Time Series Analysis using Deep LSTM Networks for predicting COVID-19 Cases in India

Shubhnesh Kumar Goyal

Associate Professor, D.S. (P.G.) College, Aligarh, India
shubhnesh[at]icloud.com

Abstract: ***While COVID has taken a toll globally, its imperative damage in India has been serious because of a large population base. In this scenario, daily prediction of COVID cases can help concerned authorities better brace themselves of the upcoming effect. Since cases form a time series data, their prediction remains a challenge due to inherent order in data points, which is tough to capture in statistical regressions. In addition, number of cases depends on a numerous factor in practical life, and to arrive on an exhaustive list for the purpose of modelling poses another challenge. To tackle these problems, we present a study spanning January 2020-April 2020, outlining way of using LSTMs for predicting 1-3 days in advance the number of cases in India and present a comparative analysis over inclusion of different factors in the prediction and its effect on accuracies. We achieved a R2 score of over 0.9 for short periods spanning 1-5 days, but model fails to capture long term (over 15 days) trend. Similarly, adding cases from Top 5 states as input factors increased the accuracy significantly for lookback = 4 to 0.99.***

Keywords: Deep Neural Networks, LSTM, Epidemiology, COVID-19 prediction

1. Introduction

COVID-19 is the disease which crippled whole of the world in 2020. It started with a pneumonia like symptoms and developed and evolved into a pandemic within months. Governments all across the globe tried to mitigate the disaster as best as they could, but with increasing number of cases it became tough to fight. WHO also issued warning, aids and recommendations in response, along with fast paced research in the field we were able to take hold of the scenario. Still, the load healthcare industry saw was unprecedented. One way to ease up the congestion was better preparedness which can be achieved through prediction of cases ahead. Many simulations were used to model the spread but they are generally computationally demanding. Statistics paved the way for simpler models for predicting cases[9], along with Deep Learning. We follow a Deep Learning based route in this paper using LSTM to predict the advent of disease in near futures, since statistical methods provide reasonably good results for longer windows but neural nets are extremely efficient in capturing near term trends. We hope that study will help all the concerned authorities in dealing with the situation better and important lives will be saved.

Dataset preparation

International data was taken from John Hopkins maintained database [5,4,6] and national data was taken from the government-maintained dataset on the COVID-19 Tracker website [7,8]. MinMax scaling of the data is done in all of the models and the scaler is fitted only on the training data to prevent the forward bias. Two different sets of features are also used with details given afterward. Data till 12th April (from 14th March for India) was used for reports and data till 13th April was available when the report was being prepared.

Long Short Term Memory Neural Networks (LSTM):

Deep Learning is a method used to find underlying complex relationships between the data that can be exploited to predict the result when given new input. LSTM is a type of Neural network that takes into account the sequential nature of data and thus this is used for the prediction of cases infected with coronavirus that will be reported the next day given we know the number of cases till now.

Dataset for testing and training purposes was prepared to keep in mind the 3-d requirement of LSTM model in Keras, and thus the first dimension has a number of observations we have, the second dimension has the timesteps or lookback i.e., how many points before till data are used to train for as one input and the third dimension has features of the dataset. **Averaged results for three separately trained models are reported.**

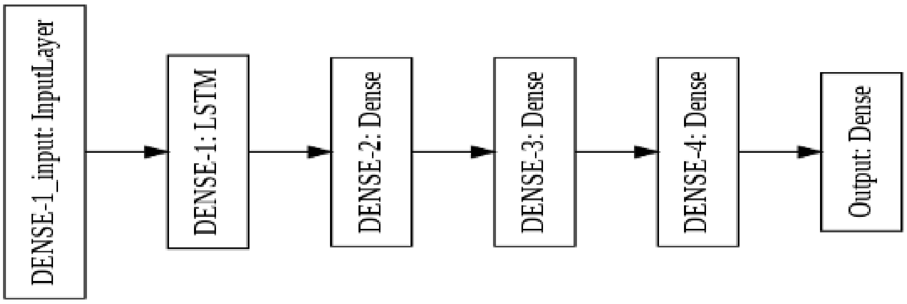

Figure 1: Represents the architecture of the model, consisting of 5 Dense Layers with linear activation

Model architecture (fig 1) has five layers of **[256, 128, 64, 32, 1] neurons with linear activation**. A split of **0.9 is used for training** as the rapidly evolving nature of cases requires current datapoint for learning. Validation loss is used as metrics and **validation split of 0.1 is used along with 200 epochs** for training on the training dataset. Only one feature has been included and that is the cumulative confirmed cases due to the largely dependent nature of prediction on various physiological, biological and societal factors making it highly susceptible if only some features are used. Results are reported for various lookbacks (fig 2), with **future predictions obtained by fitting the trained model on the whole data again.**

	Lookback=1	Lookback=2	Lookback=3	Lookback=4
R2 Train	0.99	0.99	0.99	0.99
R2 Test	0.87	0.87	0.96	0.83
12 April (9212)	8887	8983	9125	8161

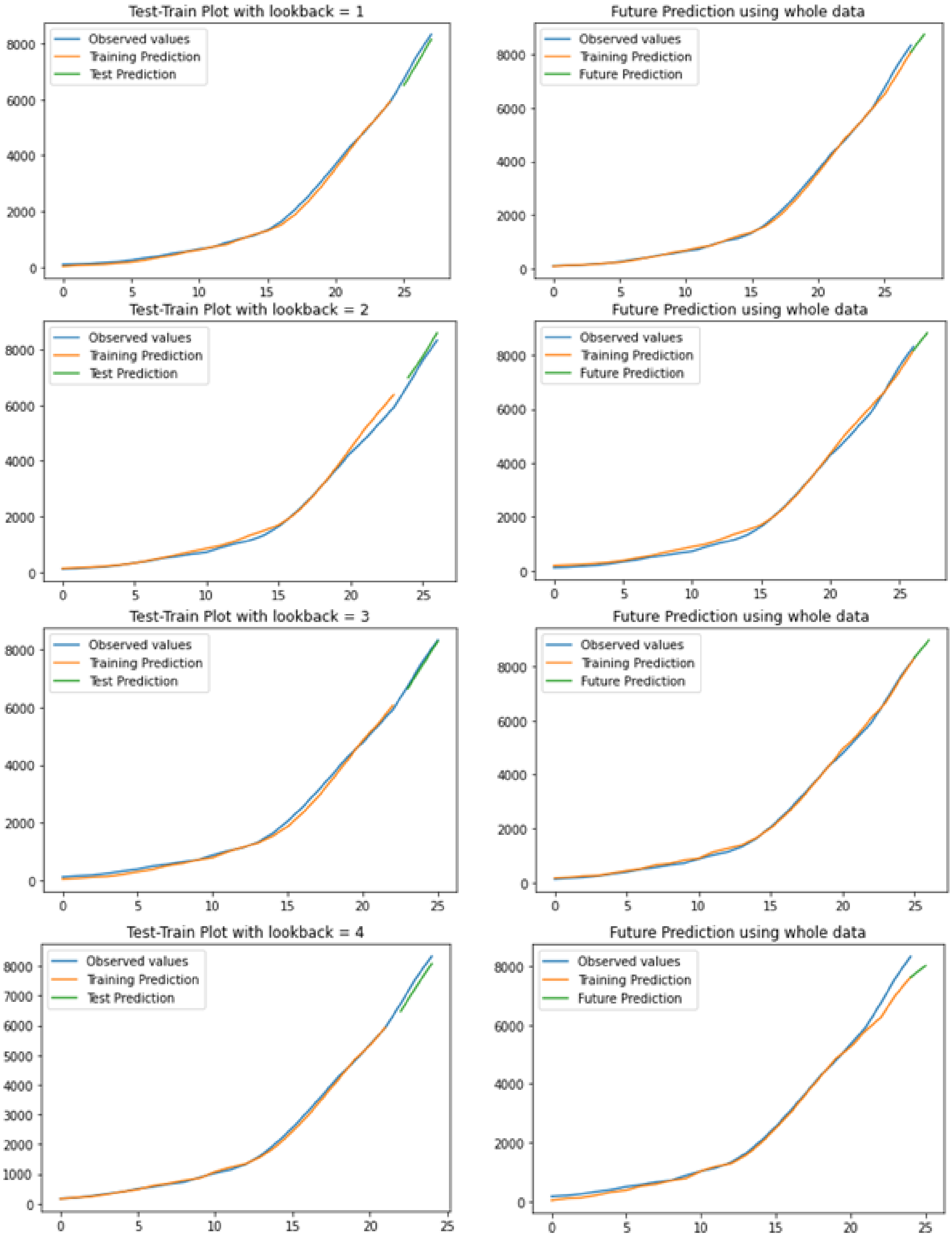

Figure 2: Represents training and prediction plots for four lookback cases with number of cases on y-axis and number of days on x-axis

LSTM without additional features can be used to predict for multiple future days ahead, by giving the previous predicted result as the input repetitively. We have shown the prediction for the next five days i.e., of 12th, 13th, 14th, 15th and 16th April using data till 11th April. Doing it for a longer stretch results in deviation as factors change a lot each day like government policies, test rate increase and many others, and error in each step directly propagates into the next. Since the best results were obtained for lookback=3, the results shown are for that (fig 3). Three similar models were separately trained and tested, and an average of all the results is presented.

12th April	13th April	14th April	15th April	16th April
9276 (9212)	10210(10454)	11088(not yet)	11963(not yet)	12836(not yet)

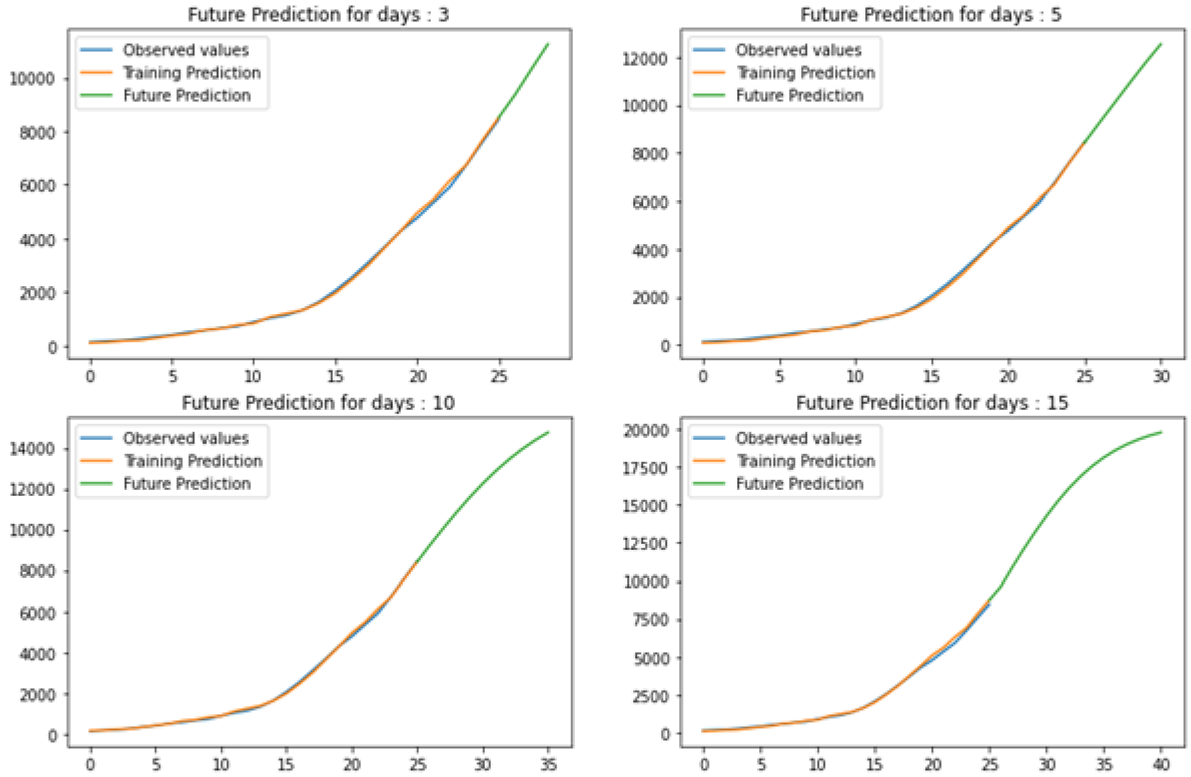

Figure 3: Represents training and prediction plots for lookback=3, with varying future prediction length

Thus, for smaller durations from 1-5 days, we arrive at a good approximation but as the number of days increases, models tend to saturate at a value and provide no good approximation. Saturation will come based upon various factors and we are still not clear if the model accounts for them or not.

To better capture the changing demographics, **we tried to add different features** as well. We tried two different sets of features,

1) Six different features: **Total confirmed cases, New confirmed cases, Total recovered cases, New recovered cases, Total deaths, and new deaths**.
2) Six different features: **Total confirmed cases and top-five state totals based on the highest daily growth that particular day.**

For the first set of features, results are presented here:

	Lookback=1	Lookback=2	Lookback=3	Lookback=4
R2 Train	0.99	0.98	0.99	0.96
R2 Test	-0.2	0.88	0.74	0.94
12 April (9212)	9431	9817	8803	8619

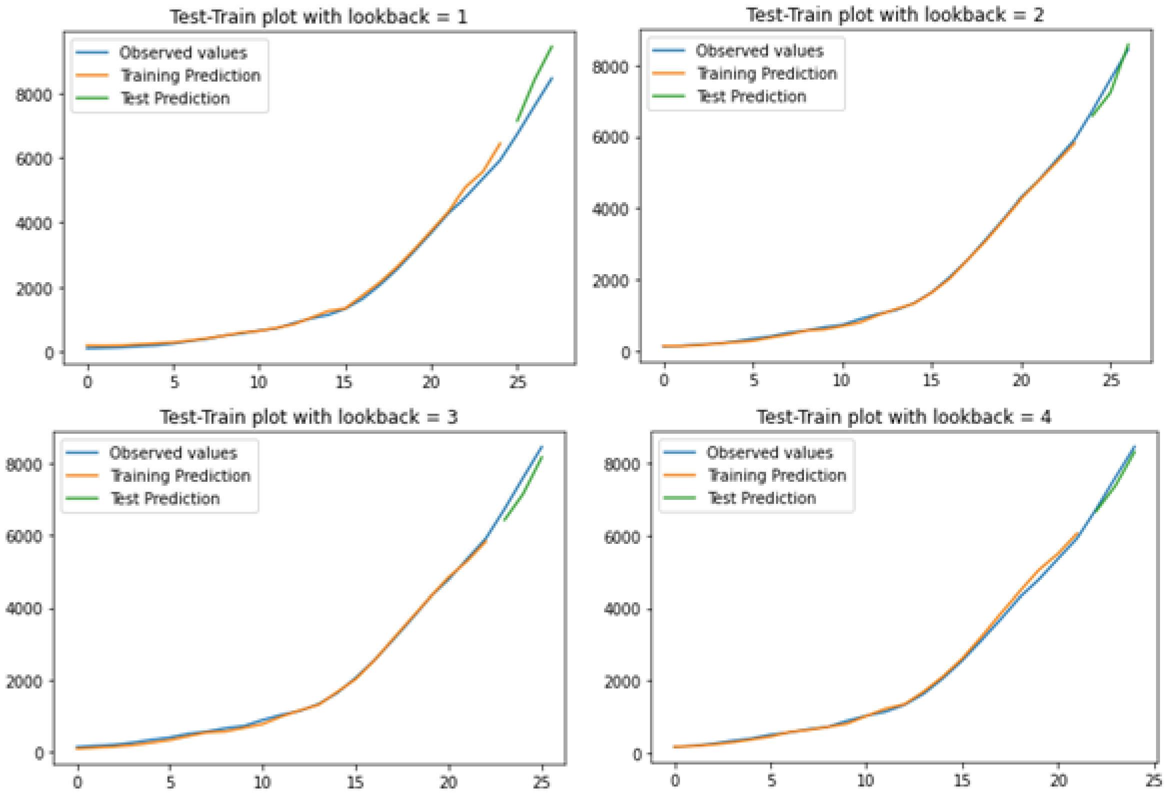

Figure 4: Represents training and prediction plots for four lookback cases with number of cases on y-axis and number of days on x-axis using first set of features.

For the second set of features, results are presented here:

	Lookback=1	Lookback=2	Lookback=3	Lookback=4
R2 Train	0.99	0.99	0.8	0.99
R2 Test	0.89	0.39	0.69	0.99
12 April (9212)	8644	8834	7911	8995

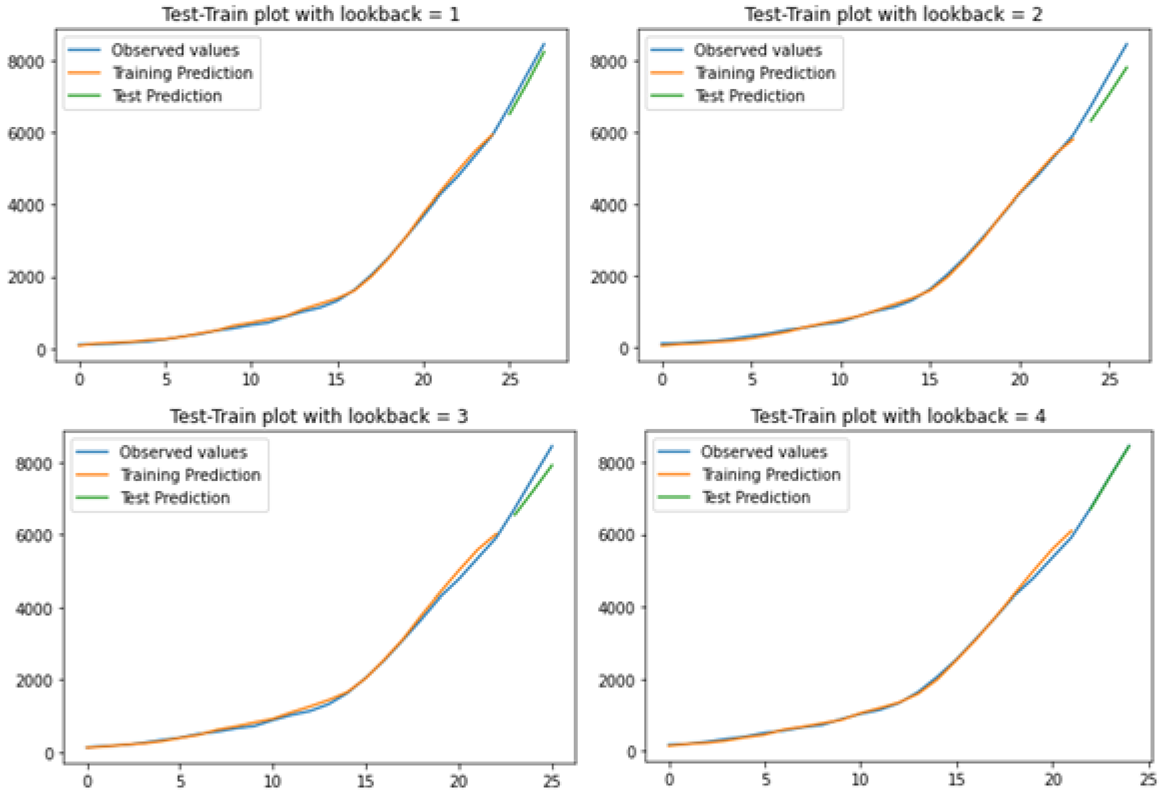

Figure 5: Represents training and prediction plots for four lookback cases with number of cases on y-axis and number of days on x-axis using second set of features.

In general, considering more features in LSTM makes the input noisy and output very susceptible to these changes and thus the overall quality of the prediction is reduced.

Also, **this puts a tight ceiling on advance days we can predict since we cannot predict more than one**.

These results represent the **demographic factors such as local state data that can be used in the future to make the model more exhaustive** as well as other input factors that can be included.

2. Discussion

We started our work by choosing the LSTM models for prediction, and showed that given proper architecture, they have excellent chances of predicting the near future number of cases. Our report shares guidelines and results for one such model which has performed really well and we think that can be used for preparation. We further found out that for large durations LSTM model tend to saturate out. Adding additional features to the data increases the accuracy significantly and paves way for further exploration. Due to accumulating error in long term predictions, we see how after 5 days, results tend to saturate out. This limits the usage of this method for longer predictions but for short spans results are in great agreement, and thus this can be used to model infrastructure requirements by individual hospitals for short spans.

Adding six additional input features gives more insight to the model and thus predictions increase slightly. This also encourages the team to explore more related inputs to capture linger trends.

References

[1] Chae, S.; Kwon, S.; Lee, D. Predicting Infectious Disease Using Deep Learning and Big Data. Int. J. Environ. Res. Public Health **2018**, 15, 159
[2] Chiou-Jye Huang, Yung-Hsiang Chen, Yuxuan Ma, Ping-Huan Kuo,Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China, https://doi.org/10.1101/2020.03.23.20041608
[3] Joceline Lega, Heidi E. Brown, Data-driven outbreak forecasting with a simple nonlinear growth model, Epidemics, Volume 17, 2016, Pages 19-26, ISSN 1755-4365, https://doi.org/10.1016/j.epidem.2016.10.002.
[4] https://github.com/CSSEGISandData/COVID-19
[5] https://github.com/datasets/covid-19
[6] https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
[7] https://docs.google.com/spreadsheets/d/e/2PACX-1vSz8Qs1gE_IYpzlkFkCXGcL_BqR8hZieWVi-rphN1gfrO3H4lDtVZs4kd0C3P8Y9lhsT1rhoB-Q_cP4/pubhtml#
[8] https://www.covid19india.org/
[9] Shubhnesh Kumar Goyal, "Predicting COVID-19 Cases in India using Global Data", International Journal of Science and Research (IJSR), https://www.ijsr.net/search_index_results_paperid.php?id=SR21720164321, Volume 10 Issue 7, July 2021, 1387 - 1394