COVID-19 Outbreak In India: Exploratory Data Analysis And Prediction

Chandrabhan[1], Deepanshu Singh[1], and Anuj Kumar[1]

[1]Department of Software Engineering, Delhi Technological University (formerly DCE) Shahbad Road, Bawana, Delhi-110042
{rishikra5, deepanshusingh223, chaudharyanuj963, }@gmail.com

Abstract. As of 30th April 2020, the cumulative number of confirmed and deceased cases for the COVID-19 (Coronavirus) outbreak in India was 33050 and 1074. It is a disturbing scenario, as India will enter stage-3, termed as the community level transmission, of COVID-19, with such a massive population within a few days. In this study, we consider machine learning models on COVID-19 transmission that combines lockdown effect and transmission variation between populations. We have strived to explore the COVID-19 data to visualize the spread of the virus in the country. We dissected the data to extract state-wise insights. Finally, we use advanced forecasting models to simulate and predict the number of confirmed cases and deaths in the next 15 days. Using daily reported cases of COVID-19 from India, we analyze the impact of the two lockdowns in terms of reduction in the number of cases.

Keywords: COVID-19, Infectious Disease, Machine Learning, Lockdown, Forecasting, Pandemic.

1. Introduction

On 31st December 2019, the city of Wuhan in China reported an outbreak of a new strain of coronaviruses (COVID-19) that has since put to death over 238,198 people. Till 30th April 2020, over 3,096,626 infections—spanning 210 countries and territories—have been confirmed by the World Health Organization (WHO) [1]. The WHO has already declared outbreak as a pandemic. Coronaviruses are wrapped non-divided positive-sense RNA infections that belong to the Coronaviridae family and the Nidovirales and are broadly disseminated among people and different warm-blooded animals [2]. The coronavirus, COVID-19 began in the territory of China, with a land accentuation at Wuhan, the capital city of the Hubei region, and has generally spread everywhere throughout the world. Huge numbers of the underlying cases were typically acquainted with the wholesale fish market of Hunan, which additionally listed live creatures. Clinical preliminaries of hospitalized patients found that patients display side effects predictable with viral pneumonia at the beginning of COVID-19, most regularly, fever, hack, pharyngitis, and exhaustion. A few patients showed detailed changes in their ground-glass lungs, typical or lower than normal white lymphocyte platelet tallies and platelet checks, hypoxemia, and unsettled liver and kidney work [2]. Most were said to be topographically identified with the wholesale fish market of Hunan. Extreme episodes happen in USA (3,67,004 cases), Italy (1,32,547 cases), Spain (1,36,675 cases), Germany (103,375 cases), France(98,010), China (81,708 cases) thus numerous nations and the illness keeps on spreading comprehensively [3]. This has been proclaimed a pandemic by the WHO (World Health Organisation). It is the third zoonotic human coronavirus that has emerged in the current century, after the 2002 serious intense respiratory condition coronavirus (SARS-CoV), which spread to 37 nations and the 2012 Middle East respiratory disorder coronavirus (MERS-CoV), which spread to 27 nations [2]. The 2019 pandemic novel coronavirus was first affirmed in India on 30th January 2020, in the province of Kerala [4]. A sum of 4778 acclimated cases, 382 recuperations, and 136 passings in the nation have been accounted for on 06th April 2020 [5]. The Indian government has presented social separation as a precautionary measure to keep away from the chance of an enormous scope of populace development that can quicken the spread of the malady. India government executed a 14-hour pre-determined lockdown on 22nd March 2020. Besides, the Prime Minister of India likewise requested an across the country 21-day lockdown at midnight on 24th March to slow the spread of COVID-19, influencing India's whole 1.3 billion population.

Notwithstanding no immunization, social separating has distinguished as the most usually utilized anticipation and control methodology. The reason for these activities is the limitation of social connection in work environments, schools, and other open circles, aside from essential open administrations, for example, fire, police, and medical clinics. Presumably, the spread of this infection episode has genuinely disturbed the life, economy, and wellbeing of residents [6]. This is an incredible worry for everybody, to what extent this situation will last, and when will the sickness be controlled. Numerical studies dependent on the arrangement of differential conditions may give a far-reaching component for the elements of COVID-19 transmission. A few ongoing research additionally determined that around 2.68 is the essential conceptive number for COVID-19 [7]. Scientists found that the measure of control generation number might be as high as 6.47 and that techniques for mediation, including serious touch followed by isolation and detachment, would adequately limit COVID cases [7]. For the fundamental regenerative number, scientists revealed an estimation of 3.1 dependent on the information fitting of an SEIR model, utilizing a supposition of Poisson-appropriated day by day time increases [8,23]. A report by Cambridge College has shown that India's countrywide three-week lockdown would not be satisfactory to forestall a resurgence of the new coronavirus pestilence that could come back in months and cause a large number of diseases. They proposed that a few lockdowns can flatten the curve with days in the middle or a solitary 49-day lockdown. Information driven scientific reporting holds a crucial position in disease counteraction, making arrangements for future flare-ups, and deciding the viability of control. A few information-driven tests have been performed in different locales. Right now, there are extremely fewer studies that contemplated the effect of lockdown on COVID-19 transmission elements in India.

2. Data

About Data

- The time-series data is collected from the Indian Health Ministry Website (https://www.mohfw.gov.in/). Each row contains data for each state with at least one confirmed case. There are different columns in the dataset, such as Date, State, Latitude, Longitude, number of Confirmed, Cured, and Death cases. It contains 1464 instances in the dataset spanning from 30-01-2020 to 30-04-2020.

- For comparison with other countries, we have used the time-series data from the website https://github.com/datasets/covid-19 which is maintained by Johns Hopkins University Center for Systems Science and Engineering (CSSE).

3. Model description and analysis

3.1 Prophet Forecasting Model

The Prophet is a strategy for gauging time arrangement information dependent on an added substance model where nonlinear patterns are fit with yearly, week by week, and day by day regularity, in addition to occasional impacts. It works best with time arrangements that have substantial occasional impacts and a few periods of verifiable information. Prophet is hearty to missing information and moves in the pattern, and ordinarily handles exceptions well [9].
Exact and quick: Prophet is utilized in numerous applications across Facebook for delivering robust conjectures for arranging and objective setting.
Tunable conjectures: The Prophet technique incorporates numerous opportunities for clients to change and modify figures. One can utilize human-interpretable parameters to improve their conjecture by including the area information.
Completely programmed: Get a reasonable estimate on muddled information with no manual exertion. The Prophet is vigorous about anomalies, missing information, and emotional changes in the time arrangement.
At its center, the Prophet strategy is an added substance relapse model with four primary segments:

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

g(t) models pattern, which depicts long haul increment or decline in the information. Prophet joins two pattern models, a soaking development model, and a piecewise straight model, contingent upon the kind of determining the issue, s(t) models regularity with Fourier arrangement, which depicts how information is influenced via occasional factors, for example, the season (for example more scans for eggnog throughout the winter occasions), h(t) models the impacts of occasions or huge occasions that sway business time arrangement (for example new item dispatch, Black Friday, Superbowl, and so forth.), ϵ_t speaks to a final error term [10].

The prophet model likewise considers the regularity part of the pandemic, but because of the brief time of the episode, that usefulness is not utilized here for anticipating the spread of the infection. Figure 3 shows the vulnerability levels of the forecast is low after utilizing this model.

3.2 ARIMA Model

ARIMA represents **A**uto**R**egressive **I**ntegrated **M**oving **A**verage.

AR (Autoregression): A model that utilizes the reliant connection between perception and some number of slacked perceptions. p is a parameter of what number of slacked perceptions to be taken in.

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

I (Integrated): A model that utilizes the differencing of crude perceptions (for example, taking away a perception from the past time step). Differencing in measurements is a change applied to time-arrangement information to make it fixed.

$$\begin{aligned} y_t^* &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

MA (Moving Average): A model that utilizes the reliance between a perception and a lingering mistake from a moving normal model applied to slack perceptions. q is a parameter of what number of slacked perceptions to be taken in. Despite the AR model, the limited MA model is consistently fixed.

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Assumptions: ARIMA model depends on various presumptions, for example, data does not contain irregularities, model parameters, and mistake term is consistent, Historic timepoints direct conduct of present timepoints which probably will not hold in focused on advertise information conditions, Time arrangement is fixed [11].

Parameters of the ARIMA model p (slack request) is the number of slack perceptions remembered for the model, d (level of differencing) is the number of times that the crude perceptions are differenced, q (order of the moving average)is the size of the moving normal window.

3.3 Random Forest Regressor

The Random Forest is an algorithm equipped for performing both joining and categorizing undertakings with the use of more than one decision trees and a procedure called “Bootstrap Aggregation", also called “bagging” [12]. This gives more reliable result because now we do not depend on a single decision tree.
We can tune various hyper-parameters to get better output. The random forest model also avoids over-fitting because of grouping characteristic of the model by a subset of features.

3.4 Gradient Boosting Frameworks

Gradient Boosting frameworks significantly increase the functionality of regular decision trees. It produces a model based on a lot of weak models and the results are generalized and combined using a differentiable loss function [13]. Here, we will use two gradient boosting frameworks, namely, XgBoost and Light GBM.
XgBoost represents Extreme Gradient Boosting. It has perhaps quickest execution using decision trees. It is a choice tree-based group Machine Learning calculation that utilizes an angle boosting structure [14]. In prediction scenarios, including unstructured information (pictures, content, etc.) ANN systems will, in general, beat every other model or structure. Be that as it may, with regards to little to-medium organized/plain information, choice tree-based calculations are viewed as the best-in-class at present. Light gradient boosting is also has many plus points, such as its ability for faster training, high accuracy and efficiency, and its capability to handle large datasets. We utilized these algorithms for our predictions.

4. Methodology

The datasets we used, as discussed in Section 2, for the Indian dataset, we grouped the data by dates to get the numbers of all the daily cured, recovered, and deceased cases from 30-01-2020 to 30-04-2020. We used this to visualize the trend in the increase in the number of cases (see figure 1). We also extracted some state-wise insights from the data by grouping the dataset according to the state or union territories (see figure 2). The international data from John Hopkins University (refer section 2) is used to compare the stage of transmission of the virus within India to other countries of the world that are ahead of us in the pandemic (see figure 3). According to the daily surge in the count of cases, the growth factor is calculated and using that, we forecast the number of cases for the next 15 days, i.e., from 01-05-2020 to 15-05-2020. The data was organized to be suitable as per the input requirements of the forecasting models used (refer section 3) and the number of cases was forecasted using these models. Traditional machine learning algorithms like random forest and gradient boosting are used and evaluated according to their error rates and optimal algorithms are suggested. The two lockdowns in India, are studied in three stages (a) the period from the first recorded case (30-01-2020) to the start of the first lockdown (25-03-2020). This period could be used as a benchmark to analyze the effectiveness of lockdown, (b) Lockdown 1.0, started from 25-03-2020, and lasted 21 days and (c) Lockdown 2.0, started from 14-04-2020 and lasted 19 days [20]. In each case, the growth factor from the number of cases are calculated and are compared with the forecasted models used by us. The growth factor is calculated by taking the average of differences in the number of cases on each day with the previous day. This gives us a good idea of the effectiveness of the measures taken by the government and whether or not it should be continued.

5. Data Analysis

In this study, we have attempted to understand the trend of the COVID-19 pandemic within India. We use the time series data (refer section 2) to gather insights that could help in taking measures and allocating resources in activities that can curb the further spread of the COVID-19 virus. We start by visualizing the trend in the increase in the number of cases (see figure 1).

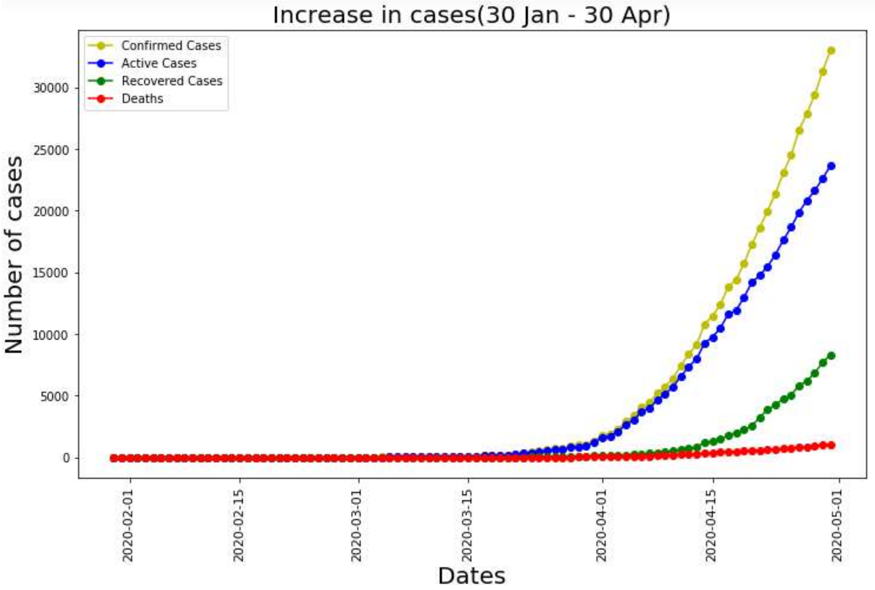

Fig.1. In the figure, according to our data analysis, the blue line represents the confirmed cases of COVID-19 within India, the green line represents the recovered cases, and red line represents the number of deaths due to the pandemic.

The curves in the above figure show an exponential increase in the number of confirmed and active cases from 30-10-2020 to 30-04-2020, which has put a strain on the medical institutions of the country. The recovered cases are gradually increasing too [15], which show some success in combating this pandemic. The death rate of this virus is low; this can be inferred by the near-constant red curve. However, this should not be taken lightly because of the enormous population of India because even a small percentage of the deceased could result in a huge death toll [16]. We also engaged our research in extracting state-wise insights from the dataset.

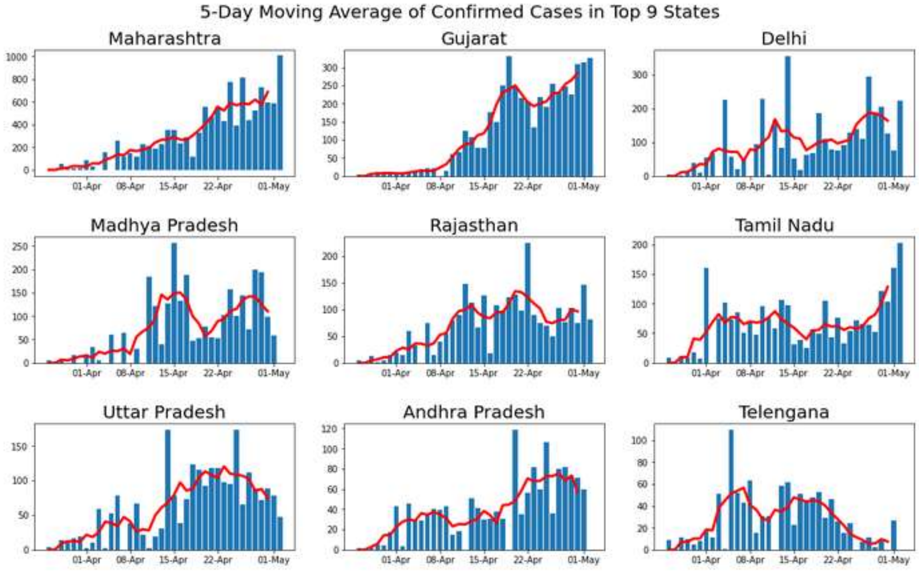

Fig.2. In the figure, the number of cases in the top 9 worst-affected states of India is plotted against the period from 30th January 2020 to 30th April 2020. The red curve shows the moving average, which points to non-stationary data.

We can understand the stage at which each state is by looking at the average curve. From the above figure, we can infer that states like Maharashtra, Delhi, and Gujarat high number of cases as compared to other states and are worst affected by the COVID-19 virus. The different rates of transmission in various states can be a result of various factors such as lockdown policy, population density, and testing strategy of the state. We also used the data from other countries from the John Hopkins University dataset (refer section 2) and grouped it by dates to plot the active, recovered, and deceased cases of each country (see figure 3).

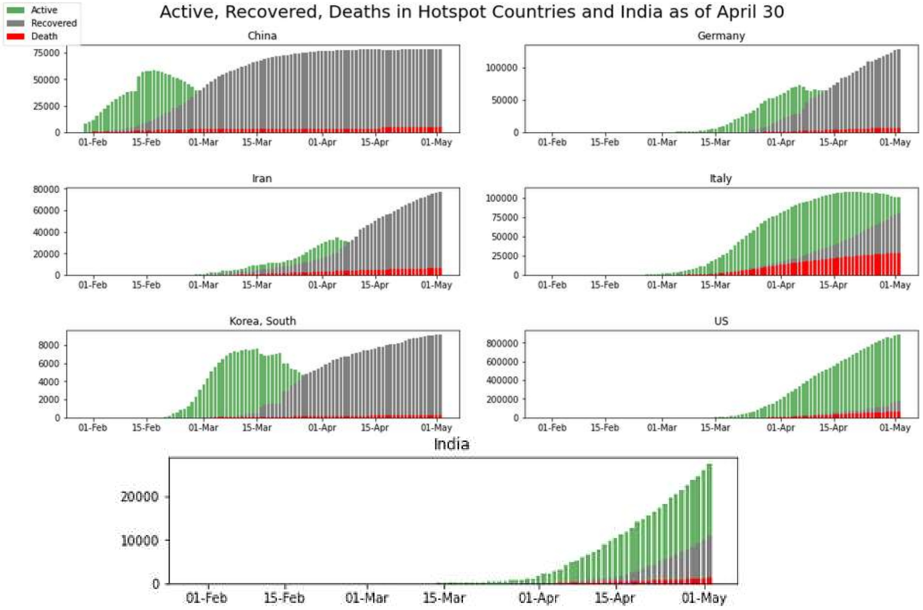

Fig.3. In the figure, we use the data from other countries with active (green), recovered (gray) and deceased (red) cases.

The rate advancement of the COVID-19 pandemic is different in different countries. The above figure suggests that countries like India are in the initial stages of the pandemic. Even with a massive population like India, the number of reported cases are low. This can be attributed to the testing strategies of the country. Another reason for this can be the lockdown measures taken by the government of India which is discussed in section 6.1 of the text.

6. Results and Discussion

The growth factor model fitting for the COVID-19 pandemic to the daily reported cases in India is shown in figure 4.

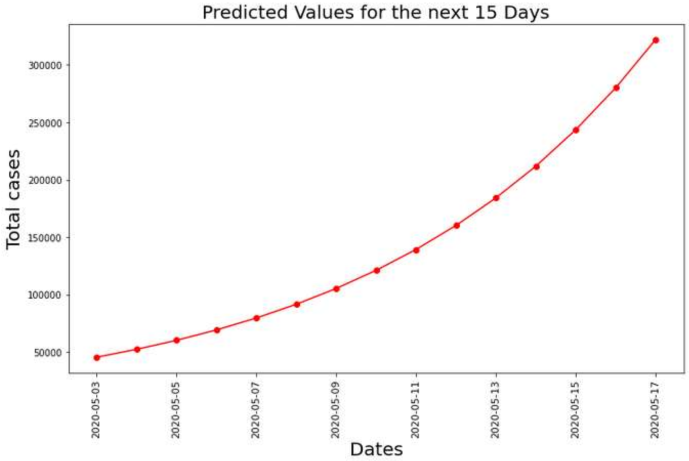

Fig. 4. The plot in the figure forecasts the growth in the number of cases in the next 15 days (starting 01st May 2020), given that the growth factor of the virus remains constant throughout the pandemic.

After analyzing the data, the calculated growth factor turned out to be 1.5625 (which is less than the agreed-upon range of transmission rates, i.e., between 2 and 3; source WHO). Therefore, figure 4 shows the exponential growth curve and the surge in the number of cases if the growth rate remains constant. This assumption of a constant growth rate is not accurate to assume since several preventions and curing measures taken by the government tend to decrease the growth factor [17]. So, we use the Prophet model for forecasting the number of cases because of its ability to incorporate complex behavior of data, ease with time-series data, and produce forecasts with high confidence levels. The yearly seasonality in the model is turned off because of the limited period of the existence of the virus.

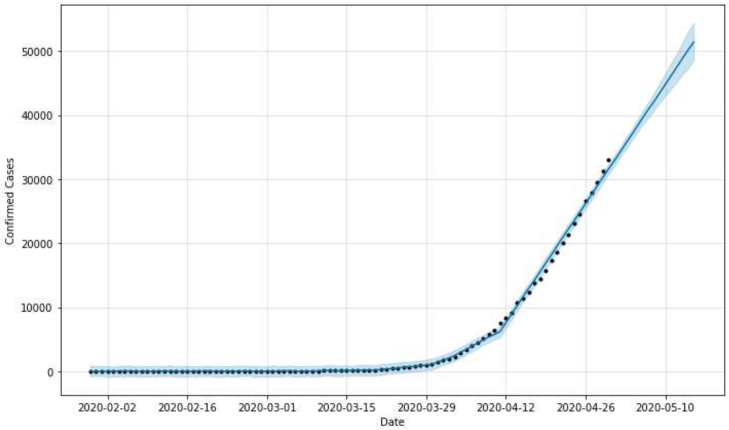

Fig. 5. In the figure, the black dots represent the actual data points; the blue line represents the

predicted exponential growth in the number of confirmed COVID-19 cases; the shaded blue region represents the uncertainty in the forecast.

Therefore, we use Prophet (see figure 5) forecasting model to predict the number of cases for the next 15 days, i.e., from 01-05-2020 to 15-05-2020. The model forecast shows very low uncertainty in the prediction values as can be extrapolated by the plot above. The model predicts with 95% confidence that the number of confirmed cases in the country will surpass 50,000 by 12-05-2020.

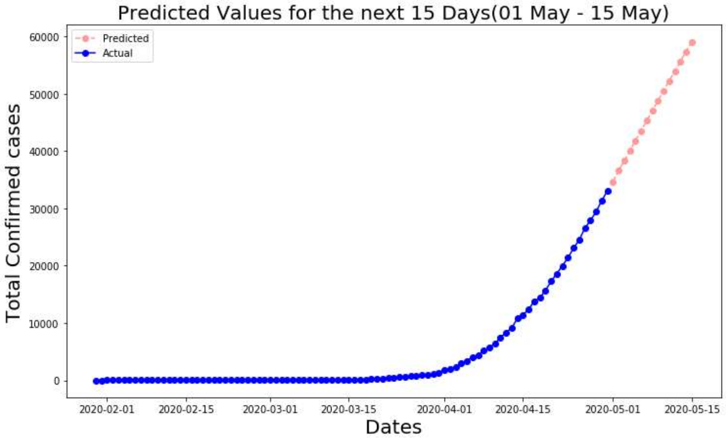

Fig.6. In the figure, the dotted blue curve represents the real data points of the COVID-19 cases in India, and the dotted orange curve represents the predicted number of cases for the next 15 days (starting from 01st May 2020).

ARIMA model is excellent for time-series non-stationary data. Our data has the same characteristics; therefore, we use ARIMA to simulate the number of cases in the country for future dates. We have used the p-value of 5 for the autoregressive model and the degree of difference, d (refer section 3.2) is set to 1. Therefore, the ARIMA (5,1,0) model is used in our case. As shown in the above figure, actual data and predicted curves seem to follow similar trends. We also used some decision tree machine learning regressor models to see how well they work on our data prediction. We used Random Forest Regressor, Light Gradient Boosting Model, and XGBoost Regressor to test their performance. We use RMSE values to show the best-fitted model (see figure 7) out of all three traditional machine learning models.

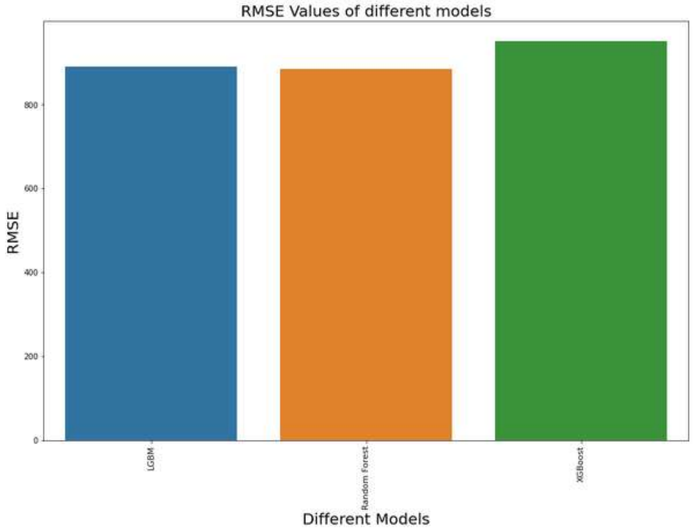

Fig.7. Root Mean Square Error for each Model

The regressors were fit on the training data which was chosen to be from 30-01-2020 (The first case recorded in India) to 12-04-2020. The regressors were tested on the rest of the data, i.e., till 30-04-2020. The above figure gives the root mean square error values of all three models and observe that random forest and light gradient boosting algorithms show similar results. Moreover, the Xgboost algorithm gives a higher error rate in the prediction of the number of cases.

6.1 Effects of Social Distancing

To investigate the lockdown effect in India, we estimate a daily increase of cases in different stages of measures taken by the government of India. Information on growth factors for India under different lockdown scenarios is depicted in Table 1. In the initial days of the pandemic i.e., without the lockdown period, the growth factor was very high as compared to other countries. With the massive population like that of India, this would have been disastrous. But thankfully the government imposed strict police enforced lockdowns all across the country to curb the pandemic. Estimates of growth factors clearly show that it falls dramatically during the first lockdown period (25/03/2020 till 14/04/2020) and in the second lockdown period. The growth factors remain very close to unity during both the lockdown periods.

Table 1. The above table shows the growth factor in each stage of the pandemic in India. *The lockdown 2.0 lasted till 03rd May 2020, but this paper uses data till 30th April 2020.

Stage	Start Date	End Date	Growth Factor
Without Lockdown	30th January 2020	24th March 2020	18.943
Lockdown 1.0	25th March 2020	14th April 2020	1.093
Lockdown 2.0	15th April 2020	30th April 2020*	1.073

This indicates that the number of newly notified cases will decrease after the lockdown periods are over in India. This result further supports the fact that both lockdowns were effective and should be extended further.

7. Conclusion

In this paper, the relevant models are chosen by examining the current data of the Indian pandemic situation, and then the simulation is performed. Moreover, we forecasted the evolution of the current situation and observed that the enforcement of controls would have a major effect on the outbreak. We have explored the situation of Covid-19 spread in India in practical scenarios. The effect of lockdown has been discussed with different stages of lockdown. It should be noted that a sudden increase in infected people may give rise to a severe situation in terms of total cases, which can overwhelm the medical institutions. The studies also show that the effect of lockdown has been significant in dealing with COVID-19. Infectious disease is a social concern in that it can inflict both personal and societal damage on a population. Therefore, numerous studies in different fields are being carried out to minimize the social losses by predicting the spread of infectious diseases. This study aimed to analyze the data of the COVID-19 outbreak and produce some intelligent forecasts to understand the trend of an increase of reported cases. The results of the analysis show that the number of cases will increase in the next 15 days. Therefore the government must extend the lockdown period to reduce the strain on the medical organizations and help in flattening the curve. This research also has a drawback, which is a fairly short duration of data collection. India is at an elevated risk of entering community-level transmission (Stage 3) due to failure in following the quarantine guidelines by citizens, as well as other social and demographic challenges [21]. The results and forecasting of the evolution of the pandemic made using the available data by our models in the current study will not hold in such a case. For the COVID-19 situation to end as soon as possible, every individual needs to be responsible for their safety and others. Finally, the results that we get using our models solely depends on the data. Because of real-time changes in data daily, the predictions could increase or decrease accordingly. Consequently, the results and predictions of this paper should only be used for qualitative interpretation and rational assessment of the complexity of the pandemic. Finally, we hope that this study can make some contributions to India's response to this pandemic and that we succeeded in making some useful insights for the policymakers of the country to use.

Future Scope and Research

The modeling and simulation of epidemics through a population is a complex process. Studies have shown that the infection caused by virus outbreaks closely resembles Gaussian distribution [18, 22]. This mathematical tool can be used to construct an effective model. Deep learning methods like DNN (Deep Neural Networks) and LSTM (Long Short Term Memory) methods can prove very useful in capturing the complex behavior of the pandemics like COVID-19 [19].

Acknowledgments We would like to acknowledge Dr. Rajni Jindal, HOD, Dept. of CSE at Delhi Technological University for her useful inputs in the making of this paper and guiding us through the project.

References

1. "Coronavirus Disease 2019" , https://www.who.int , Retrieved: [30 April, 2020].
2. Mattia Mori, Clemente Capasso, Fabrizio Carta, William a Donald &Claudiu T Supuran (2020). "A deadly spillover: SARS-CoV-2 outbreak", Expert Opinion on Therapeutic Patents, DOI: 10.1080/13543776.2020.1760838
3. "COVID-19 Coronavirus Pandemic",https://www.worldometers.info/coronavirus/, Retrieved : [30 April, 2020].
4. David Reid, Jan 30, 2020, "Coronavirus (COVID-19)" , https://www.cnbc.com, Retrieved : [01 May, 2020]
5. "COVID-19" , AVAILABLE: https://www.mohfw.gov.in/, Retrieved : [30-04-2020]
6. Vaishnavi Chandrashekhar, 30 March, 2020, "Coronavirus Disease 2019" , https://www.sciencemag.org/ , Retrieved : [01-04-2020].
7. Zhai, Pan, et al. "The Epidemiology, Diagnosis and Treatment of COVID-19." International Journal of Antimicrobial Agents, 2020, p. 105955.
8. Hongjun Zhu, "Transmission Dynamics and Control Methodology of COVID-19: a Modeling Study", medRxiv, 2020.03.29.20047118
9. Taylor SJ, Letham B. 2017. "Forecasting at scale". PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2
10. Steven Liu, Dec 21, 2018, "Forecasting with Prophet". [On-line]. Available: https://towardsdatascience.com . Retrieved: [Apr 30, 2020].
11. "ARIMA models for time series forecasting", Available: https://people.duke.edu/~rnau/411arim.htm . Retrieved: [Apr 30, 2020].

12. Krishni , Nov 27, 2018, “Guide to Random Forest Regression”, [On-line]. Available: https://medium.com , Retrieved: [May 01, 2020].
13. Jason Brownlee, 21st August 2019, "A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning", Available: https://machinelearningmastery.com Retrieved : [30th April 2020].
14. Chen, Tianqi & Guestrin, Carlos. (2016). “XGBoost: A Scalable Tree Boosting System”. 785-794. 10.1145/2939672.2939785.
15. “COVID-19 Stats”, Available: https://www.covid19india.org/ .
16. Nick Triggle ,16 April,2020, “Coronavirus: How to understand the death tol”l , Available: https://www.bbc.com. Retrieved : [April 30,2020].
17. Samantha Sault, 21st March 2020," Why lockdowns can halt the spread of COVID-19", Available: https://www.weforum.org.Retrieved : [02nd May 2020].
18. Lixiang Li, Zihang Yang, Zhong kai Dang, Cui Meng, Jingze Huang, Haotian Meng, Deyu Wang, Guanhua Chen, Jiaxuan Zhang, Haipeng Peng, Yiming Shao, "Propagation analysis and prediction of the COVID-19", Infectious Disease Modelling, Volume 5, 2020, Pages 282-292.
19. Chae, Sangwon et al. “Predicting Infectious Disease Using Deep Learning and Big Data.” International journal of environmental research and public health vol. 15,8 1596. 27th July. 2018, doi:10.3390/ijerph15081596.
20. Kirti Pandey, 01 May, 2020, “Lockdown Timeline”, Available: https://www.timesnownews.com Retrieved: [May 02, 2020].
21. Amit Mudgill, 25th March 2020, "How will lockdown play out for India", Available: https://economictimes.indiatimes.com Retrieved: [02nd May 2020].
22. Predictions for COVID-19 Outbreak In India Using Epidemiological Models, Rajesh Ranjan, doi: https://doi.org/10.1101/2020.04.02.20051466
23. M.K., Arti. (2020). Modeling and Predictions for COVID 19 Spread in India. 10.13140/RG.2.2.11427.81444.