Learning from Others: Daily COVID-19 Cases Prediction in India using Ensembles of LSTM-RNNs

Debasrita Chakraborty
Machine Intelligence Unit
Indian Statistical Institute
Kolkata, India
debasritac@gmail.com

Debayan Goswami
Department of Computer Science and Engineering
Jadavpur University
Kolkata, India
debayang.ju@gmail.com

Susmita Ghosh
Department of Computer Science and Engineering
Jadavpur University
Kolkata, India
sushmitaghoshju@gmail.com

Jonathan H. Chan
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
jonathan@sit.kmutt.ac.in

Ashish Ghosh
Machine Intelligence Unit
Indian Statistical Institute
Kolkata, India
ash@isical.ac.in

***Abstract*—Accurate prediction of the number of COVID-19 infected cases per day is fast becoming a critical necessity globally to mitigate the burden on the various health systems. In a densely populated country like India which has currently the second highest number of infections and limited medical support, it is a need for the authorities to know the statistics beforehand to address these issues more effectively. In this article, a data driven transfer learning based model is proposed that takes into account the conditions of different countries which have witnessed the COVID-19 infection. We have taken four countries to be the source domain for transfer learning scenario namely, the United States of America, Spain, Brazil and Bangladesh. We have pre-trained four different LSTM-RNN models with each of the country's data and have re-trained (fine tuned) each of the models using only a very small portion of Indian data on COVID-19. Predictions of these four models are averaged to get the actual prediction. It is seen that such an ensemble model outperforms all the compared models. This may be due to the fact that the four LSTM-RNNs used here could successfully take into account the diversities of conditions. As India is a diverse nation with variety of climates, it makes more sense to incorporate such transfer learning techniques.**

***Index Terms*—LSTM-RNN, Transfer learning, COVID-19 prediction**

I. Introduction

The Indian Government had imposed a nationwide lockdown [4] on 24 March, 2020 to curb the spread of the corona virus pandemic. Several other countries are also trying their best to predict the total active cases their country would face on a particular day [2]. The governments are in dire need of witnessing the actual statistics of the predictions of the spread of the virus. This article presents a novel prediction system based on transfer learning [5] framework.

In case of viral contagion like that of the novel corona virus, different countries have taken different actions and imposed strict measures. Even under such conditions, there are certain countries which have been adversely affected by the virus.

Worldometer has been sharing a country-wise live statistics of manually analyzed and validated data regarding the COVID-19 cases all over the world. The proposed method has used the data from the Worldometer website, https://www.worldometers.info/coronavirus/, to analyse the data of four countries and use the knowledge from these countries for prediction of cases in India [7]. The four countries that have been chosen are the United States of America, Spain, Brazil and Bangladesh. These countries have been carefully selected to incorporate four different scenarios that might be relevant for the prediction of Indian COVID-19 cases.

There are four different situations in these four countries. The number of active cases from the USA have been seen to reach a short plateau region during the last week of May 2020, only to be spiked again in June 2020. The short plateau again occurred during the month of September 2020 and again the spike was witnessed during the month of December 2020. This clearly shows short bursts of repeated contagions in waves. India is also witnessing a second wave of infections now similar to what USA experienced. On the other hand, cases in Brazil show a peak during the first week of August 2020, and then a plateau during the first week of November 2020, and then shows a steady exponentially rising trend of a starting second wave. There are increasingly more number of cases with each wave in Brazil. We cannot guarantee that the subsequent waves in India may not take this form. So, we need to take this situation into account as well. The third country chosen is Spain, where the contagion reached a relatively longer plateau between May 2020 to July 2020 and then there was a second wave that peaked around the end of October

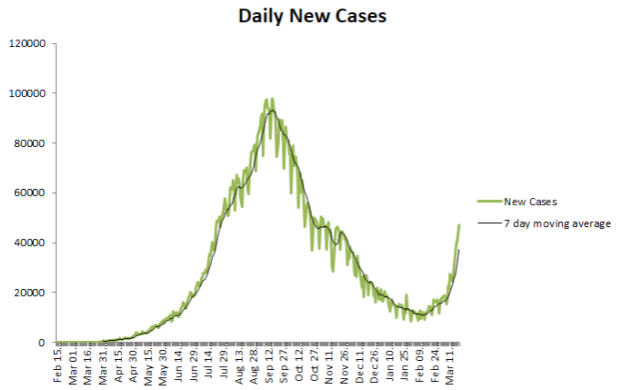

Fig. 1: Daily new COVID-19 cases in India

which was much worse than the first one. This was the worst hit country in the first wave where the number of daily deaths was very high. This may be another scenario for the Indian subcontinent during its second wave of contagion. Bangladesh has been witnessing a decline in COVID-19 cases daily even though being a much densely populated country than India [Figure 1]. Since, India and Bangladesh seem to have passed its first wave and are already witnessing the second one, such findings can be incorporated into the present model as India shares its border with Bangladesh and having similar habitat.

Researchers [3] are now sceptical about the situation of Indian population because:

- people are now in the "Unlock Phase" and are going out of their homes and
- second waves in other countries are much worse than the first one.

Some experts have researched on the spread of COVID-19 infections and have found that the effect of the population density is negligible under strict lockdown policies [10]. However, by dealing with the virus for over a year, people are gradually returning to their normal lives but wearing masks. As people are moving out of their houses again, population density factor that will play its part now. So, it is very critical to know how the virus will spread through the community in a densely populated country as India [1].

To handle such a problem, a transfer learning based multivariate prediction model is proposed using deep LSTM networks. Such a model has been already proposed by Gautam [9]. However, we are incorporating a diversity in the proposed method by taking the average of four transfer-learnt models for four different scenarios.

II. Proposed Method

The proposed method is an attempt to forecast an estimated value of the possible COVID-19 cases in India while learning the behavior of the disease spreading curve from other countries. This model incorporates transfer learning in predictive analysis using LSTM networks [8].

A. Transfer Learning

Transfer learning has been used in machine learning in cases where the model can be efficiently built using a source domain data and then it could be applied for a target domain data. As a rule of thumb, usually the source domain and the target domain should have a similar behaviour. In such cases, the source domain acts as the provider for the majority of the training data and then, the target domain can be used just for fine tuning the model.

In the present scenario of learning from COVID-19 data, since the infection in India surpassed its first wave and is witnessing a more severe second wave, it might be better to learn from the data of other countries. However, it must also be useful to incorporate information about multiple possibilities into the model.

As stated earlier, this was the reason for choosing data from four countries with four different scenarios. The United States of America has been the worst hit with more than 22 million cases till date. The COVID-19 infection has been seen to be recurring in multiple waves with each wave being stronger than the former one. Although the population density of the USA is lower than that of India, it had a higher infection rate during the first wave. Bangladesh is a heavily populated country but has less COVID-19 cases than India. Brazil has started to witness the second wave recently and Spain has already been hit harder by the second wave than the first one.

As India is now undergoing an "Unlock phase", it would be obvious that the lockdown restrictions on movement of people is not there anymore and the population density will now come into picture. We have, therefore, considered the population density of each country in our consideration too. Table I shows the population density (retrieved from https://www.worldometers.info/world-population/population-by-country/).

Sl. No.	Country	Population Density (per km^2)
1	USA	36
2	Spain	94
3	Brazil	25
4	Bangladesh	1265
5	India	464

TABLE I: List of countries and their respective population densities.

B. Training and testing

It would be worth mentioning that this proposed model has been used to investigate transfer learning scenario in COVID-19 prediction. We have taken the data of one whole year from 15 February, 2020 to 14 February, 2021 for training purposes. The rest of the data (from 15 February, 2021 to 21 March, 2021) is used for testing. The model is constructed to provide only predictions for the next day using the data of the past five days. We have trained four models depending on the source country it has been trained on. We have used sliding window model for training the LSTM network. In all the cases, we

have trained the model on the source country (The United States of America, Spain, Brazil or Bangladesh) for the said period and fine tuned on the target data of India for the period of 1 January, 2021 to 14 February, 2021. We have then tested each of our models on India's data from 15 February, 2021 to 21 March, 2021.

The fine tuning is required as India has started mass vaccination recently and therefore, its rate of COVID-19 infection has been fluctuating. Moreover, Indian government policies and overall immunity of people are different from the other source countries. This is why the prediction models which have been pre-trained on the source data, need to be fine tuned on India's data to incorporate these factors into account. Without this fine tuning, it would be conceptually meaningless to run India's data on the model as different countries have different environmental and societal settings.

III. Results

The proposed model has been compared with four major time series prediction models namely, Autoregressive Integrated Moving Average (ARIMA) [6], Vector Auto-Regression (VAR), Facebook's Prophet [2] and a Standard LSTM-RNN trained on Indian data. To keep the comparison fair, all the four models are re-trained on India's data as mentioned earlier (from 15 February, 2020 to 14 February, 2021) and tested for the period 15 February, 2021 to 21 March, 2021. The standard LSTM-RNN trained in the traditional manner is compared with the proposed transfer learning setting of pre-training using different countries and fine tuning using a small amount of India's data to show the efficacy of our proposed method. We have used four variables in our consideration for prediction: Total cases, Daily New cases, Total Recovered cases and Active cases. We have not incorporated total deaths and daily new deaths into comparisons. It is because the information about total deaths can be easily obtained from the Total cases, Total Recovered cases and Active cases. The visual illustrations for only one predicted variable, Daily New Cases, using a few methods, have been given here in Figures 2 - 5 due to the space constraint.

It is seen that even though ARIMA is capable of capturing the variables of total cases and total recovered, it fails to capture the daily new cases (Figure 2) and the active cases. The failure for predicting the number of active cases is quite challenging as it is highly uncertain as to how many people would recover and how many would die. The virus spreads uniformly, but it is not fatal to everyone who gets infected. The conditions of co-morbidity is a big factor that plays in the circumstances of death or recovery and thus is indirectly affecting the prediction of active cases.

It is quite evident that the proposed method involving four Transfer-learnt LSTM-RNN models is much efficient in predicting the cases as can be seen in Figure 6.

Table II shows the standard deviation and mean of the relative absolute error percentages (MAE) and the root mean squared error (RMSE) over all the test days. It is seen that the proposed model is the best w.r.t. relative RMSE and the second

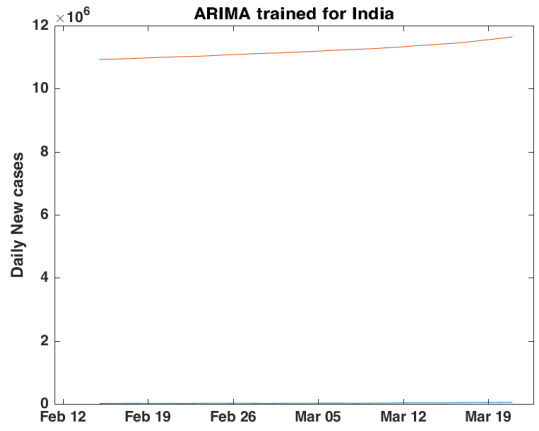

Fig. 2: Daily New cases actual (blue) vs predicted (red) by ARIMA trained on India data.

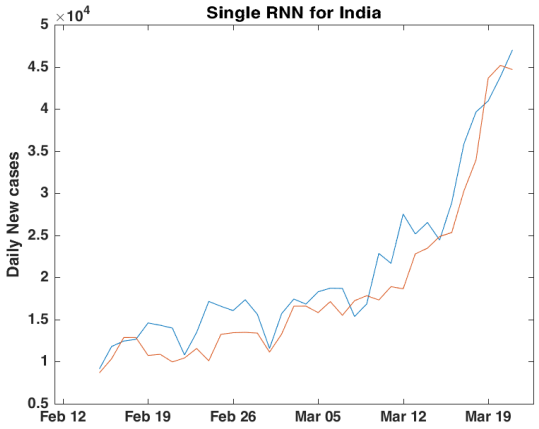

Fig. 3: Daily New cases actual (blue) vs predicted (red) by Standard LSTM-RNN trained on India data.

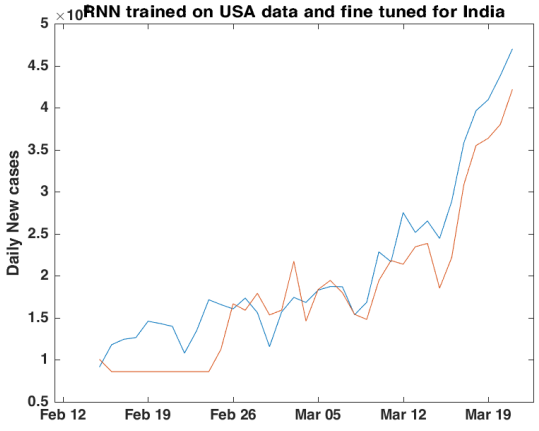

Fig. 4: Daily New cases actual (blue) vs predicted (red) by LSTM-RNN pre-trained on USA data and fine tuning on India data.

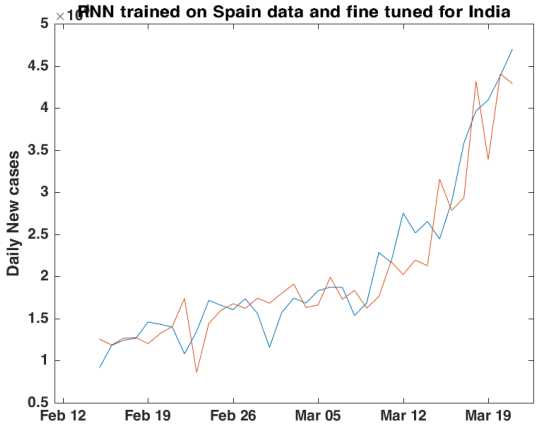

Fig. 5: Daily New cases actual (blue) vs predicted (red) by LSTM-RNN pre-trained on Spain data and fine tuning on India data.

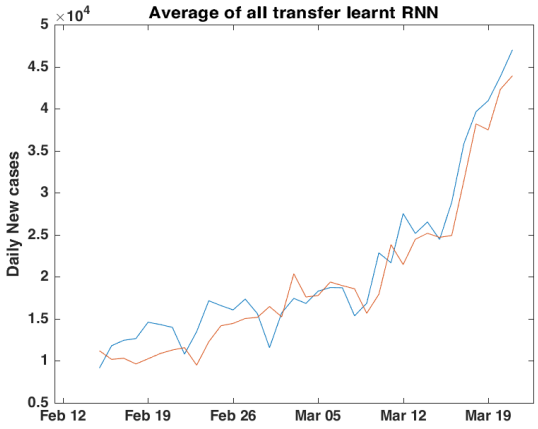

Fig. 6: Daily New cases actual (blue) vs predicted (red) by Proposed model.

best w.r.t. relative MAE; and for both the cases, the model produces least standard deviation indicating more stability. It is to be noted that although ARIMA is showing the best MAE, it is unable to predict the daily cases.

IV. Conclusions and Future Directions

The proposed transfer learning based model has established its supremacy in predicting the COVID-19 infection over the compared models. We would explore different ensembling methods in future and might propose an even stronger prediction model than the traditional LSTM-RNNs which are presently the most popular models used in the literature. The present model can only give a one day ahead prediction of the cases. In future, this work would be extended to give atleast a week ahead prediction. That would be much more helpful in statistical analysis of the existing cases. Moreover, as already said, the present framework is completely data driven and has not incorporated any domain knowledge.

Models	Relative MAE%		Relative RMSE	
	Mean	STD	Mean	STD
ARIMA	3.37	2.64	0.117417	0.097899
Prophet	43.45	8.71	0.929017	0.125566
VAR	36.2	14.98	0.788258	0.260255
Standard LSTM-RNN	5.32	2.62	0.132809	0.095494
LSTM-RNN pre-trained on USA data	5.3	3.53	0.182555	0.135259
LSTM-RNN pre-trained on Spain data	4.32	3.79	0.14098	0.140341
LSTM-RNN pre-trained on Bangladesh data	6.3	3.34	0.203276	0.1339
LSTM-RNN pre-trained on Brazil data	4.55	3.53	0.141086	0.131503
Proposed	**4.04**	**2.53**	**0.109257**	**0.087776**

TABLE II: Comparison of the standard deviation and the mean of relative errors for various models

V. Acknowledgements

Support from the Science and Engineering Research Board (SERB) via sanctioning an ASEAN-India collaborative research project "A Multi Modal Approach to Medical Diagnosis Embedding Deep and Transfer Learning" (IMRC/AISTDFICRD/2019/000151 dated 29 June 2020) is gratefully acknowledged.

References

[1] A. Bhadra, A. Mukherjee, and K. Sarkar. Impact of Population Density on Covid-19 Infected and Mortality Rates in India. *Modeling Earth Systems and Environment*, 7(1):623–629, 2021.

[2] G. Battineni, N. Chintalapudi, and F. Amenta. Forecasting of COVID-19 Epidemic Size in Four High Hitting Nations (USA, Brazil, India and Russia) by Fb-Prophet Machine Learning Model. *Applied Computing and Informatics*, 6(1): 1-10, 2020.

[3] K. Chauhan, Y. Mistry, and S. Mullan. Analysis of Compliance and Barriers for Hand Hygiene Practices among Health Care Workers during COVID-19 Pandemic Management in Tertiary Care Hospital of India- An Important Step for Second Wave Preparedness. *Open Journal of Medical Microbiology*, 10(4):182–189, 2020.

[4] P. Mahajan and J. Kaushal. Epidemic Trend of COVID-19 Transmission in India during Lockdown-1 Phase. *Journal of Community Health*, 45(6):1291–1300, 2020.

[5] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan. Transfer Learning. *Cambridge University Press*, 2020.

[6] R.J. Hyndman and Y. Khandakar. Automatic Time Series Forecasting: The Forecast Package for R. *Journal of Statistical Software Articles*, 27(3):1-22, 2008.

[7] S. Ghosh. Predictive Model with Analysis of the Initial spread of COVID-19 in India. *International Journal of Medical Informatics*, 143:104262, 2020.

[8] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735-1780, 1997.

[9] Y. Gautam. Transfer Learning for COVID-19 Cases and Deaths Forecast using LSTM Network. *ISA Transactions*, 2021.

[10] Z. Sun, H. Zhang, Y. Yang, H. Wan, and Y. Wang. Impacts of Geographic Factors and Population Density on the COVID-19 Spreading under the Lockdown Policies of China. *Science of The Total Environment*, 746:141347, 2020.