Analysis and Prediction of Covid-19 in India

D. Deepika[1], M. Keerthi Reddy[2]
[1]Assistant Professor, Dept of CSE
[2]Dept of CSE
[1, 2] Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad.

Abstract- *The COVID-19 is an infection of the human lungs with illness, cold, pneumonia being the significant clinical symptoms. The novel coronavirus was discovered at Wuhan in China during December 2019. Around 4.27 million people across the 187 countries are affected, and more than 20 percent of the worldwide population is under lockdown condition as on May 15, 2020.Presently, DNA amplification-based polymerase chain reaction (PCR) is used for identification of COVID-19 with the help of genetic material. However, the efficiency and accuracy of PCR results can be affected by the quantity of viral ribonucleic acid (RNA) that increases the chances of false ineffective results.Early prediction of COVID-19 can be helpful in reducing the huge burden on healthcare systems by helping to diagnose COVID-19 patients.*

The task is to build a model which provide a better identification rate for COVID-19 prediction in the earlier stage and provide greater help emergency of patients in earlier treatment. For training, a model namely the FbProphet (Facebook Prophet) model, which is for time-series forecasting based on an additive modelis used.

The analysis of the COVID-19 progression in India is done by considering the time duration from 30-Jan-20 to 01-Jun-21 and observed that the most affected Indian states (viz. Maharashtra, Karnataka, Tamil Nadu and Kerala) as of 01-Jun-21 and developed a prediction model to forecast the behaviour of COVID-19 spread in the future months. And till the time there is no vaccination, for the states that have already reached their peak and 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, FbProphet model, India, Time series.

I. INTRODUCTION

Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-2) also known Novel Coronavirus or COVID-19 is an infectious and contagious viral respiratory disease that will analyse the magnitude of outbreak in India and future projections of the same [4]. The COVID-19 is an infection of the human lungs with illness, cold, pneumonia being the significant clinical symptoms. The novel coronavirus was discovered at Wuhan in China during December 2019.

The Coronavirus disease 2019 (COVID-19) was declared as a global pandemic by World Health Organization (WHO) on 11-Mar-20. The disease has severely impacted globally across 188 countries and territories, with the confirmed cases count reaching 24.7 million. The disease was first detected in India on 30-Jan-20[3].

1.1 Problem Definition

Future Projection plays an important role in mitigating the burden on the healthcare system, as well as providing the best possible care for patients, efficient diagnosis and information gathering on the prognosis of the disease.The aim is to perform analysis and also provide a better identification rate for COVID-19 prediction in the earlier stage and provide greater help emergency of patients in earlier treatment.

Prediction of covid-19 disease is advantageous to identify patients at a risk of health and can be very useful to overcome the shortage of availability of doctors and physicians in remote places.

1.2 Existing System

Machine learning is commonly being used in every field. To manage forecasting challenges, many prediction approaches are widely used. For example, a comparative study of five machine learning standard models like Linear regression (LR), decision tree, least absolute shrinkage and selection operator (LASSO), random forest and support vector machine (SVM) are used to forecast the threatening variables of COVID-19.

Another model has considered ARIMA and SARIMA models, and identifies the suitability of the model for prediction purposes. The limitation of this model is that it can be handy in calculating linear relationships which are not affected by multiple factors but there are multiple factors that can affect the spreading of COVID.

1.3 Proposed System

The proposed model is based on a FB Prophet model, which is for time-series forecasting based on an additive model which was made open source by Facebook in 2017.The dataset contains date, state names, number of cured, deaths and confirmed cases. By training and fitting the model, will try to predict the covid-19 cases in India for the next 60 days.

II. LITERATURE SURVEY

Potential research work carried out for analysis and prediction of covid19 using different techniques and has been discussed in the following paragraphs. Various researchers have employed different mechanisms for predicting the covid19 cases in India.

[1] Narayana Darapaneni, Praphul Jain, Manish Chawla,Anwesh Reddy Paduri published a paper titled "Analysis and Prediction of COVID-19 Pandemic in India,2020".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. In this paper, further performed the comparative analysis of the prediction results from SIR and FbProphet modelsandconcluded 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.

[2] Yasin Khan, Pritam Khan, Sudhir Kumar Jawar Singh, Rajesh M. Hegde published a paper titled "Detection and Spread Prediction of COVID-19 from Chest X-ray Images using Convolutional Neural Network-Gaussian Mixture Model". The existence of novel coronavirus disease 2019 (COVID-19) is diagnosed using the chest X-ray images of patients and additionally, for the prediction of COVID-19 pandemic, linear regression of the components of Gaussian mixture model (GMM). Using the chest X-ray pneumonia dataset from Kaggle and the University of Montreal, will obtain training and testing classification accuracies of 100% and 96.66% respectively using our CNN model. Further, will obtain the linear regression equations that predict the COVID-19 spread from the GMM.

[3] Siddharth Singh, Piyush Raj, Raman Kumar, Rishu Chaujar published a paper titled "Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models".In this paper, the difference between number of actual reported confirmed cases and approximate number of actual cases, due to insufficient number of tests being conducted, is highlighted based on a unique approximate mathematical formula, thereby establishing relationship between Death Count due to disease and number of people infected with it.Further, utilizing ICMR's available data about COVID-19 patients in India and employing an Enhanced Version of SIR Epidemic Model also known as SIRD devised by generating optimal parameter values and taking number of deaths due to pandemic into account, the time dependence of Outbreak's Intensity in India forecasting maximum number of confirmed active cases of COVID-19 present in a day (Peak Value) and also predicting total number of deaths in India due to the outbreak.

[4] Yifan Yang, Wenwu Yu, Duxin Chen published a paper titled "Prediction of COVID-19 spread via LSTM and the deterministic SEIR model".This paper deals with the Long Short Term Memory algorithm at first to predict the infected population in China. However, it does not explain the dynamics of diffusion process, and the long-term prediction error is too large. Therefore, the widely-accepted SEIR model is introduced to capture the spread process of COVID-19. By using a sliding window method, which suggest that the parameter estimation and the prediction of the infected populations are well performed. This may provide some insights for epidemiological studies and understanding of the spread of the current COVID-19.

[5] R.G Babukarthik, V. Ananth Krishna Adiga, G. Sambasivam, D. Chandramohan, J. Amudhavel published a paper titled "Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN)". In this paper, the state-of-the-art techniques used is Genetic Deep Learning Convolutional Neural Network (GDCNN). It is trained from the scratch for extracting features for classifying them between COVID-19 and normal images.A dataset consisting of more than 5000 CXR image samples is used for classifying pneumonia, normal and other pneumonia diseases. Training a GDCNN from scratch proves that, the proposed method performs better compared to other transfer learning techniques. Classification accuracy of 98.84%, the precision of 93%, the sensitivity of 100%, and specificity of 97.0% in COVID-19 prediction is achieved.

III. DESIGN METHODOLOGY

The proposed model for analysis and prediction of covid-19 in India is based on a FB Prophet model, which is for time-series forecasting based on an additive model which was made open source by Facebook in 2017.The dataset contains date, state names, number of cured, deaths and confirmed cases. By training and fitting the model, will try to predict the covid-19 cases in India for the next 60 days. The overall scope of work to forecast the cases may look like the following:

- Data Acquisition
- Data Pre-processing
- Analysis and Visualization
- Training Model
- Future Prediction

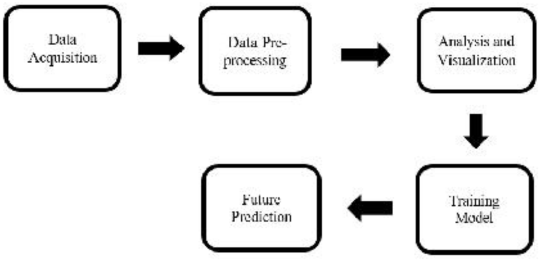
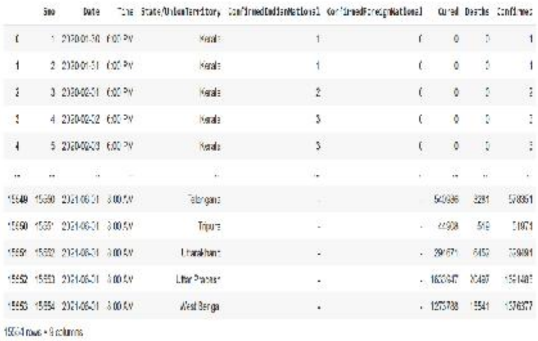


Figure 1 Block diagram of Proposed work

3.1 Data Acquisition

Data acquisition is the step where the dataset and the required libraries are imported. Pandas and NumPy are the libraries which are used for data manipulation and processing.

The required data for developing the prediction model is time-series based. It consists of 15554 rows × 9 columns from January 30,2020 to June 01,2021.



	Sno	Date	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
0	1	2020-01-30	6:00 PM	Kerala	1	0	0	0	1
1	2	2020-01-31	6:00 PM	Kerala	1	0	0	0	1
2	3	2020-02-01	6:00 PM	Kerala	2	0	0	0	2
3	4	2020-02-02	6:00 PM	Kerala	3	0	0	0	3
4	5	2020-02-03	6:00 PM	Kerala	3	0	0	0	3
...	...	...	...	...	...	...	...	...	...
15549	15550	2021-06-01	8:00 AM	Telangana	-	-	540986	3281	578351
15550	15551	2021-06-01	8:00 AM	Tripura	-	-	44908	519	51971
15551	15552	2021-06-01	8:00 AM	Uttarakhand	-	-	294671	6452	329491
15552	15553	2021-06-01	8:00 AM	Uttar Pradesh	-	-	1630647	20487	1691485
15553	15554	2021-06-01	8:00 AM	West Bengal	-	-	1273788	15541	1376377

15554 rows × 9 columns

Figure 2 Snapshot of the dataset being used

3.2 Data Pre-processing

In data pre-processing step, the data is cleaned i.e, the data is made free from missing values, outliers, etc, and the data is changed according to the user, by changing the datatype, by renaming the column names etc,.

3.3 Analysis and Visualization

The data analysis and visualization is done and is observed that the most affected Indian states (viz. Maharashtra, Karnataka, Tamil Nadu and Kerala) as of 01-June-21.

3.4 Training Model

The model is trained using FbProphet model. FbProphet is for time-series forecasting which was made open source by Facebook in 2017.Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

Prophet follows the sklearn model API. We create an instance of the Prophet class and then call its fir and predict methods.

The input to Prophet is always a dataframe with two columns: ds and y, The da(datastamp) column should be of a format expected by Pandas, ideally YYYY-MM-DD for date or YYYY-MM-DD HH:MM:SS for a timestamp. The y column must be numeric, and represents the measurement we wish to forecast.

Fitting of the model is done by instantiating a new Prophet object. Any settings to the forecasting procedure are passed into the constructor. Then its fit method is called and pass in the historical dataframe. Fitting of the model takes 1-5 seconds.

Predictions are then made on a dataframe with a column ds containing the dates for which a prediction is to be made. The helper method Prophet.make_future_dataframe is used for a dataframe that extends into the future a specified number of days. By default it will also include the dates from the history.

The predict method will assign each row in future a predicted value which it names yhat. A new object is created, which is a new dataframe that includes a column yhat with the forecast, as well as columns for components and uncertainty intervals.

The plotting is done by calling the Prophet.plot method and passing in your forecast dataframe. If we want to see the forecast components, use the Prophet.plot_components method.

By default, the trend, yearly seasonality, and weekly seasonality of the time series is obtained.An interactive figure of the forecast and components is then created with plotly.

3.5 Future Prediction

Once the model is built and trained, the future prediction plots are obtained for the next 60 days and there might be a slight variation between the actual and the predicted numbers.

IV. TESTING AND RESULTS

The analysis and prediction of covid-19 in India isdone by considering the time duration from 30-Jan-20 to 01-Jun-21, gathered from the official website of the Ministry of Health and Family Welfare, Government of India and following are the results obtained for the analysis and prediction respectively.

4.1 Analysis of COVID-19 in India

By uploading the data file and pre-processing it, further analysis is done and observed that the most affected Indian states (viz. Maharashtra, Karnataka,Tamil Nadu and Kerala) as of 01-Jun-21and the results are obtained as follows:

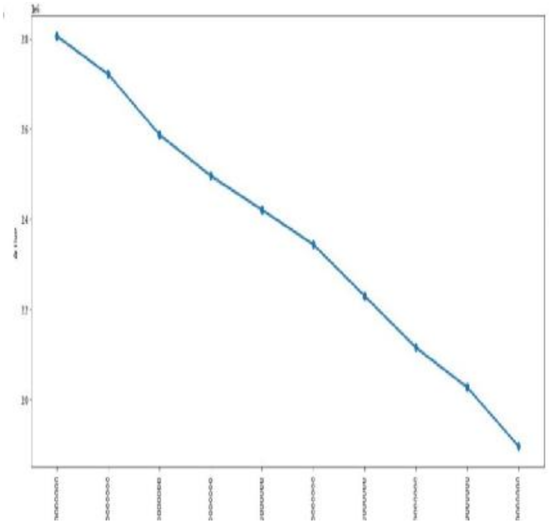


Figure 3 Point plot representing the active cases for the last 10 days

The above point plot is plotted against the Date attribute on x-axis and the sum of active cases on y-axis and it is clear that the number of cured cases decreased in the last 10 days.

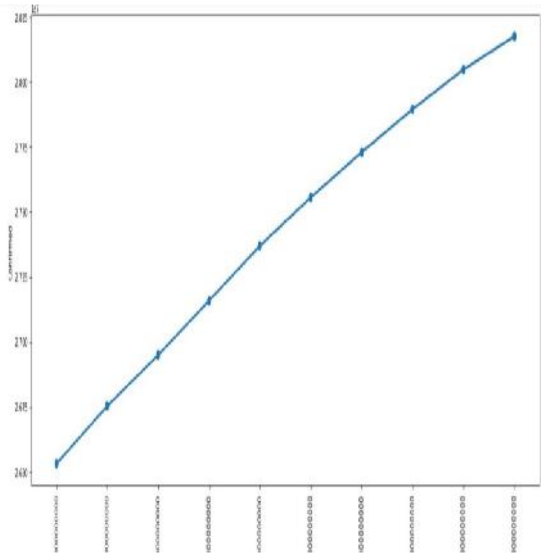


Figure 4 Point plot representing the confirmed cases for the last 10 days

From the above point plotit is clear that the number of confirmed cases increased in the last 10 days, which is plotted against the Date attribute on x-axis and the sum of confirmedcases on y-axis.

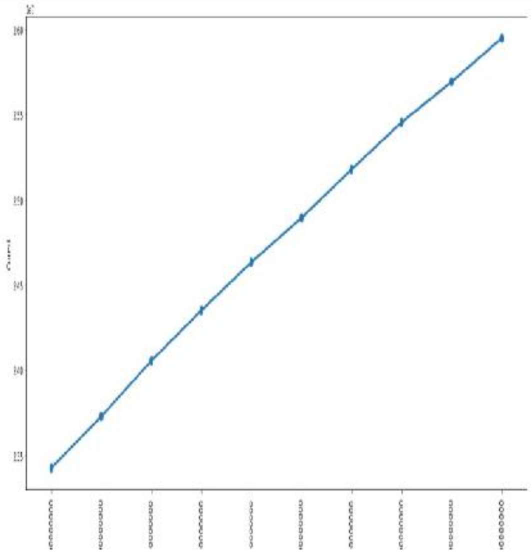


Figure 5 Point graph representing the cured cases for the last 10 days

The above point plot is plotted against the Date attribute on x-axis and the sum of cured cases on y-axis and it is clear that the number of cured cases increased in the last 10 days.

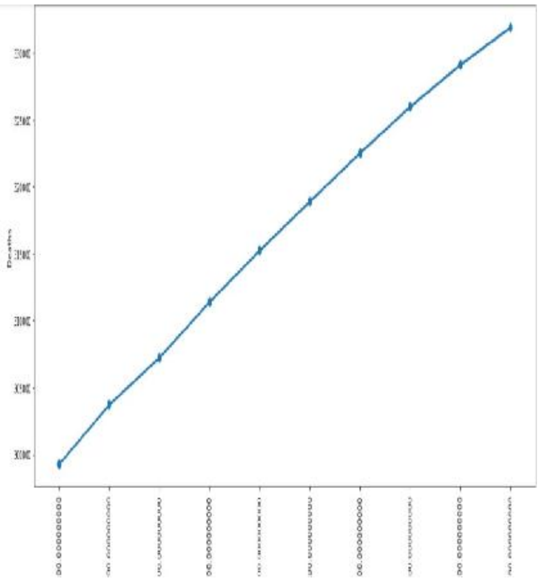


Figure 6 Point plot representing the death cases for the last 10 days

From the above point plotit is clear that the number of death cases increased in the last 10 days, which is plotted against the Date attribute on x-axis and the sum of death cases on y-axis.

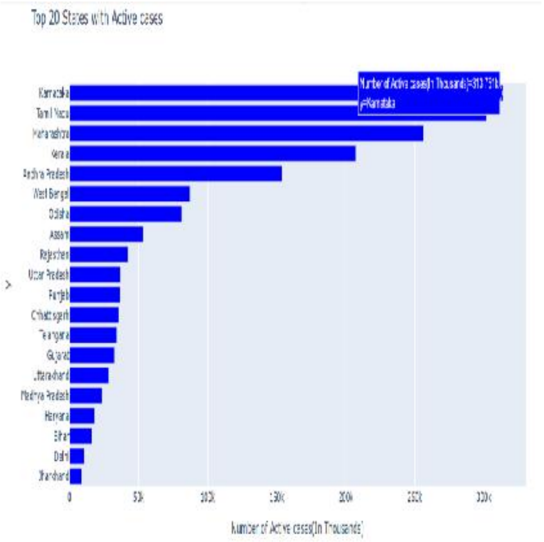


Figure 7 Bar graph representing active cases

From the above bar graph it can be concluded that among the top 20 states, Karnataka is having the highest number of active cases (313.751k) as of 01-Jun-21.

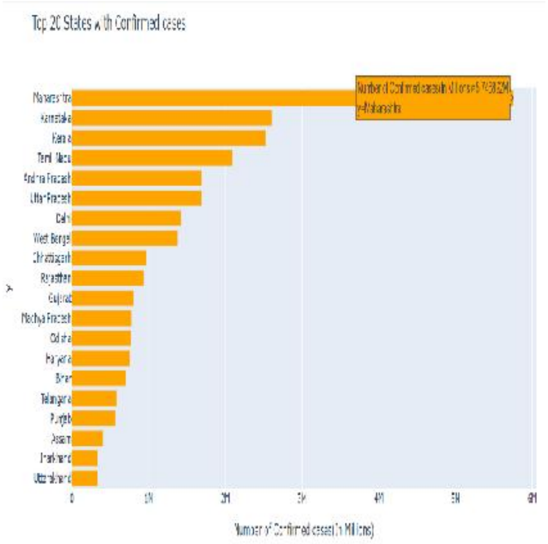


Figure 8 Bar graph representing confirmed cases

From the above graph, it is clear that among the top 20 states, Maharashtra is having the highest number of confirmed cases (5.746892M)as of 01-Jun-21.

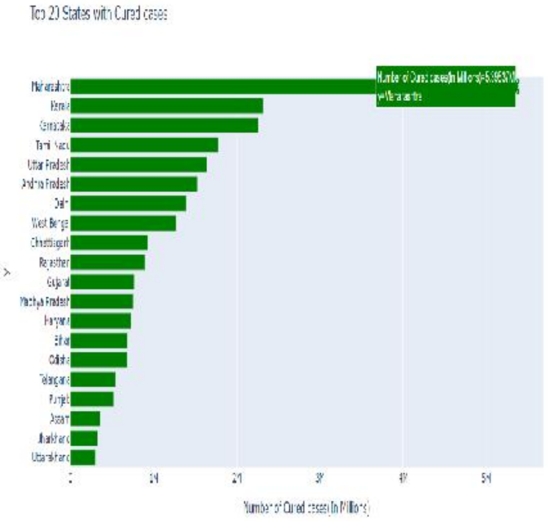


Figure 9 Bar graph representing cured cases

From the above bar graph it can be concluded that among the top 20 states, Maharashtra is having the highest number of cured cases (5.39537M)as of 01-Jun-21.

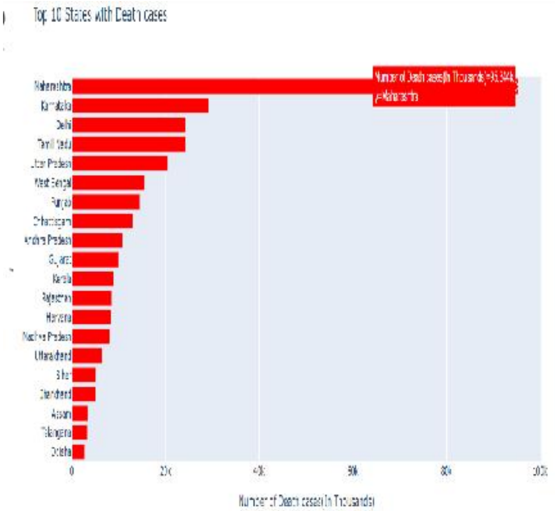


Figure 10 Bar graph representing death cases

From the above graph, it is clear that among the top 20 states, Maharashtra is having the highest number of death cases (95.344k)as of 01-Jun-21.

4.2 Prediction of COVID-19 in India

In FbProphet forecasting, the creation of a Prophet instance is made. Its fit and predict functions were then called for training and prediction respectively. The input to FbProphet model is always a time-series data with two features: date *ds* and value *y*. Here, *ds* is the date of day, and *y* is the accumulated cases for the Confirmed, Deaths and Cured cases in India. Results for the prediction of COVID-19 for next 60 days are shown below:

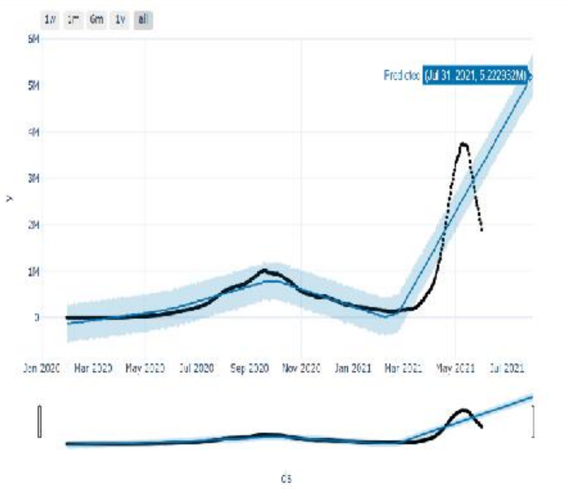


Figure11 Active cases forecasting with FbProphet Model

The abovegraph,shows the prediction of active cases for the next 60 days using FbProphet model. The predicted value for active cases as of 31-July-21 is 5.222932M.

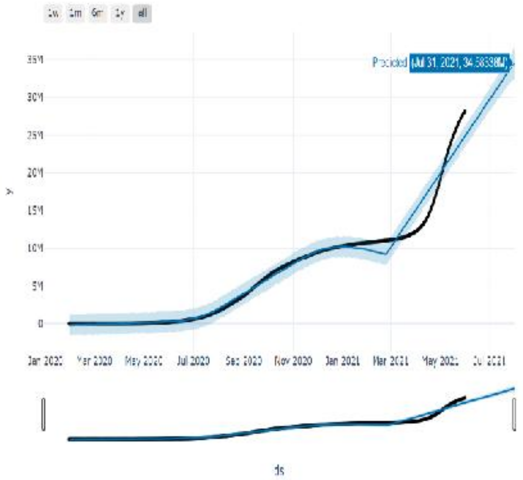


Figure 12Confirmed cases forecasting with FbProphet Model

The above graph, represents the prediction of confirmedcases for the next 60 days using FbProphet model. The predicted value for confirmedcases as of 31-July-21is 34.58338M.

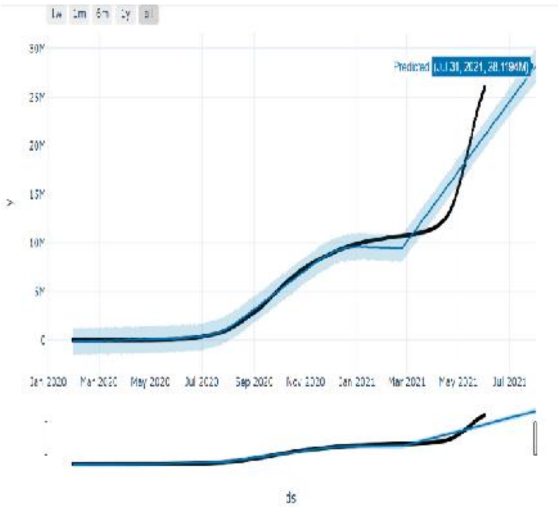


Figure 13 Cured cases forecasting with FbProphet Model

The above graph, shows the prediction of cured cases for the next 60 days using FbProphet model. The predicted value for cured cases as of 31-July-21 is28.1194M.

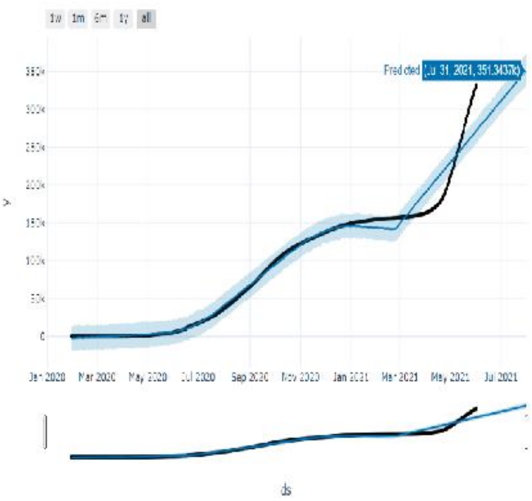


Figure 14Death cases forecasting with FbProphet Model

The above graph, represents the prediction of death cases for the next 60 days using FbProphet model. The predicted value for deathcases as of 31-July-21is 351.3437k.

V. CONCLUSION AND FUTURE WORK

5.1 Conclusion

The analysis of the COVID-19 progression in India is carried out by considering the time duration from 30-Jan-20 to 01-Jun-21 and observed that the most affected Indian states (viz. Maharashtra, Karnataka, Tamil Nadu and Kerala) as of 01-Jun-21 and developed a prediction model to forecast the behaviour of COVID-19 spread in the future months. And till the time there is no vaccination, for the states that have already reached their peak and 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.

There may be a slight variation between the actual and the predicted numbers. The possible reason for the variation in the numbers might be due to the inadequate number of tests and presence of a large number of asymptomatic patients, 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. The magnitude of spread in the 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.

5.2 Future Work

In future, some deep learning methods can be applied for forecasting time series data for getting better predictions.There is a lot of scope for machine learning in healthcare. The future work can be extended on calibrated and ensemble methods that could resolve quirky problems faster with better outcomes than the existing algorithms. Also an AI-based application can be developed using various sensors and features to identify and help diagnose diseases. As healthcare prediction is an essential field for future, A prediction system that could find the possibility of outbreak of novel diseases that could harm mankind through socio-economic and cultural factor consideration can be developed.

REFERENCES

[1] Narayana Darapaneni, Praphul Jain, Manish Chawla, Anwesh Reddy Paduri “Analysis and Prediction of COVID-19 Pandemic in India”, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).

[2] Yasin Khan, Pritam Khan, Sudhir Kumar Jawar Singh, Rajesh M. Hegde“Detection and Spread Prediction of COVID-19 from Chest X-ray Images using Convolutional Neural Network-Gaussian Mixture Model”, 2020 IEEE 17th India Council International Conference (INDICON).

[3] Siddharth Singh, Piyush Raj, Raman Kumar, Rishu Chaujar“Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models”, 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA).

[4] Yifan Yang, Wenwu Yu, Duxin Chen “Prediction of COVID-19 spread via LSTM and the deterministic SEIR model”, 2020 39th Chinese Control Conference (CCC).

[5] R.G Babukarthik, V. Ananth Krishna Adiga, G. Sambasivam, D. Chandramohan, J. Amudhavel“Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN)”, IEEE Access (Volume: 8).

[6] SinaArdabili, Amir Mosavi, Shahab S. Band, Annamaria R. Varkonyi-Koczy“Coronavirus Disease (COVID-19) Global Prediction Using Hybrid Artificial Intelligence Method of ANN Trained with Grey Wolf Optimizer”, 2020 IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE).

[7] Yuankang Zhao, Yi He, Xiaosong Zhao “COVID-19 Outbreak Prediction Based on SEIQR Model”, 2020 39th Chinese Control Conference (CCC).

[8] AlavikunhuPanthakkan, S.M. Anzar, Saeed Al Mansoori, Hussain Al Ahmad “Accurate Prediction of COVID-19 (+) Using AI Deep VGG16 Model”, 2020 3rd

International Conference on Signal Processing and Information Security (ICSPIS).

[9] Vartika Bhadana, Anand Singh Jalal, Pooja Pathak "A Comparative Study of Machine Learning Models for COVID-19 prediction in India", 2020 IEEE 4th Conference on Information & Communication Technology (CICT).

[10] Zhenyu Li, Shentong Yang, Junhong Wu "The Prediction of the Spread of COVID-19 using Regression Models", 2020 International Conference on Public Health and Data Science (ICPHDS).