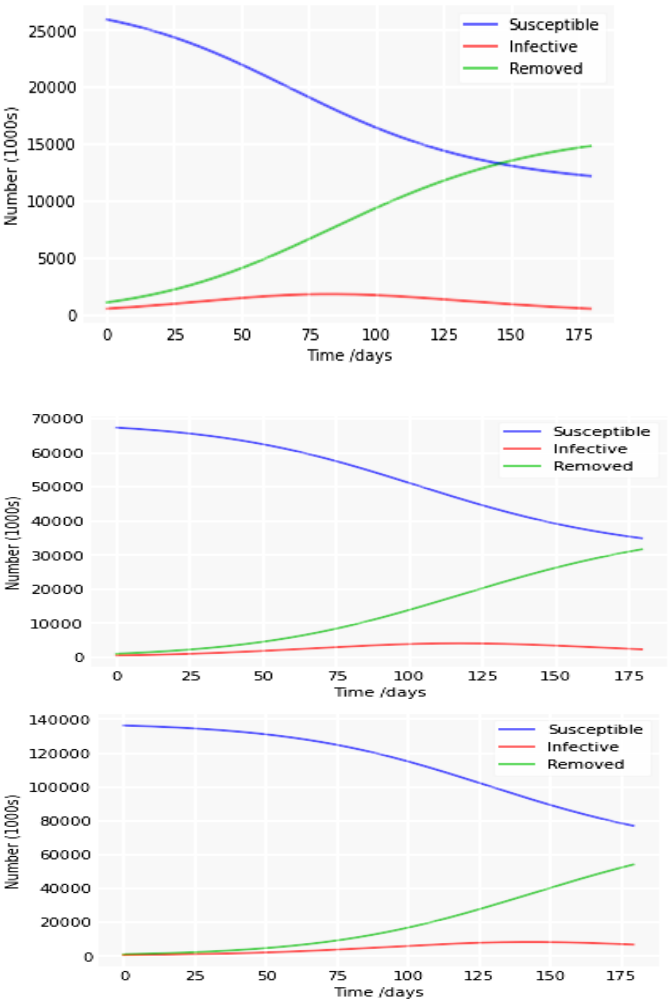

Figure 1: India forecasting with SIR model for 2%, 5% and 10% scenarios

TABLE 3: MAHARASHTRA FORCASTING WITH SIT MODEL

Scenario	2%	5%	10%
N	2,569,338	6,423,346	12,846,692
S_0	2,501,683	6,355,691	12,779,037
R_0	1.66	1.6	1.61
β	0.0726	0.0912	0.0702
γ	0.0427	0.0565	0.0433
RMSE	7324	5701	5281
MAPE(S)	0.13%	0.03%	0.01%
MAPE(I)	3.00%	2.09%	1.89%
MAPE(R)	1.40%	1.39%	1.41%
Peak Infectives	262,501	545,523	1,027,685
Peak Date	26-Sep-20	09-Nov-20	11-Dec-20

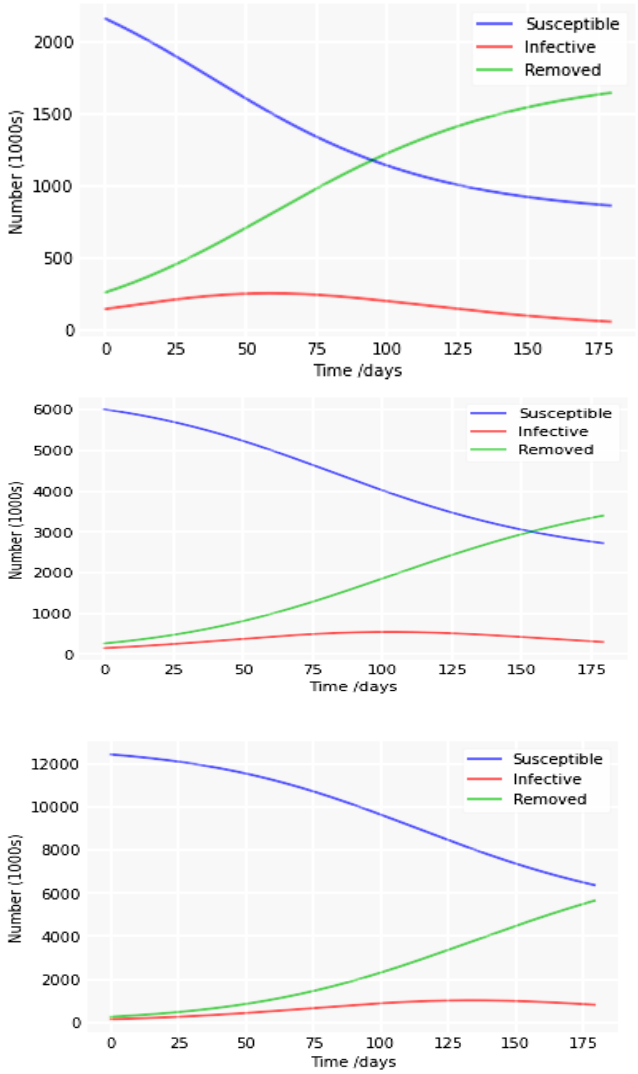

Figure 2: Maharashtra forecasting with SIR model for 2%, 5% and 10% scenarios

TABLE 4: TAMILNADU FORECASTING WITH SIR MODEL

Scenario	2%	5%	10%
N	1,674,081	4,185,204	8,370,407
S_0	1,651,748	4,162,870	8,348,074
R_0	1.47	1.42	1.41
β	0.1244	0.1242	0.1227
γ	0.0837	0.0873	0.0868
RMSE	8022	9240	9754
MAPE(S)	0.10%	0.05%	0.03%
MAPE(I)	10.16%	11.62%	12.09%
MAPE(R)	4.82%	5.09%	4.82%
Peak Infectives	93,871	197,839	388,680
Peak Date	08-Sep-20	10-Oct-20	02-Nov-20

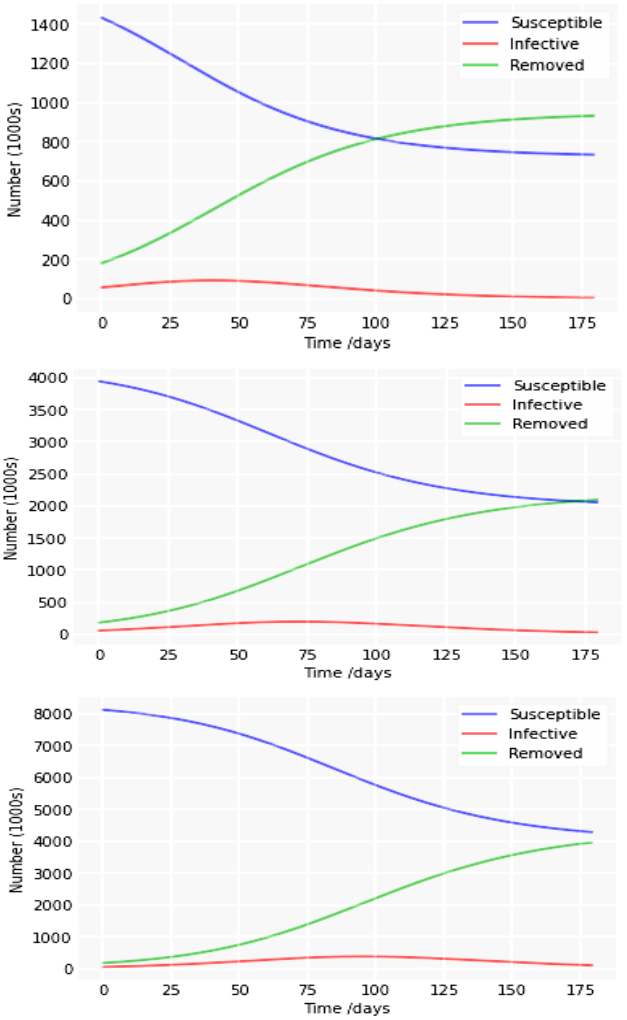

Figure 3: Tamil Nadu forecasting with SIR model for 2%, 5% and 10% scenarios

TABLE 5: ANDHRAPRADESH FORECASTING WITH SIR MODEL

Scenario	2%	5%	10%
N	1,819,195	4,547,987	9,095,974
S_0	1,815,515	4,544,307	9,092,294
R_0	2.07	2.06	2.06
β	0.1254	0.1246	0.1245
γ	0.0604	0.0604	0.0604
RMSE	7068	6717	6614
MAPE (S)	0.16%	0.06%	0.03%
MAPE (I)	9.25%	8.54%	8.57%
MAPE (R)	15.43%	15.01%	14.99%
Peak Infectives	307,155	749,938	149,385
Peak Date	15-Sep-20	01-Oct-20	13-Oct-20

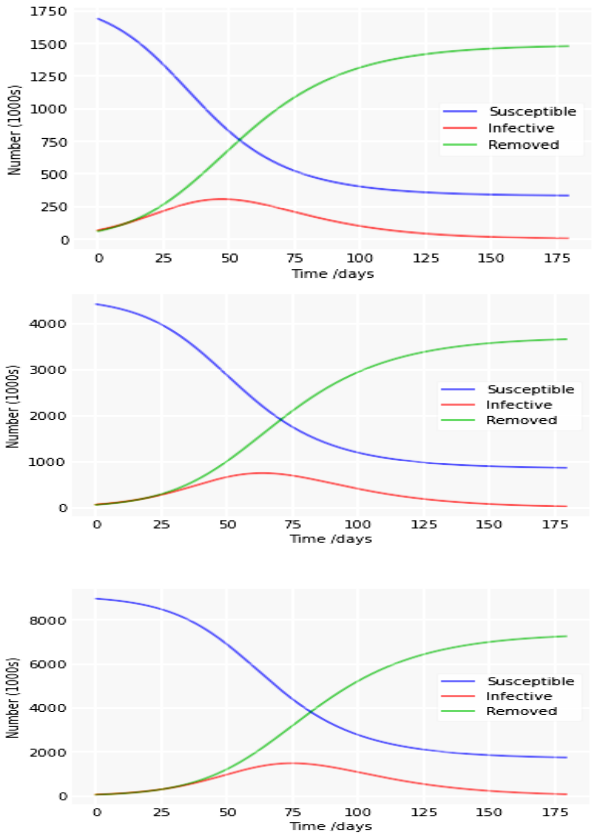

Figure 4: Andhra Pradesh forecasting with SIR model for 2%, 5% and 10% scenarios

B. *FbProphet Model Prediction Results*

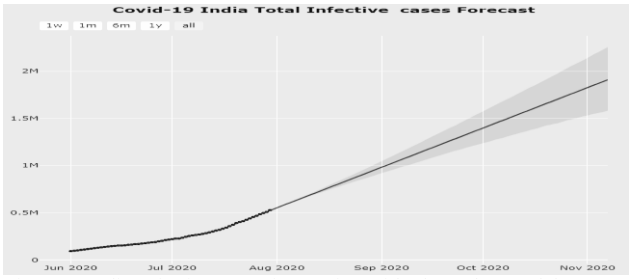

Figure 5: India Infective cases forecasting with FbProphet model

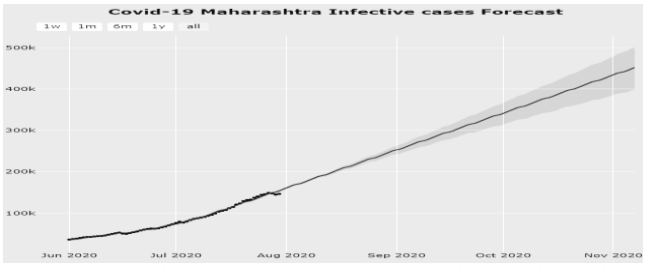

Figure 6: Maharashtra Infective cases forecasting with FbProphet model

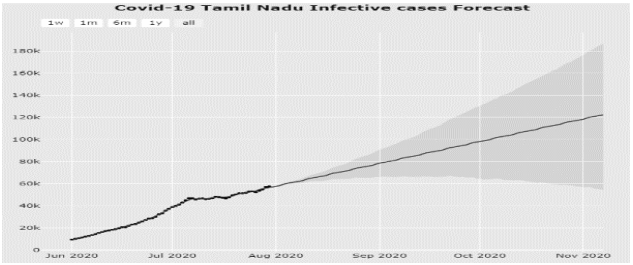

Fig 7: Tamil Nadu Infective cases forecasting with FbProphet Model

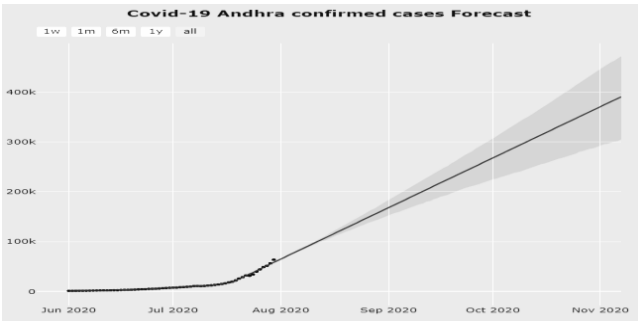

Figure 8: Andhra Infective cases forecasting with FbProphet model

C. *Results Comparison Between SIR and FbProphet Models:*

TABLE 6: SIR AND FBPROPHET MODELS COMPARISION FOR INDIA

Scenario	2%	5%	10%
SIR Peak	21-Oct-20	24-Nov-20	20-Dec-20
Infective (SIR)	1,830,144	4,136,872	8,342,276
Infective_Min (FbP)	1,394,162	1,687,145	2,505,180
Infective (FbP)	1,673,524	2,145,754	2,505,180
Infective_Max(FbP)	1,933,592	2,553,789	3,066,819

TABLE 7: : SIR AND FBPROPHET MODELS COMPARISION FOR MAHARASTRA

Scenario	2%	5%	10%
SIR Peak	26-Sep-20	9-Nov-20	11-Dec-20
Infective (SIR)	262,501	5,45,523	1,027,685
Infective_Min (FbP)	306,227	408,707	477,247
Infective (FbP)	327,390	458,859	552,390
Infective_Max (FbP)	349,080	514,227	635,467

TABLE 8: : SIR AND FBPROPHET MODELS COMPARISION FOR TAMILNADU

Scenario	2%	5%	10%
SIR Peak	8-Sep-20	10-Oct-20	2-Nov-20
Infective (SIR)	93,871	197,839	388,680
Infective_Min (FbP)	66,298	64,344	57,957
Infective (FbP)	83,330	103,807	125,949
Infective_Max(FbP)	99,792	144,152	180,228

TABLE 9: : SIR AND FBPROPHET MODELS COMPARISION FOR ANDHRAPRADESH

Scenario	2%	5%	10%
SIR Peak	15-Sep-20	01-Oct-20	13-Oct-20
Infective (SIR)	307,155	749,938	149,385

Infective_Min (FbP)	187,018	225,130	250,525
Infective (FbP)	214,566	267,760	307,528
Infective_Max (FbP)	249,753	306,752	358,524

IV. DISCUSSION AND CONCLUSION

The SIR model and FbProphet models seem to be giving similar forecast results when we consider the susceptible population as 2% of the overall population which can be seen in Tables 6, 7, 8 and 9. For SIR model prediction, in the best-case scenario (2%), India seems to peak at an active infected population of 1.83 Million, while in the worst-case scenario (10%), it could peak at 8.34 Million. We noticed a significant difference in the forecast from the two models when susceptible population is taken as 5% or 10% of the overall population. This could be due to various reasons like the FbProphet model couldn't catch interactions between the external features, which we think could improve the model's forecasting power. The FbProphet might not be suitable for the time-series with an unstable variance [15]. Moreover, the FbProphet model is still under development and might not be completely stable. [16] On the other side, it should be noted that the SIR model also takes into account some simplifying assumptions about the population. It assumes homogeneous mixing of the population which might not be the case in the real scenario. The SIR model also assumes a constant population with no births, or deaths from other than this disease.

For 5% scenario, India seems to be attaining the peak in Nov-20 while the different states are seen to be attaining the peak at different times. This could possibly be due to the difference in the scale of measures implemented by the respective state governments at different times and if these measures were adhered to by the population as expected in order to control the pandemic. Overall, the peak for India is predicted to be a result of the strict implementation of the measures taken by the central government, like compulsory face mask, social distancing, washing hands frequently, halting public transport completely during the lockdown period, limiting interstate and intrastate movements to only essential travels, etc.

TABLE 10: ACTUAL [3] vs PREDICTED INFECTIONS AS ON 29-AUG-2020

State / Country	Actual	Predicted (2%) SIR	Predicted (5%) SIR	FbProphet
India	752,424	106,1728	117,950	940,137
Maharashtra	185,131	223,803	268,934	244,243
Tamil Nadu	52,726	90,277	123,060	76,037
Andhra Pradesh	97,681	243,508	336,928	157,912

While some of the states (like Delhi) have already attained the peak, other states are expected to attain the peak within next few weeks (in Oct-20), while some other states are expected to be attaining it further in Nov-20 and Dec-20. This could be due to the commencement of the unlock period wherein the movement of people has been allowed. While the pandemic until now was limited to certain areas, it seems to be spreading in the other areas due the movement of the population in these areas. We feel that the magnitude of spread in these areas will primarily depend on the readiness of the governments to tackle the infected population, strict implementation of the control measures and the adherence of the same by the population. Table 10 presents the actual and predicted numbers for infective cases as on 29-Aug-20. The possible reason for the variation in the numbers might lie in the nature and novelty of this pandemic. Patterns have changed quickly in this pandemic. Also, modelling the result with precision becomes difficult as still we know very little about the disease at this point of time. Different aspects such as the modes of transmission, life of virus on different surfaces and virus mutation are still being researched.

REFERENCES

[1] BBC News, "Coronavirus confirmed as pandemic by World Health Organization," *BBC*, 11-Mar-2020.

[2] Wikipedia contributors, "Coronavirus disease 2019," *Wikipedia, The Free Encyclopedia*, 06-Nov-2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Coronavirus_disease_2019&oldid=987294572. [Accessed: 06-Nov-2020].

[3] "MoHFW," *Gov.in*. [Online]. Available: https://www.mohfw.gov.in. [Accessed: 06-Nov-2020].

[4] B. Malhotra and V. Kashyap, "Progression of COVID-19 in Indian states - forecasting endpoints using SIR and Logistic Growth models," *bioRxiv*, p. 2020.05.15.20103028, 2020.

[5] Y. Liu, A. A. Gayle, A. Wilder-Smith, and J. Rocklöv, "The reproductive number of COVID-19 is higher compared to SARS coronavirus," *J. Travel Med.*, vol. 27, no. 2, 2020.

[6] S. Zhao *et al.*, "Epidemic growth and reproduction number for the novel Coronavirus disease (COVID-19) outbreak on the Diamond princess cruise ship from January 20 to February 19, 2020: A preliminary data-driven analysis," *SSRN Electron. J.*, 2020.

[7] L. Tindale *et al.*, "Transmission interval estimates suggest pre-symptomatic spread of COVID-19," *bioRxiv*, p. 2020.03.03.20029983, 2020.

[8] R. Gupta, G. Pandey, P. Chaudhary, and S. K. Pal, "SEIR and Regression Model based COVID-19 outbreak predictions in India," *bioRxiv*, 2020.

[9] J. N. Dhanwant and V. Ramanathan, "Forecasting COVID 19 growth in India using Susceptible-Infected-Recovered (S.I.R) model," *arXiv [q-bio.PE]*, 2020.

[10] S. Tridip, N. Sk Shahid, R. Sourav, and C. Joydev, "Assessment of … lockdown effect in some states and overall India: A predictive mathematical study on COVID-19 outbreak," *arXiv [q-bio.PE]*, 2020.

[11] S. Azad and N. Poonia, "Short-term forecasts of COVID-19 spread across Indian states until 1 may 2020," *Preprints*, 2020.

[12] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in with logistic model and machine learning technics," *Chaos Solitons F* 139, no. 110058, p. 110058, 2020.

[13] S. J. Taylor and B. Letham, "Forecasting at scale," *Am. Stat.*, vol. 72, no. 1, pp. 37–45, 2018.

[14] "Prophet: automatic forecasting procedure, ([EB/OL])," [Online]. Available: https://facebook.github.io/prophet/docs/ or https://github.com/facebook/prophet. [Accessed 11 Aug 2020].

[15] "Topic-9-part-2-time-series-with-facebook-prophet," *Mlcourse.ai*. [Online]. Available: https://mlcourse.ai/articles/topic9-part2-prophet. [Accessed: 06-Nov-2020].

[16] elenapetrova, "Time series analysis and forecasts with prophet," *Kaggle.com*, 19-Jul-2017. [Online]. Available: https://www.kaggle.com/elenapetrova/time-series-analysis-and-forecasts-with-prophet. [Accessed: 06-Nov-2020].