EasyChair Preprint

№ 4061

Forecasting the Peak of COVID-19 Daily Cases in India Using Time Series Analysis and Multivariate LSTM

Souvik Sengupta

August 20, 2020

Forecasting the Peak of COVID-19 Daily Cases in India Using Time Series Analysis and Multivariate LSTM

Souvik Sengupta
Aliah University
ssg@aliah.ac.in

Abstract:

Since the start of COVID-19 pandemic, the main question that every country is desperately looking for an answer is when the number of daily new infection cases would decline. At the end of July 2020 (time of writing this paper), many countries have already managed to decrease the number of daily infections as well as total infections and death cases. India imposed four consecutive lockdowns which spanned over 61 days and currently it is being withdrawn phase by phase. However, the number of new cases and number of fatality is still on the rise. This paper investigates what could be the possible time required for India before the numbers of daily infected people could start declining and what could the peak value hit by then. Three different models are used independently for initial predictions using statistical ARIMA model, SAR epidemical model, and ML regression model. These three results are fit into a stacked LSTM model which makes the final prediction. It is forecasted that the number of daily new cases would keep on increasing till first week of Nov 2020 and can reach up to 90 thousands before it finally starts to decline.

Keywords: COVID-19 prediction, LSTM, Epidemic model, Polynomial Regression, ARIMA

1. Introduction:

Forecasting the number of possible infected people in next few months can be useful in assisting the policy makers for designing better strategies and in taking productive measures. At initial stages most of the measures taken were based on assumptions and previous experience on the spread of other contagious diseases. The first case in India was reported on 12th February 2020. India stopped it International flights from 10th March and imposed complete lockdown all over the country from 25th March 2020. The total number infected case was 543 at that time. Since then the country has followed 4 consecutive lockdowns between 25th March and 7th July 2020. This lockdown imposed ban on public gathering, public commute, multiplex and cinema halls, shopping malls, bars and restaurants, shops and markets other than essential commodities etc. However, according to media reports the strictness of the lockdown was between low to moderate. While lockdown was observed more rigidly on urban areas, rural areas have been more

reluctant. Unlike countries like Italy, Germany, UK where lockdown was released only after the rate of new cases has taken a dip, India has started unlocking during the growing phase for some obvious economic constraints. As a consequence the number of new cases has increased exponentially since then. It is obvious from the case history of all such viral pandemics that it takes longer time in diminishing phase than the growing phrase. Assuming that it is very unlikely that India will be compelled to impose new set of lockdowns for the entire country, the prime concern now is when the country would see a sign of retreat of this decease. The objective of this paper is to estimate a probable time period, considering the current scenario (no vaccination, minimum lockdown), in which new COVID-19 cases in India should hit the peak and then start to decline. The proposed architecture of this work employs four components: one for statistical analysis of the time-series data using Auto Regressive Integrated Moving Average (ARIMA) model, second for estimating the unconstrained growth of infected people based on Susceptible Infected Recovered (SIR) model, third for predicting the future cases in India based on the supervised learning on cases of other countries using Machine Learning (ML) regression model, and the last one for forecasting the future cases from these three components using multivariate Long Short Term Memory (LSTM) model. The rest of the paper is divided as followed: section 2 describes the review works, section 3 describes the overall methodology, section 4 analyzes the results, and section 5 concludes this work.

2. Review Works:

More than fifty thousand research papers/articles have been published till mid of July 2020 since COVID-19 has spread worldwide [12]. Scientific researchers from all disciplines are trying their best to contribute in their own way. Although major expectation is from the biological sciences especially virology and medicine, ML and data science are also being used to understand, analyze and recognize the pattern of the spread of the disease. Researchers are trying to model and analyze the infection data of different countries using ML and statistical tools to identify any latent pattern and key factors in the spreading of the disease and then predicting the future growth and decay of the pandemic. Next we review some the recent works on analysis and prediction of COVID-19 cases mostly related to India.

Singh et al. [1] forecasted the COVID-19 epidemic in India with mitigated social distancing. This work introduces a mathematical model of the infection spread in a population considering social contact between ages. It uses the social contact structure as described by Prem et al. [11] and then the impacts of social distancing measures like workplace non-attendance, school closure, and lockdown are investigated. The authors suggested that sustained periods of lockdown with periodic relaxation will reduce the number of cases to a satisfactory level. Gupta et. al [2] investigated importance of lockdown in six social components i) restaurants and cafes ii) Grocery markets and food shops iii) community parks and

gardens iv) public transports v) private and government offices and vi) residential places. This work uses exponential and polynomial regression for predicting the number of future infections and claims a significant drop could be possible with considering first five of these categories. This work also shows impact of lockdown, social distancing and mass events on the growth of infected cases. Mortality predictions have been done through binary classification using decision tree model and recorded an accuracy of 60%.

For time series modeling of COVID-19 data, ARIMA model has been the most popular among the researchers. Tandon et. al [3] proposed ARIMA based model for prediction of future cases. Autocorrelation function (ACF) graph and partial autocorrelation (PACF) graphs are used to determine the initial parameters. These models are then used to test for variance in normality and stationary of the time-series data. ARIMA (2,2,2) appeared to be the best fit model with respect to scores of Mean absolute percentage error (MAPE), Mean absolute deviation (MAD), and Mean squared deviation (MSD). The model forecasted that the infection cases are expected to greatly rise in mid of May and may start to decline after that. Chakraborty et al. [4] proposed a two-folded approach- first for generating real-time forecasts of the future COVID-19 cases in multiple countries and second for predicting fatality rate by considering different demographic and disease characteristics. It uses a hybrid model of AIRAMA and wavelet based forecasting model. It also employs a decision tree regression model for predicting risk associated with fatality rate. In another work, Deb et al.[5] proposed a time series model to analyze the trend and pattern of the of COVID-19 outbreak. The authors claim that a time-dependent quadratic trend successfully captures the incidence pattern of the disease. This work uses ARIMA model to identify if the trend changes after any point. This work estimated the average contagious rate in India to be 1.42

Many researchers worked on epidemic modeling to forecast the future COVID-19 cases. Das [6] used an extension of SIR model known as Susceptible-Infectious-Quarantined-Recovered (SIQR) along with a statistical machine learning (SML) model for prediction of infected population. This work combines two approaches, one to understand the severity of the ground situation and the second for the prediction. With a polynomial regression model this paper predicted around total 66,224 cases by May 01, 2020 in India. In a similar work, Pandey et al. [7] used SEIR (Susceptible, Exposed, Infectious, Recovered) model with polynomial regression to predict the number of cases in next two weeks in India based on data available up to 30th March 2020. The model predicted around six thousand cases within mid of April 2020.

Tomar et. al [10] have given data-driven estimation using long short-term memory (LSTM) and predicted the number of COVID-19 cases in India. It also analyzed effect of preventive measures like social isolation and lockdown on the spread of the disease. The training data for this work contains case details

up to 4th Apr 2020, and the prediction is made for next 90 days. In another work, Tiwari [8] proposed a model for prediction of infected number based on simple Ordinary Differential Equation (ODE). This model predicted that infection in India would hit the peak with daily 22 thousand active cases during the last week of April followed by decline in active cases. In [9] Tiwari et al. used the pattern of china with help of ML model to predict COVID-19 outbreak in India. The predictive model is built using WEKA to predict the day-wise number of confirmed cases, recovered cases, and death cases. The model is trained with data from China and predicts the case of India, assuming that the trend of the pandemic would be same in both countries except a time lag.

However, almost all of the above the mentioned works failed to predict the outburst of the disease that India is currently observing. Acknowledging the famous "All models are wrong, but some are useful" theory, we try to find out the thriving and failing parts of the above discussed models. It is observed that model that relied on the data of lockdown and social distancing (in whatever form that are available) the prediction failed more than the models that considered only infected, recovered and death records. It is mainly due to unreliability of the lockdown and social distancing data and also small changes in its number affect the infected predictions largely. It is also observed that hybrid models predicted better than any single model, especially when ARIMA and SIR (or its variation) is used with ML models.

3. Methodology:

Figure1 represents the architecture of the proposed methodology. The dataset is collected from the WHO Corona Virus Disease (COVID-19) website [16]. It contains country-wise information about number of new cases, cumulative cases, new deaths, and cumulative deaths on each day. We selected a time period of 1st April to 27th July 2020 for this study. For the supervised regression model some additional features of each country like total population, population density, average temperature and average humidity during the time period are also added to the dataset [17]. We explore three different models independently on the dataset for the same time-period for predicting the number of daily infected cases in India using ARIMA model, SIR model, and an ensemble of regression models. The ARIMA model gives pure statistical prediction solely based on the pattern of daily cases. The SIR model gives an idea of how far the number of daily cases can rise in an epidemic model in unconstraint environment. The regression model first learns from the growth and decay pattern of daily cases of different countries using features like daily cases, total cases, temperature, humidity, population, and population density and then make prediction on the number of cases for India. Finally, a stacked multivariate LSTM model is trained with these three predicted time-series data to match with the actual data. Then the trained LSTM model makes forecasting on the number of future cases in India.

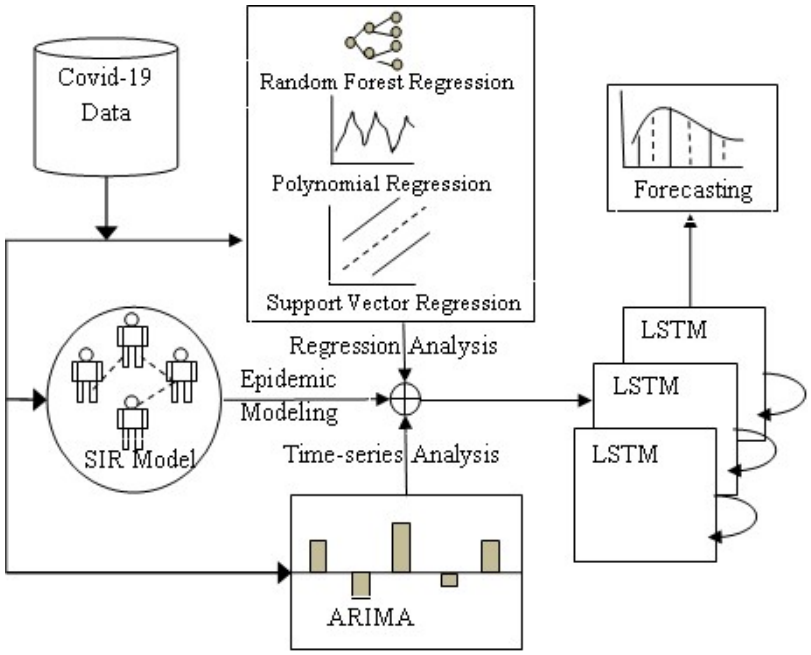

Figure1: Proposed architecture of prediction model

3.1 SIR Epidemiological model:

This work uses the classical Susceptible Infected Recovered (SIR) model, which is a compartmental model ideal for modeling epidemic cases like COVID-19 [5]. It splits the population into three compartments S, I, and R and defines the transition rates among them. People who are at the equal risk of getting infected are considered susceptible. Susceptible individuals who are getting affected by the virus are considered as infected. Infected people who are recovering (or dead) from the disease are considered as recovered. This model makes some basic assumptions: i) the total number of population (N) remains constant during the time period ii) recovered people do not become susceptible again iii) No one is vaccinated so everybody is equally susceptible, and iv) Quarantine or isolating infected people is not done (or failed badly).

The rate of change in each category is as follows:

$dS/dt = - \beta.SI$

$dI/dt = \beta.SI - \gamma I$

$dR/dt = - \gamma I$

Where, β= rate of contact and γ is the rate of recovery which is equal to (1/ number of days required to recover for an infected person). The rate of change of susceptible people (dS/dt) is always negative as the

number of susceptible people reduces with more people getting infected. If S_0 is the number of initial susceptible population, then at anytime (t) the number of susceptible (S_t) is always less than S_0. The rate dS/dt is same as the rate in which number of new infected people increases. As infected people recover at a constant rate (dR/dt), the effective rate of increase of infection (dI/dt) is equal to (- dS/dt) - dR/dt.
Epidemiological model defines a parameter R_0 as the basic reproduction number which denotes the contagious ability of the disease. It is defined as the average number of people who can get infected from a single infected individual. It is formulated as β/γ. If the value of $R_0 < 1$, it means that the disease would not spread. If the value is $R_0 = 1$, it signifies the spread is stable or endemic. If the value of $R_0 > 1$ it means the spread would be pandemic.

3.2 ARIMA model:

It is a combination of two models namely Auto Regressive (AR) and Moving Average (MA) models. The 'I' stands for integration of the two models. An ARIMA model is represented with three parameters p, d and q, where p is the autoregressive lag, d is the order of differencing, and q is the moving average. ARIMA model is formally expressed as [18]

$$y_t = \theta_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

Where, the predicted value y_t = Constant + Linear combination Lags of y_{t-i} (upto p lags) + Linear Combination of Lagged forecast errors ε_{t-j} (upto q lags)

ARIMA model requires data to be stationary. A stationary time series has the mean and variance constant over time. If the data is non stationary i.e., the data has trend or seasonality, then we need to transform it into stationary series using differencing. The integration parameter *d* refers to the number of times differencing is required to get the data stationary. We can test stationarity with Augmented Dickey-Fuller test (ADCF). The p value of ADCF test should be much less than 1 for stationary data. Since the data for India has a high increasing trend we performed differencing of order 2 to make the data stationary (Figure 2(a) and 2(b)) and obtained p-value of 0.004. For estimation of ARIMA parameters p and q we use the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots (Figure 2(c) and 2(d)).

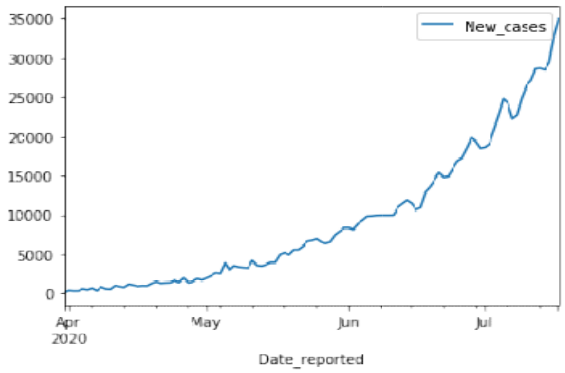

Figure 2(a): Original data (Non-stationary) with p-value:1.0

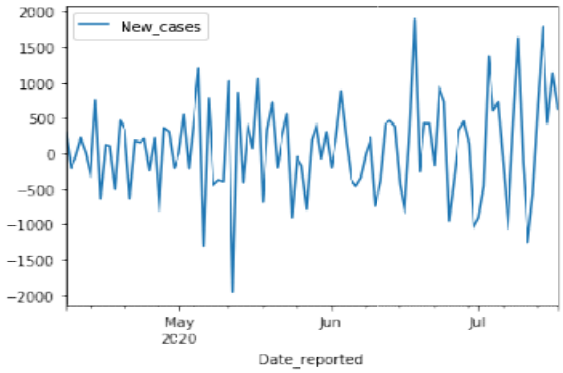

Figure 2(b): Differenced data (Stationary) with p-value: 0.00047

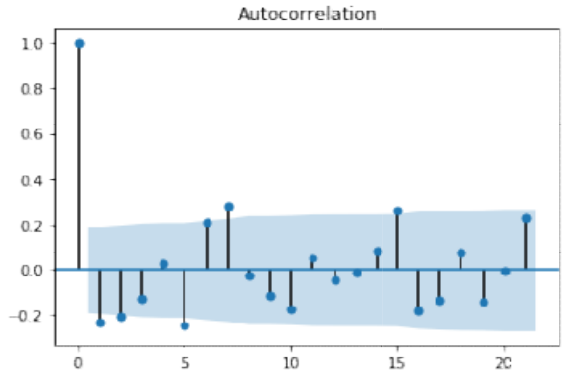

Figure 2(c):ACF graph after differencing(2)

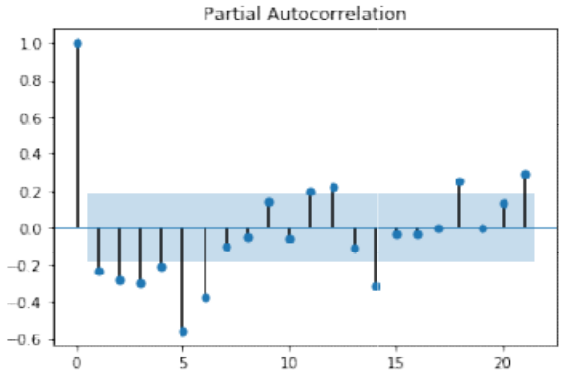

Figure 2(d): PACF graph after differencing(2)

3.3 Ensemble Regression Model:

For the regression model we employ Ensemble learning where multiple models are trained to solve the same problem independently and then combined to get the final result. We use three regression models namely Polynomial Regression, Random Forest regression and Support Vector Regression which are trained on same data and then use the voting regressor model to aggregate the results.

Polynomial regression: Polynomial regression is a special case of linear regression in which the relationship between the independent variable x and the target variable y is modeled as nth degree polynomial of x.

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \cdots + \theta_n x^2 + \varepsilon$$

The objective is to minimize the error cost (J), which can be written as:

$$J = \frac{1}{n}\sum_{i}(\hat{y}^{(i)} - y^{(i)})^2$$

Where for n number of points, $y^{(i)}$ is the true value, and $\hat{y}$ is the predicted value.

Non-Linear SVR: Support Vector Regression works similar to SVM but only for regression problem. It gives the flexibility to deal with an acceptable error within the model. The basic objective is to find an appropriate hyperplane that fit the data. For the polynomial features, the kernel function transforms the data into a higher dimensional feature space to find out the separation plane.

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*).K(x_i, x) + b$$

Where, the kernel in use (K) is Gaussian Radial Bias Function $K(x_i, x) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$

Random Forest Regression: Random Forest itself is an ensemble method that aggregates the learning of multiple Decision Trees. To use a decision tree for regression problem, we need an impurity metric that is suitable for continuous variables. Therefore, instead of Entropy we use the weighted mean squared error (MSE) of the children nodes as the impurity measure:

$$MSE(t) = \frac{1}{N_t}\sum_{i \in D_t}(y^{(i)} - \hat{y}_t)^2$$

Where, N_t is the number of training samples at node t, D_t is the training subset at node t, $y^{(i)}$ is the true value, and $\hat{y}_t$is the predicted target value.

3.4 LSTM model:

LSTM is a type of Recurrent Neural Network (RNN) that allows the network to retain long term dependencies at a given time from many previous time-steps. It has the ability to perform well both in univariate and multivariate time-series forecasting. A multivariate time series data has more than one observation for each time step. Many researchers have advocated that multivariate analysis gives better performance in forecasting than by studying just one variable [13,14,15]. In this work, time-series data predicted by ARIMA model, Regression Model and SIR model learn three different aspects of same input data and forecast differently. These three forecasted time-series are used as multivariate time-series to predict the future new cases in India with help of LSTM model.

4. Result and Analysis:

Figure 3 depicts the SIR predictions in India. The recovery period is taken as 14 which implies $\gamma = 1/14$. We consider the social contact rate=2.5 which makes the effective contact rate β=0.15. Since the dataset in use counts infected cases only from who have been tested, instead of considering the entire population in SIR model we consider the number of people who can come under test during the period under consideration (d=230 days). Considering the fact that the total number of test done up to July is 18.5M and the highest number of test in a day to be 320 thousand (t), we consider the total population to be tested up to Nov 2020 would be approximately 73.6M (d*t). The basic reproduction number R_0, which is the ratio of β and γ, is calculated to be 2.1. The selections of parameters are based on the result having best match with the current (up to July) record of the daily cases in India. The result predicts that the number of infected people (active cases) in India can reach up to 0.9 M by December 2020 and then would start reducing, if current trend continues. However, we consider this could be an overestimation as protecting measures like lockdown, social distancing, and use of mask and sanitizer which would inevitably work as resisting factors against the exponential spread of the disease, are not considered in this case. The SAR model gives a simplistic idea of what would be the growth of the disease in India in an unconstrained environment. Finally, the time-series prediction of daily cases is calculated with help of differencing consecutive active cases, recovered cases and fatality cases.

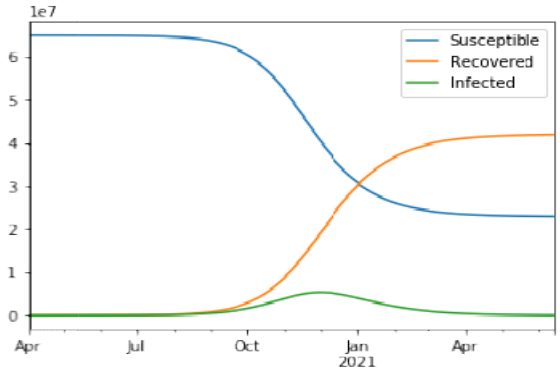

Figure 3: SIR prediction

The original non-stationary data for daily cases in India is differenced to make it stationary (Figure 2(b)). We used ACF and PACF graph on this stationary data for an early estimation of the parameters (p,d,q) of the ARIMA model (Figure 2(c), 2). Then the best combination (5,2,1) was found by applying grid-search method on the early estimated values. Figure 4(a) and 4(b) depict the ARIMA prediction on training data and future data respectively. Since ARIMA is solely based on previous time stamps and the training data has only the growing phase of cases in India, it predicts an exponential growth in the future time period.

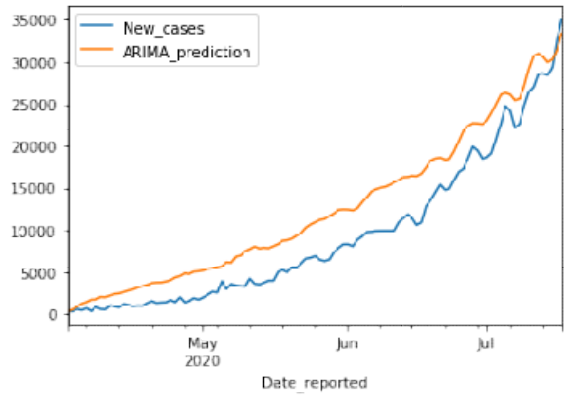

Figure 4(a): ARIMA on training data

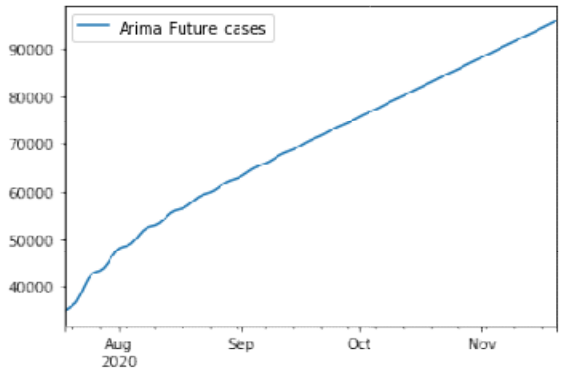

Figure 4(b): ARIMA prediction

For regression analysis using ML models, we introduce a new feature set (temperature, humidity, population and population density) to the dataset in addition to daily new cases with a trail of previous 5 days. We selected six countries from Europe and Asia namely Italy, UK, Russia, Spain, Saudi Arab, and Pakistan where the number of daily cases are already in the reducing phase. In order to ensure uniform scaling, instead of using number of daily cases we used number of daily cases per 1M of population. The average temperature and humidity of each country during the time period is also considered. The validation split = 0.3 is done on the dataset for testing the efficiency of the models. For polynomial regression we use linear regression with polynomial feature transformation of order=3. For RF regression we use 50 decision tree estimators with maximum depth=10. In SVR we use RBF kernel with gamma coefficient=0.1. Finally, an ensemble method – Voting-Regressor is used on top of these three base estimators. Table 1 shows comparative performances of different regression techniques. Mean Squared Error (MSE) is used as a common measure of performances.

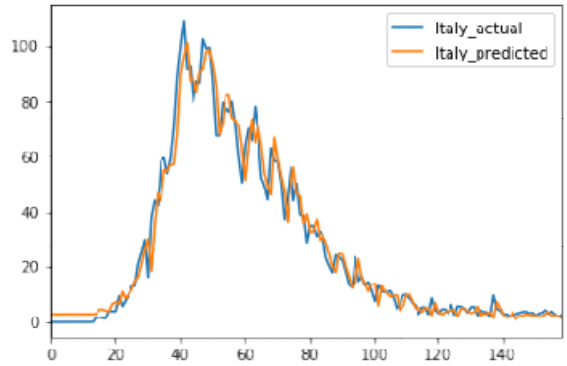

Figure 5(a): Ensemble Regression on training data

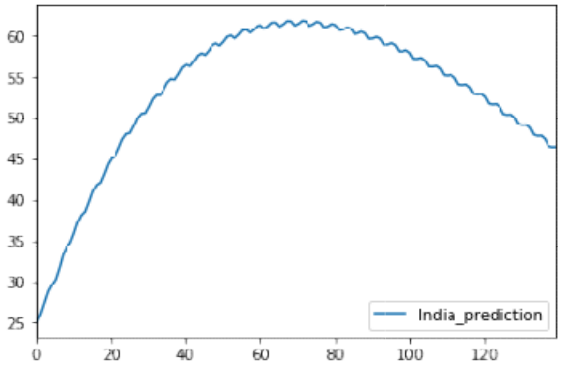

Figure 5(b): Prediction of Regression Model

Table1: Performance of Regression Models

Model	Hyper-parameters	MSE
Polynomial Regression	Degree=3	0.91
Random Forest Regression	n_estimators=50, max_depth=10, criterion='mse'	0.48
Support Vector Regression	kernel='rbf', gamma=0.1	0.21
Ensemble Regression	class=VotingRegressor, estimators=[LR,RF,SVR]	0.38

The original dataset has daily COVID-19 cases in India between 2nd April and 17th July (107 days). All the three models predicted daily cases for India for a span of 230 days (2nd April to 18th Nov). We split this range into *current_data* (2nd April to 17th July) and *future_data* (18th July to 18th Nov). The multivariate LSTM model is trained on the *current_data* with the original dataset as truth value [Figure 6(a) 6(c)]. Then it uses the *future_data* as test data to forecast the future cases [Figure 6(d)]. We prepare three series input data with 5 days previous time-steps for each day. Thus the input shape of data becomes (5,3) which is fed to a stacked LSTM model of 50 nodes and 3 layers. The training of the model is done on first 107 days of data and forecasting is made on data of next 123 days. The result reveals that India can hit the peak at 102nd day and reach as high as 90 thousand new cases of infections per day [Figure 6(d)]. Now 102 days after 23rd July 2020 (18th July + 5days lag) is 2nd Nov 2020. Therefore, the forecast of our model for India to hit the peak of daily cases is in the first week of Nov 2020 with 90 thousand daily cases. The model shows a decline after that.

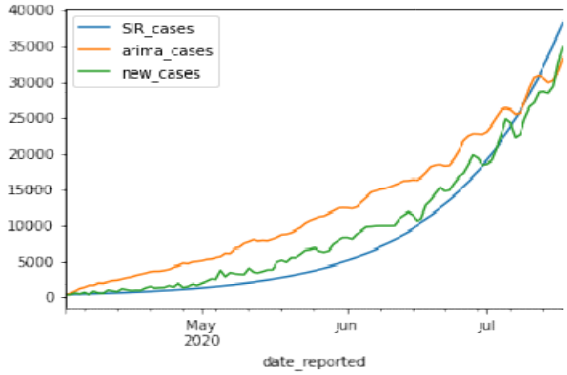

Figure 6(a): LSTM input series

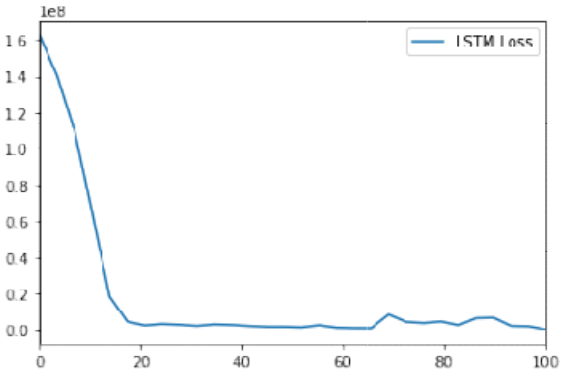

Figure 6(b): LSTM training loss

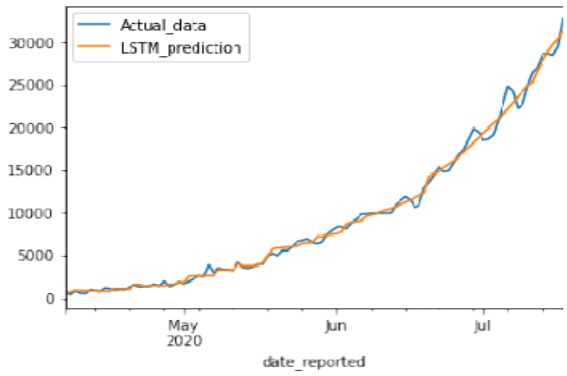

Figure 6(c): LSTM training data

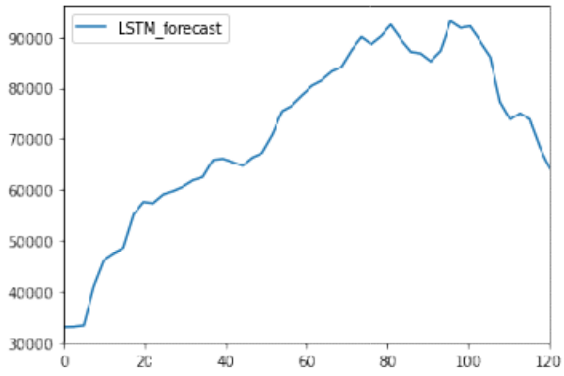

Figure 6(d): LSTM Forecasting

5. Conclusion:

This paper presents a prediction on probable time period for daily new cases to hit the peak in India. Three different modeling techniques namely SIR, ML Regression, and ARIMA are used independently on the same data to produce three different time-series for next four months. Then these data are used on a stacked multivariate LSTM model for forecasting the future cases in India. It is observed that the number of daily cases can reach at 90 thousand by the first week of November 2020 and then start declining. This work transfers the learning on infection pattern from other countries into the context of India. However, instead of relying solely on this prediction, we also consider prediction from two other models SIR and ARIMA which explores only data of India and hence showed exponential growth. Therefore, the LSTM forecasting is more reliable as its input represents three different aspects of the spreading of the disease. The limitation of this work is not considering issues like social distancing, quarantine, and isolation despite of the fact that it is already proven that they can have big impact on the numbers of infected cases if maintained properly. However, at present there is no reliable data available on these attributes but the proposed model has the capability of adopting it when it will be available in future.

References:

1. Singh, R., & Adhikari, R. (2020). Age-structured impact of social distancing on the COVID-19 epidemic in India. arXiv preprint arXiv:2003.12055.
2. Gupta, R., Pal, S. K., & Pandey, G. (2020). A Comprehensive Analysis of COVID-19 Outbreak situation in India. medRxiv.

3. Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. arXiv preprint arXiv:2004.07859.
4. Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel corona virus (COVID-19) cases: A data-driven analysis. Chaos, Solitons & Fractals, 109850.
5. Deb, S., & Majumdar, M. (2020). A time series method to analyze incidence pattern and estimate reproduction number of COVID-19. arXiv preprint arXiv:2003.10655.
6. Das, S. (2020). Prediction of COVID-19 disease progression in India: Under the effect of national lockdown. arXiv preprint arXiv:2004.03147.
7. Pandey, G., Chaudhary, P., Gupta, R., & Pal, S. (2020). SEIR and Regression Model based COVID-19 outbreak predictions in India. arXiv preprint arXiv:2004.00958.
8. Tiwari, A. (2020). Modeling and analysis of COVID-19 epidemic in India. medRxiv.
9. Tiwari, S., Kumar, S., & Guleria, K. (2020). Outbreak Trends of Coronavirus Disease–2019 in India: A Prediction. Disaster medicine and public health preparedness, 1-6.
10. Tomar, A., & Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Science of The Total Environment, 138762.
11. K. Prem, A. R. Cook, and M. Jit, "Projecting social contact matrices in 152 countries using contact surveys and demographic data," PLoS Comp. Bio 13, e1005697 (2017)
12. Nature Index Blog: https://www.natureindex.com/news-blog/the-top-coronavirus-research-articles-by-metrics
13. Cai, Y.; Wang, H.; Ye, X.; An, L. Multivariate Time Series Prediction Based on Multi-Output Support Vector Regression. In Knowledge Engineering and Management; Springer: Berlin/Heidelberg, Germany, 2014; pp. 385–395.
14. Jin, X.; Yu, X.; Wang, X.; Bai, Y.; Su, T.; Kong, J. Prediction for Time Series with CNN and LSTM. In Proceedings of the 11th International Conference on Modelling, Identification and Control (ICMIC2019), Tianjin, China, 13–15 July 2019; Springer: Singapore, 2019; pp. 631–641.
15. Du, S.; Li, T.; Yang, Y.; Horng, S.J. Multivariate time series forecasting via attention-based encoder–decoder framework. Neurocomputing 2020, 388, 269–279.
16. URL: https://covid19.who.int/
17. URL: https://www.weather-atlas.com/en/climate
18. Sowell, F. (1992). Modeling long-run behavior with the fractional ARIMA model. Journal of monetary economics, 29(2), 277-302.