Data driven covid-19 spread prediction based on mobility and mask mandate information

Sandipan Banerjee[1] · Yongsheng Lian[1]

Abstract

COVID-19 is one of the largest spreading pandemic diseases faced in the documented history of mankind. Human to human interaction is the most prolific method of transmission of this virus. Nations all across the globe started to issue stay at home orders and mandating to wear masks or a form of face-covering in public to minimize the transmission by reducing contact between majority of the populace. The epidemiological models used in the literature have considerable drawbacks in the assumption of homogeneous mixing among the populace. Moreover, the effect of mitigation strategies such as mask mandate and stay at home orders cannot be efficiently accounted for in these models. In this work, we propose a novel data driven approach using LSTM (Long Short Term Memory) neural network model to form a functional mapping of daily new confirmed cases with mobility data which has been quantified from cell phone traffic information and mask mandate information. With this approach no pre-defined equations are used to predict the spread, no homogeneous mixing assumption is made, and the effect of mitigation strategies can be accounted for. The model learns the spread of the virus based on factual data from verified resources. A study of the number of cases for the state of New York (NY) and state of Florida (FL) in the USA are performed using the model. The model correctly predicts that with higher mobility the cases would increase and vice-versa. It further predicts the rate of new cases would see a decline if a mask mandate is administered. Both these predictions are in agreement with the opinions of leading medical and immunological experts. The model also predicts that with the mask mandate option even a higher mobility would reduce the daily cases than lower mobility without masks. We additionally generate results and provide RMSE (Root Mean Square Error) comparison with ARIMA based model of other published work for Italy, Turkey, Australia, Brazil, Canada, Egypt, Japan, and the UK. Our model reports lower RMSE than the ARIMA based work for all eight countries which were tested. The proposed model would provide administrations with a quantifiable basis of how mobility, mask mandates are related to new confirmed cases; so far no epidemiological models provide that information. It gives fast and relatively accurate prediction of the number of cases and would enable the administrations to make informed decisions and make plans for mitigation strategies and changes in hospital resources.

Keywords COVID-19 · Covid-spread · Prediction · Mobility · Mask mandate · Data-driven · Machine learning · LSTM

This article belongs to the Topical Collection: *Artificial Intelligence Applications for COVID-19, Detection, Control, Prediction, and Diagnosis*

✉ Yongsheng Lian
 yongsheng.lian@louisville.edu

 Sandipan Banerjee
 sandipan.banerjee@louisville.edu

[1] Department of Mechanical Engineering, University of Louisville, Louisville, KY 40292, USA

1 Introduction

The Corona-virus pandemic has infected more than 12.5 million people across the world, of which more than 3.2 million are in the United States. The number of deaths across the globe is more than 559000, of which more than 136000 deaths were in the United States. These values are as of July 10th, 2020 from the website worldometer. info/coronavirus [2]. It can be clearly seen that the United States bear a major brunt of the COVID-19 pandemic. For prediction purposes, the daily new cases is a significant

parameter especially if we consider a correlation with mobility. The daily new cases in the US is shown in Fig. 1 along with a 7-day moving average.

Administrations provided guidelines to stay at home, which resulted in remote work for many people, remote learning for school and college students, cancellation/postponement of many events, and commercial flights. Social distancing guidelines between the population were also delineated across multiple websites, social media and, news media to make people aware of mitigation steps like maintaining a 6ft distance with another being, wearing a form of face covering or masks to limit the virus spread from infected to non-infected people. Closure of non-essential facilities and business were also mandated by multiple nations while keeping only essential services open. All these mitigation strategies resulted in a reduction of cases, however there is a delay attached to the decrease. This delay will be discussed later in the paper. Since the number of cases were successfully reduced by social distancing measures and a mask mandate from the administrations, it is imperative to have a correlation between the overall new confirmed cases each day to the amount of mobility that people exhibit and the mask mandate information. This could lead to a prediction of the daily new confirmed cases based on the mobility and mask mandate information.

Prediction of confirmed cases of COVID-19 has been studied by using mathematical-epidemiological models. Typically, these models are named 'SEIR' or in some cases 'SIR' [9, 19, 20]. The abbreviations arise from the fact that the models divide the entire population into 'Susceptible', 'Exposed', 'Infected' and 'Recovered' categories. The model consists of equations that govern the rate at which the values of these categories change over time. Mostly the rate of change for these sub-divisions are estimated from available data of either the same disease or data from previous similar diseases like SARS or MERS [27]. Although in some cases these models provide reasonable approximations, there are some drawbacks. These models assume that the disease would spread in a certain way which is defined by three or four 1st degree equations depending on whether SEIR or SIR model is used. The models assume homogeneous mixing among the populace, some predictions claim that up to 90% of the population could eventually become infected unless the contact rate is minimized by social distancing measures [24]. Most models use reported case numbers as the input values for the equations, but this may not be accurate as testing rates vary across places, and delay in testing results could be observed in multiple cases. This would erroneously estimate an under-reported contact rate. Estimating an accurate contact rate parameter in the model is a challenge. The incubation rate could also be a source of erroneous prediction for these models unless accurate estimation is provided. For this strain of virus, the incubation period could be in the range of 2-14 days as per the Centers of Disease Control and Prevention website [1].

It has been shown that Machine Learning models have been effective in their predictive capabilities across multiple fields. Computer Vision [26], Natural Language Processing [28], Software Engineering [12], Direct Numerical Simulation (DNS) of Turbulence [21] (This is significant since DNS is computationally very expensive specifically for

Fig. 1 Daily new confirmed cases in US as a function of time along with seven day moving average
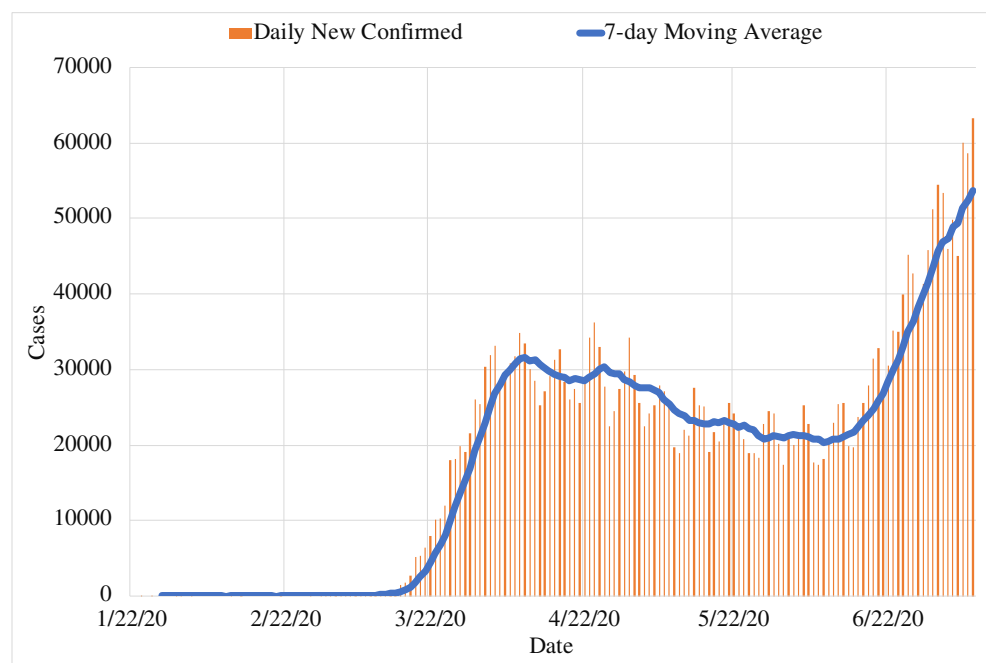
turbulent flows [8]). Some hybrid SEIR-Data driven models have also been reported in the literature and online pre-prints [23, 27] for COVID spread prediction. One of the models tries to estimate the contact rate in the 'SEIR' model from mobility data [23]. However, the base drawbacks of a 'SEIR' based model still persists in these models too. Further, none of the models consider the effect of mask mandate which our results predict to have a large impact on the spread of the virus. Data driven prediction of the spread of COVID-19 have also been reported. [25] used LSTM for predicting the number of cases in India, but they did not include the effect of social distancing measures by considering the actual mobility of the population, neither did they consider the effect of mask mandate on the spread of the virus.

In this paper, we report a completely data driven approach to predict the daily new confirmed cases based on the actual mobility of the population and actual mask mandate information. Unlike the SEIR based models, no prior assumption is used to account for the spread of this work. The number of new confirmed cases on a certain day is considered to be dependent on: (a) the mobility data from a few days before, (b) the mask mandate information, and (c) the confirmed cases from previous few days. We use a LSTM neural network model to predict the number of new confirmed cases up to 75 days in the future. The mobility data is gathered from cell-traffic information provided by Google LLC [3]. The confirmed cases information is gathered from Johns Hopkins Center for Systems Science and Engineering [4, 5]. The remainder of the paper is divided into the following sections: A description of the method, data analysis, and the model is provided in Sections 2, 3, and 4. Results and discussion are provided in Section 5, followed by a conclusion in Section 6.

2 Method

We use LSTM neural networks to predict the 7-day average daily new confirmed cases based on mobility data from cell phone traffic information and mask mandate information. In this section, we provide a brief explanation of a LSTM cell.

LSTM is a type of Recurrent Neural Network (RNN) capable of learning long term temporal dependencies in a time-series. This neural network was introduced by [17]. LSTMs have an advantage over regular RNNs as it does not exhibit the vanishing gradient and exploding gradient issues which are typical drawbacks associated with RNNs.

In an LSTM cell, there are three types of gates, each performing a specific task.

– **Forget Gate:** Forget gate decides which information the network should keep and which it should forget

or remove. The equation solved by the forget gate is:

$$f_t = \sigma \left(W_f \cdot [h_{t-1}, x_t] + b_f \right) \tag{1}$$

where W_f are the weight values, and b_f are the bias values. x_t corresponds to the inputs at time t, and h_t corresponds to the output.

– **Input Gate:** Input gate decides what new information the network should add to the cell state. Once that is done, the old cell state is updated with the new values. The calculation for the input gate and the consequent updating of the cell state is done by:

$$i_t = \sigma \left(W_i \cdot [h_{t-1}, x_t] + b_i \right) \tag{2}$$

$$\tilde{C}_t = \tanh \left(W_C \cdot [h_{t-1}, x_t] + b_C \right) \tag{3}$$

and,

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

– **Output Gate:** Output gate decides what should be the output of the cell, the output depends on both the input and cell state values. The calculations are performed by:

$$o_t = \sigma \left(W_o \cdot [h_{t-1}, x_t] + b_o \right) \tag{5}$$

$$h_t = o_t * \tanh (C_t) \tag{6}$$

3 Data analysis

Three sets of time-series data are used in the model. First is the average mobility data which is calculated from cell-phone traffic information of five different categories of public areas. The areas are Retail and Recreation, Grocery and Pharmacy, Parks, Transit Station, and Workplaces. The data from Google reports a percentage change from pre-COVID time cell-traffic for each day. As a first step, the mean percentage change is calculated between all five categories. Next, the mean percentage change is converted to an actual value.

$$v = 1.0 + pc/100.0 \tag{7}$$

where v represents the final calculated value, and pc represents the percentage change from pre-COVID times. Next, a seven day average of the mobility information is calculated to smooth the data. Then the 7-day averaged data is scaled so that all values lie between 0 and 1 as this helps the model to learn faster. Similar conversion has also been applied by [23]. The second set of time-series data used for training is the 7-day average daily new confirmed cases for the previous few days. The reason for using the previous data of cases is a that there is a time-dependent trend in the daily new cases which can be shown by a stationarity test.

3.1 Stationarity test: augmented dickey fuller test

In order to check the time-dependence of the daily new cases data, we perform a stationarity analysis by using the well-known "Augmented Dickey Fuller (ADF) test". As per the test, if the p-value of a time-series is less than 0.05, then the data is stationary, and if the p-value is greater than 0.05 then the time-series is non-stationary. Stationary time-series means there is no time-dependent trend in the data, and non-stationary means there is a time-dependent trend. Typically for stationary time-series conventional machine learning regression algorithms like ARIMA, ARIMAX are used. LSTMs are generally favorable for non-stationary data [22]. We applied the ADF test to the daily new cases time-series and found the p-value to be 0.1 which concludes that the time-series is non-stationary, hence shows that there is a time-dependent trend. This can be considered as the reason for using LSTM network as well as the reason for using the previous few days of daily new cases as inputs along with mobility data in order to predict the daily new cases for the next day.

Finally, the mask mandate information is provided as a binary option. 0 is set for dates when no mask mandate was applicable, and 1 is set for dates when mask mandate was applicable in that state. This is the third and final time-series of input data provided to the model for training purposes.

3.2 Problem formulation and training data-set

From the available data, it is evident that a reduction in mobility would result in fewer interactions between people, which would lead to reduced transmission resulting in fewer new cases and vice-versa. However, the response in the change of cases due to change in mobility is expected to exhibit a lag or delay. Once awareness about the rising number of cases increases among the populace, along with mitigation procedures enacted by administrations, the mobility starts to reduce sharply. The lag could be attributed to a combination of two main reasons: the virus incubation period (about 2-14 days as per [1]) and a delay in testing and reporting of cases (about 25 days). Since the virus incubation period is 2-14 days, the symptoms can show any time in that time window, so we chose 15 days of data prior to the 25 days of lag for the model to be trained on in reference to the mobility and immediate prior 15 days in terms of the cases.

The daily new case prediction is treated as a supervised learning problem. It is considered that the number of new cases averaged over 7 days for each particular day would depend on the number of new cases in the past 15 days, the mobility information from 15 days prior to the number of lag days, and the mask mandate information. The problem for the neural network model is formulated as: Given the

mobility and mask mandate at day $= t - (15 + n_{lagdays}), t - (14 + n_{lagdays}), \cdots, t - (1 + n_{lagdays}$, and the number of new cases at $(t - 15), (t - 14), \cdots, (t - 1)$, predict the number of new cases at day $= t$. This would make the total number of features to be (15 days mobility + 15 days case count +15 days mask mandate = 45) on which prediction of one day's case count would depend on. The neural network model would assign weights to each of these features and would learn from the data. The functional mapping between the provided input and the estimated output would be approximated by changing the weight values and then comparing the estimated output to the actual output of daily number of cases. The goal is to keep the mean squared error or MSE between the model estimation of the number of cases at day $= t$ (also called prediction), to the actual value (also called label), to be minimum. MSE is the loss function used in this model. Their are two versions of the model implemented. For Version 1, the model is trained on 129 days (Feb 15th - June 24th), and validation forecast is made for the next twenty days (June 25th-13th July). Prediction is made for each day of the twenty days using the non-cascading option where the actual data of the previous days are used as input. Version 1 is only for validation purposes. The only difference between Version 1 and Version 2 is that all 149 days (Feb 15th-July 13th) are used as training sample for Version 2, and prediction for 75 days into the future is made (14th July-25th Sep).

4 Model architecture and parameters

An overview of the neural network model used in this work is provided in Fig. 2.

Although the first layer is shown to be an LSTM layer, in reality there is a zeroth layer as well. A masking layer is applied before the LSTM layer, but it not shown on the architecture as the masking layer does not perform any learning. Since data from 15 prior days is used for training, that information is not available in the data-set for the initial few training samples. For this the conventional approach of padding those unavailable data has been used. The padded values are typically selected such that they do not occur anywhere on the actual training set. Since the training data was scaled to be between 0 and 1, a padding value of -1 was chosen. The benefit is that the masking layer prohibits the model to learn or train on those samples which have values of -1. Thus, both the input shape requirement of the data for the LSTM is maintained, and training on padded values is also prohibited.

The first layer in the model is an LSTM layer with 100 units, followed by another LSTM layer with 50 units. Both LSTM layers have been applied with a recurrent dropout of 0.5 to prevent over-fitting. The next layer in the model is a
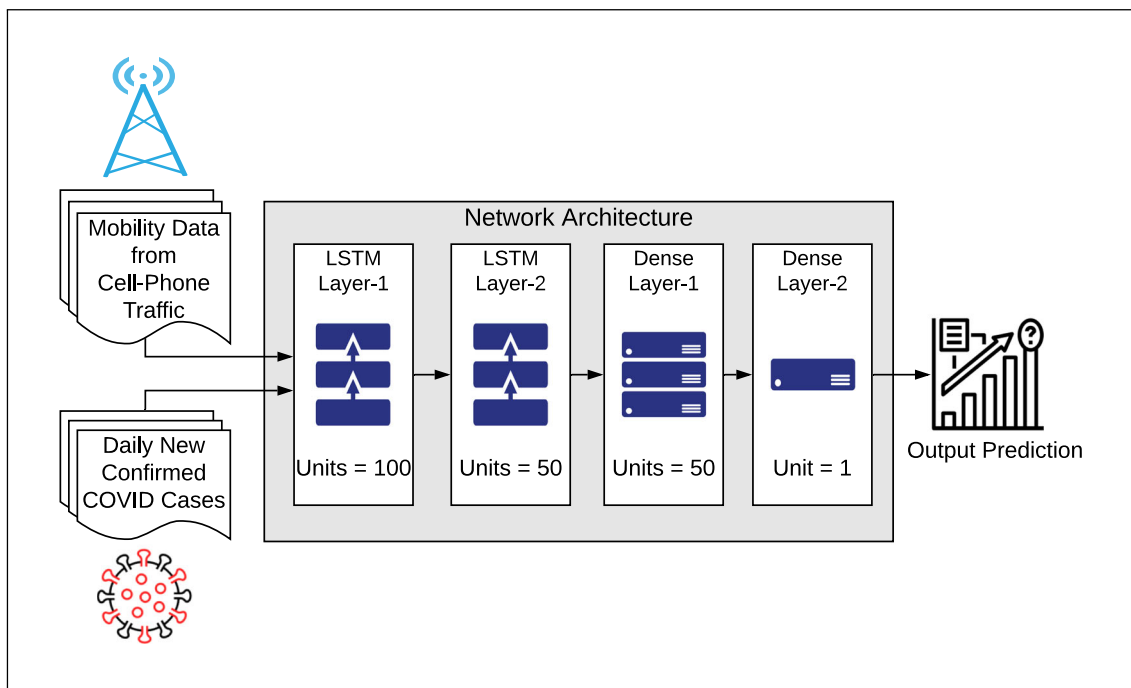
Fig. 2 Architecture of network

Dense layer with 50 units. Dense layer is the conventional 'MLP' or Multi Layer Perceptron. The final layer consists of another dense layer with one unit corresponding to the output of each sample. Since the training data-set is relatively small, only one batch is used for all the samples. The well-known 'Adam' optimizer has been used with an initial learning rate of 0.001, and a decay of 1e-6 has been applied. Adam was introduced by [18]. As mentioned earlier, mean squared error loss function has been employed for the model. Hyper-parameter tuning was performed to check for optimum performance. Both larger and smaller learning rates did not seem to produce a significant change in prediction. The above mentioned settings for the hyper-parameters produced the most optimum performance. The model has been created in Python using Tensorflow [7] and Keras [13] framework.

5 Results and discussion

In this section we report the results predicted by our model. The results for NY and FL are provided in Section 5.1. Prediction using our model and comparison of our results with other models such as ARIMA-based model of [16] is provided in Section 5.2 for Italy and Turkey.

5.1 Results for NY and FL

We start with the case for state of NY. Once the model is trained on samples from 129 days, prediction is made for the next twenty days for the 7-d average case count on each day. The prediction for the twenty days along with the actual NY data are shown in Fig. 3.
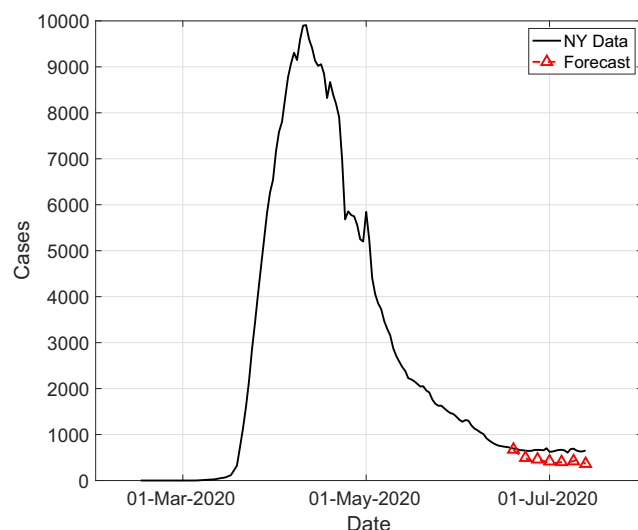

Fig. 3 Validation Forecast for NY

Figure 3 shows that prediction is in good general agreement with the actual prediction. This shows that the functional mapping performed by the model is accurate. It should be noted that the prediction for the validation is performed with Version 1. Next, in order to perform forecast of next 75 days, model is trained for all 149 days (Version 2). Prediction for different mobility values of 0.25 and 1.0 are made with both with mask (value=1) and without mask (value=0) options. Here 1.0 mobility corresponds to pre-COVID mobility. 75 day forecast results for 7-d average daily new cases are shown in Fig. 4a. Total number of cases prediction are shown in Fig. 4b.
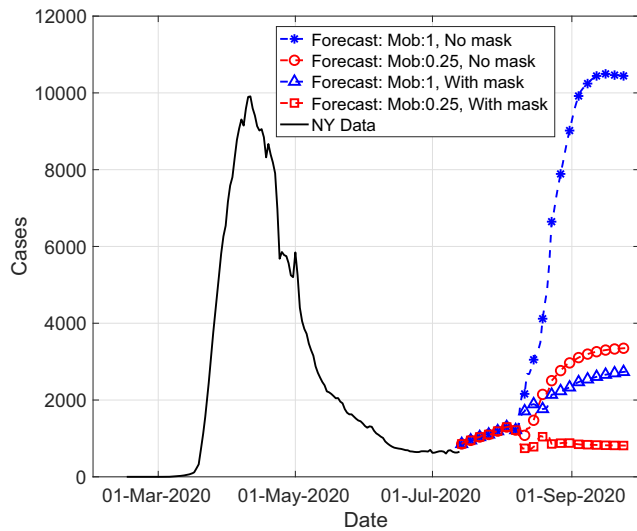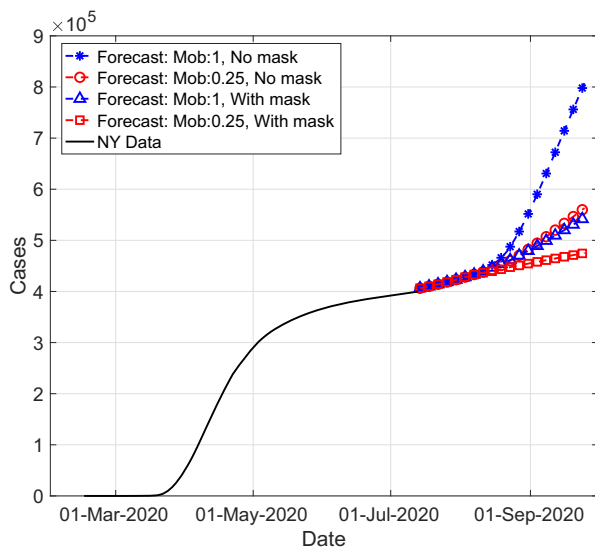
The results are in agreement with leading epidemiologists, that with decrease in mobility, the virus spread will decrease. It should be noted that the initial similar prediction for all different cases is due to the number of lag days. The change in mobility will show in the change in cases after the lag period, the initial similar prediction corresponds to this lag. The model also accurately predicts that using masks reduces the spread, as can be seen from Fig. 4a & b, that for the same mobility, wearing masks produces much less cases. However, the most significant conclusion that can be drawn from the projections is that the spread is lower for mobility of 1.0 with mask than for mobility of 0.25 without mask. This means that if there is mask mandate, the spread will be lower even with a certain degree of mobility, which would enable the economy to be opened to a significantly higher extent with out a significant rise in cases.

The model which has been pre-trained on the NY data is then used to further train on the data for state of FL. However, since FL has no mask mandate option at the time of writing this paper, the mask mandate option cannot be used for FL, due to lack of training data. For FL, the correlation with the mobility is provided. Similar to the NY model, first the FL model is trained on 129 days of training data which ranges for the same dates as in the NY model. The validation prediction of the FL model along with actual FL data is provided in Fig. 5. It can be observed that the model makes accurate prediction of the cases to rise sharply for the twenty days.

Similar to the NY model, Version 2 is used for forecast of FL. Data from all 149 days for FL is used to train the model, and forecast is made for 75 days in future for mobility rates of 0.25 and 1.0. The 75 day forecast results for 7-d average daily new cases are shown in Fig. 6a. Total number of cases prediction are shown in Fig. 6b.
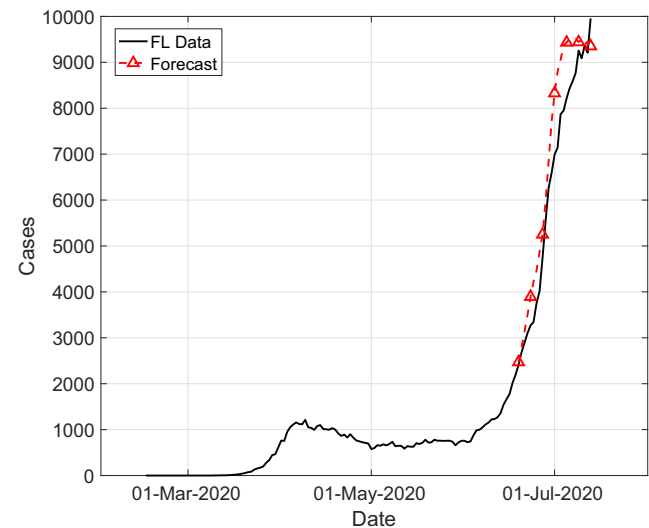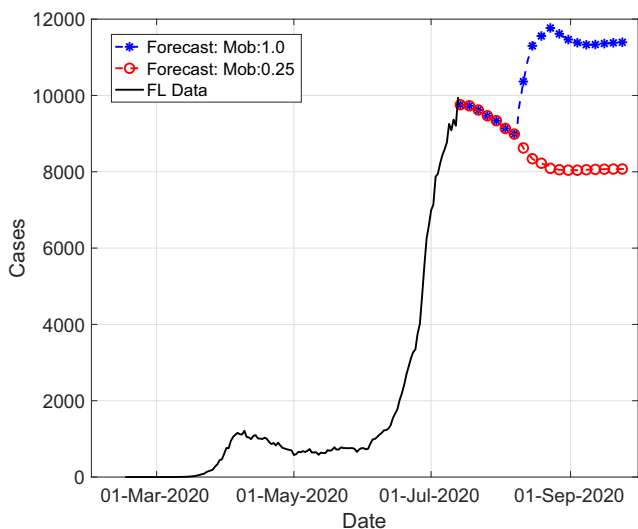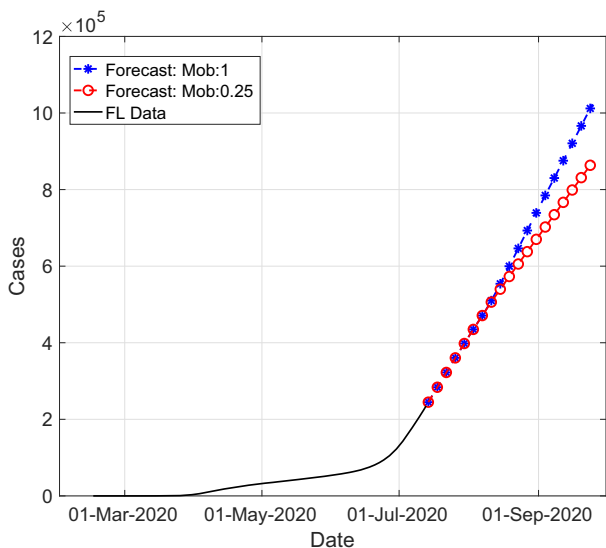

(a)

(b)

Fig. 4 Forecast for 7-d averaged new cases for 75 days **a** and Forecast for total cases for 75 days **b** for different mobility values and mask options for NY

Fig. 5 Validation Forecast for FL

(a)

(b)

Fig. 6 Forecast for 7-d averaged new cases for 75 days **a** and Forecast for total cases for 75 days **b** for different mobility values and mask options for FL

As seen in the NY model, the FL model projection shown in Fig. 6 also reports initial similar prediction for both mobility values due to the lag period for the cases to reflect the change in mobility. The model is in general agreement that higher mobility will cause more cases, and lower mobility will cause less. The important factor that

Table 1 RMSE and average percent difference for NY and FL

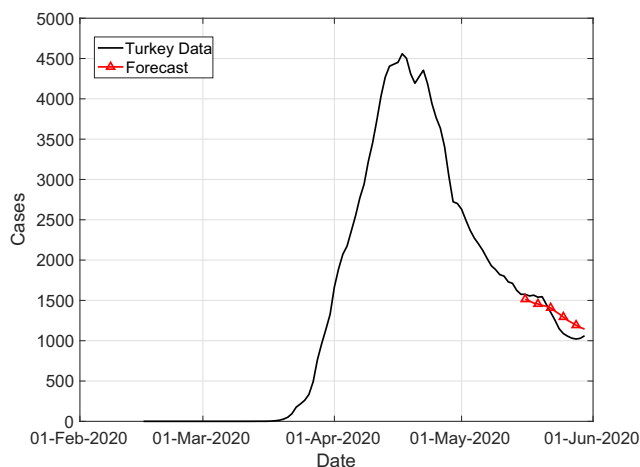| State	RMSE	Average Percent Difference
NY	2554.16	0.5
FL	10528.56	5.17

Fig. 7 Validation forecast for Turkey

needs to be considered here is that since there is no mask mandate in FL, the drop in cases due to lower mobility is not very sharp. This has been verified by the NY projections. Next, we include the RMSE (Root Mean Square Error) and percent differences for NY and FL. Here RMSE and percent difference are calculated based on the total confirmed cases calculated from the prediction by the model and the ground truth data of confirmed cases. RMSE is calculated as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad (8)$$

where $\hat{y}$ is the predicted value, y is the actual value, and n is the number of samples. The results for RMSE and percent difference for NY and FL are reported in Table 1.

RMSE for FL is higher than NY as the prediction dates for the FL observed a sharp increase in the actual data. This can be observed in Fig. 5. The model does a good job with predicting a sudden increase with a percent difference
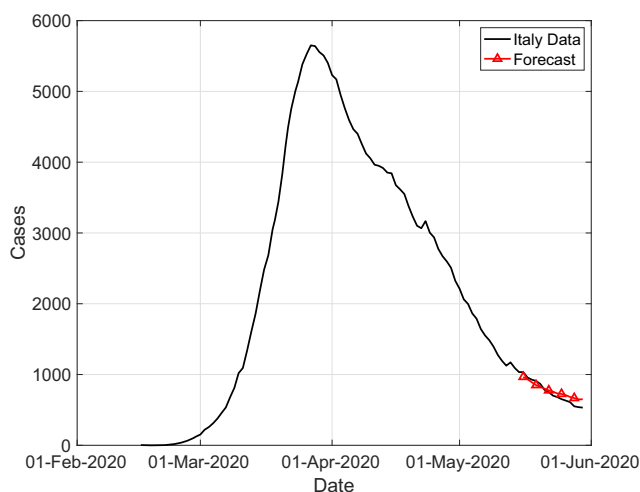

Fig. 8 Validation forecast for Italy

Table 2 RMSE comparison of our model with [16] for Italy and Turkey

| Method	RMSE–Turkey	RMSE–Italy
Our model (LSTM)	710.5	229.4
Hernandez et al. (ARIMA based)	1892.3	566.8

of about 5%. The RMSE for NY data results in a percent difference of 0.5%.

## 5.2 Comparison with a another model for predictions of other countries

Different models have been proposed in the literature for predicting COVID-19 spread such as Fong et al. [14, 15],

Hernandez et al. [16]. It should be noted that Catelli et al. [10, 11] performed interesting studies on Italy dataset. We compare our results for Italy and Turkey with the results of [16], who used an ARIMA-based model. In order for the comparison to be accurate, we use data until May 28th, 2020 which is same as used by Hernandez et al. [16]. First we provide the prediction comparison of 15 days for each countries. It should be noted that for the dates we are comparing (i.e. until May 28th) a mask mandate policy was not yet active in either Italy or Turkey. Prediction for 15 days in comparison to the ground truth for Turkey and Italy are provided in Figs. 7 and 8.
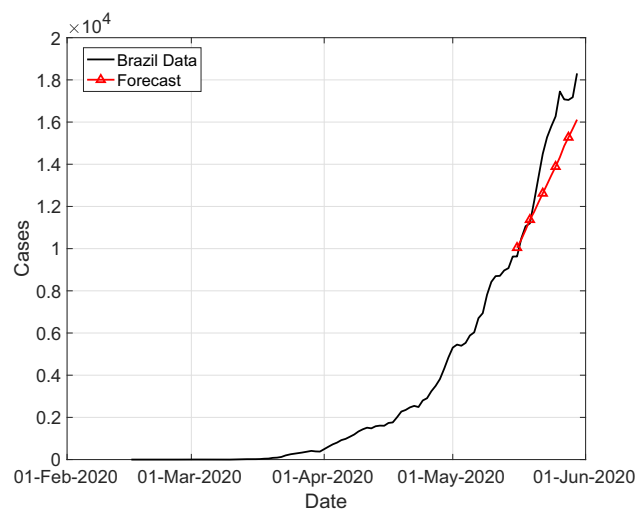
As can be seen from Figs. 7, and 8, the model predictions are in good agreement with the ground truth. Similar results were also reported by Hernanzdez et al. [16], using an ARIMA based model. We compare our model's RMSE values with their results in Table 2.
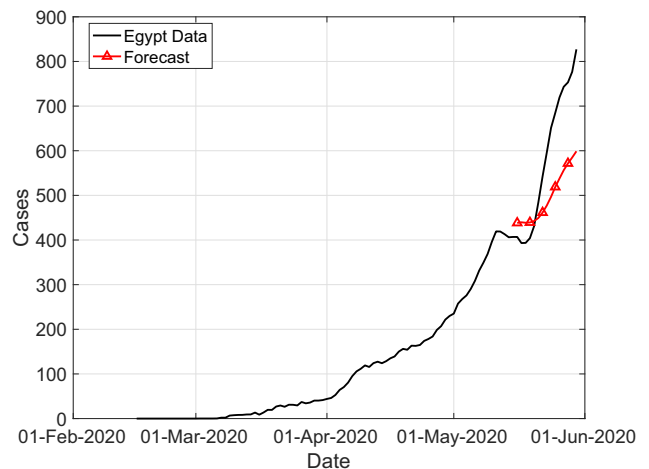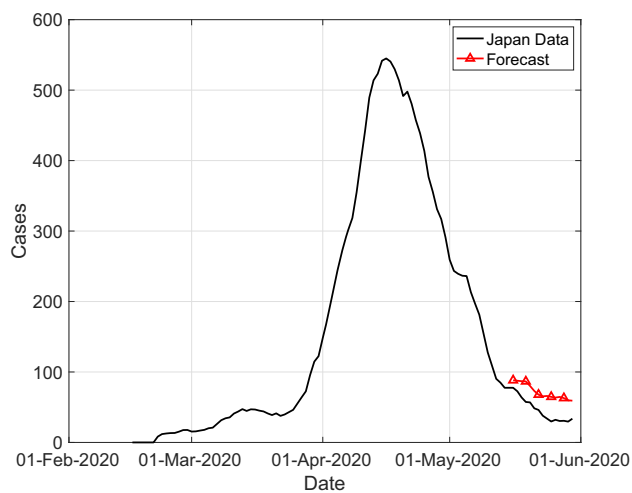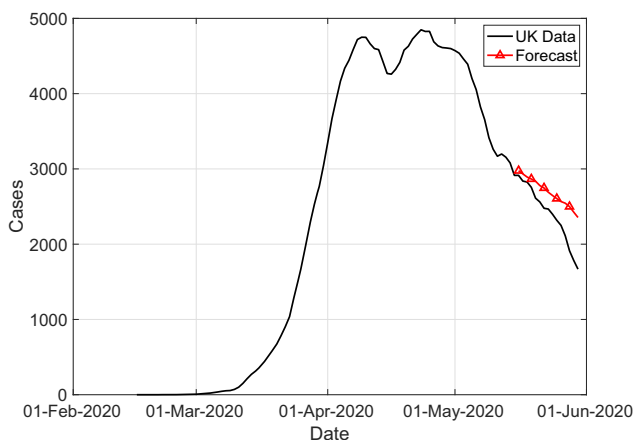

(a)
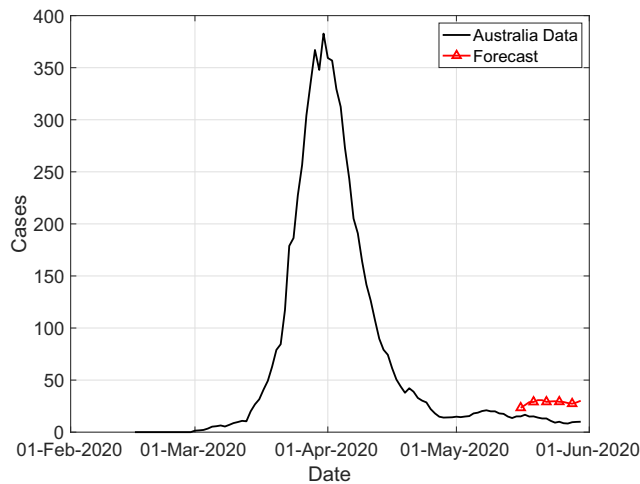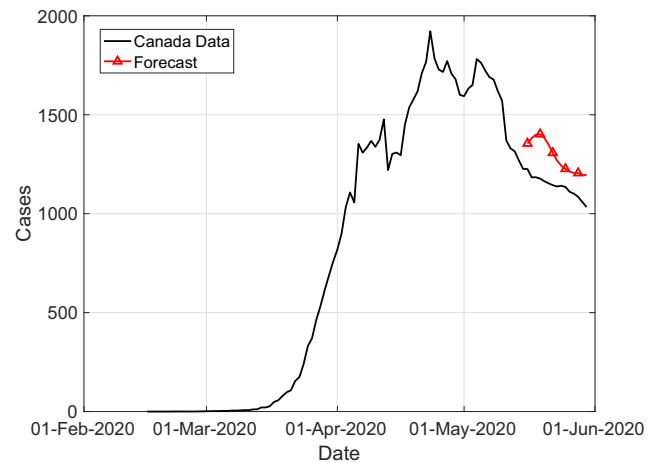
(b)

Fig. 9 Prediction of 7-day average of daily new cases for **a** Australia and **b** Brazil

(a)

(b)

Fig. 10 Prediction of 7-day average of daily new cases for **a** Canada and **b** Egypt

(a)

(b)

Fig. 11 Prediction of 7-day average of daily new cases for **a** Japan and **b** UK

From Table 2, it can be observed, that our model prediction is more accurate than [16] for both Italy and Turkey due to considerably lower RMSE values.

Table 3 RMSE comparison of our model with [16] for different countries

Country	RMSE-Our model (LSTM)	RMSE- Hernandez et al. [16] (ARIMA)
Australia	148.86	637.01
Brazil	10636.22	94302.48
Canada	1641.90	8863.83
Egypt	570.07	2067.55
Japan	236.79	6543.16
UK	2198.53	24835.98

Additionally, we include prediction results of total confirmed cases for same period of 15 days (May 12th-May 27th), as in Hernandez et al. [16] for six other countries, namely: Australia, Brazil, Canada, Egypt, Japan and United Kingdom (UK). Based on [6], none of the countries mentioned above had a mask mandate during the range of days of training data as well as prediction data. The prediction results for Australia and Brazil are provided in Fig. 9.

The prediction results for Canada and Egypt are provided in Fig. 10.

The prediction results for Japan and UK are provided in Fig. 11.

Prediction results for all of the countries are in good agreement with the actual data. The RMSE comparison of our model with that of Hernandez et al. [16] is given in Table 3.

The results in Table 3 shows that our LSTM based model reports considerably lower RMSE for all the countries in question. This shows that the LSTM based model learns the temporal correlations in the data. It also shows that mobility data is a significant input parameter for predicting the spread of the COVID-19 irrespective of the country.

6 Conclusion

A novel data driven approach is used to predict the spread (daily new cases) of Corona-virus using mobility data from cell traffic information and mask mandate information. In contrary to epidemiological models, this data driven approach does not use any pre-defined model specific equations for the prediction. We propose a LSTM based model that incorporates mobility and mask mandate information, which cannot be done by the existing epidemiological models. The proposed model does not assume homogeneous mixing among the populace for the spread prediction, instead it is trained on factual data gathered from verified resources. LSTM neural network has been implemented to capture the long term temporal dependencies in the data. Model predictions show that with a mask mandate implemented, the virus spread would be reduced, and vice-versa. Results are in agreement with opinions from medical experts that a reduction in mobility would reduce the spread. The model also predicts that a mask mandate would produce lower cases for a mobility of 1.0 (i.e. pre-COVID mobility) than the case of no mask mandate with a lower mobility of 0.25. This would correspond to a reduction by hundreds of cases per day with a higher mobility (for the state of NY) depending on the population number. Results for states of NY and FL confirm the efficacy of the model by reporting a low percentage difference between prediction and real data.

Further comparison with an ARIMA-based model shows that our model is more accurate for Australia, Brazil, Canada, Egypt, Italy, Japan, Turkey, and the UK due to lower values of RMSE.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Centers for disease control and prevention https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html
2. Corona virus case counts https://www.worldometers.info/coronavirus/country/us/
3. Google mobility data https://www.google.com/covid19/mobility/
4. Johns Hopkins center for systems science and engineering https://systems.jhu.edu/
5. Johns Hopkins university coronavirus map https://coronavirus.jhu.edu/map.html
6. Mask mandate information https://www.aljazeera.com/news/2020/8/17/which-countries-have-made-wearing-face-masks-compulsory
7. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X (2016) Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on operating systems design and implementation, OSDI'16. USENIX Association, USA, pp 265–283
8. Banerjee S, Ayala O, Wang LP (2020) Direct numerical simulations of small particles in turbulent flows of low dissipation rates using asymptotic expansion. In: 5th thermal and fluids engineering conference (TFEC), pp 659–668. https://doi.org/10.1615/TFEC2020.tfl.032308
9. Binti Hamzah FA, Hau C, Nazri H, Ligot D, Lee G, Shaib M, Zaidon U, Abdullah A, Chung M, Ong C, Chew P (2020) Coronatracker: world-wide covid-19 outbreak data analysis and prediction. https://doi.org/10.2471/BLT.20.255695
10. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M (2020) Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. Applied Soft Computing 97:106779. https://doi.org/10.1016/j.asoc.2020.106779 https://www.sciencedirect.com/science/article/pii/S1568494620307171
11. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M (2021) A novel covid-19 data set and an effective deep learning approach for the de-identification of italian medical records. IEEE Access 9:19097–19110. https://doi.org/10.1109/ACCESS.2021.3054479
12. Chatterjee P, Damevski K, Pollock L (2021) Automatic extraction of opinion-based Q&A from online developer chats. Proceedings of the 43rd International Conference on Software Engineering (ICSE)
13. Chollet F et al (2015) Keras. https://keras.io
14. Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E (2020) Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Appl Soft Comput 93:106282
15. Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E (2020) Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak. arXiv:2003.10776
16. Hernandez-Matamoros A, Fujita H, Hayashi T, Perez-Meana H (2020) Forecasting of covid19 per regions using arima models and polynomial functions. Appl Soft Comput 96:106610
17. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
18. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization
19. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM (2020) Early dynamics of transmission and control of covid-19: a mathematical modelling study. medRxiv. https://doi.org/10.1101/2020.01.31.20019901
20. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). Science 368(6490):489–493. https://doi.org/10.1126/science.abb3221
21. Mohan AT, Gaitonde DV (2018) A deep learning based approach to reduced order modeling for turbulent flow control using lstm neural networks. arXiv:1804.09269
22. Preeti, Bala R, Singh RP (2019) Financial and non-stationary time series forecasting using lstm recurrent neural network for short and long horizon. In: 2019 10th international conference on computing, communication and networking technologies (ICCCNT), pp 1–7
23. Soures N, Chambers D, Carmichael Z, Daram A, Shah D, Clark K, Potter L, Kudithipudi D (2020) Sirnet: understanding social distancing measures with hybrid neural network model for covid-19 infectious spread
24. Thomas T, Benjamin J, Jha A (2020) American hospital capacity and projected need for covid-19 patient care. Health Affairs Blog. https://doi.org/10.1377/hblog20200317.457910
25. Tomar A, Gupta N (2020) Prediction for the spread of covid-19 in India and effectiveness of preventive measures. Science of The Total Environment 728:138762. https://doi.org/10.1016/j.scitotenv.2020.138762
26. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: a brief review. Comput Intell Neurosci 2018:1–13
27. Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, Liu P, Cao X, Gao Z, Mai Z, Liang J, Liu X, Li S, Li Y, Ye F, Guan W, Yang Y, Li F, Luo S, Xie Y, Liu B, Wang Z, Zhang S, Wang Y, Zhong N, He J (2020) Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. J Thoracic Disease 12(3):165–174. http://jtd.amegroups.com/article/view/36385
28. Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine 13(3):55–75

Sandipan Banerjee is a Ph.D student.

Yongsheng Lian is a Professor.