The Forecasting of the 2019 Novel Coronavirus (COVID-19) Epidemic in Canada

Shuoyuan Chen[20683201] and Xinyi Ma[20814843]

Department of Electrical and Computer Engineering
University of Waterloo
200 University Ave W, Waterloo, Canada
{s487chen,x258ma}@uwaterloo.ca

Abstract. The COVID-19 pandemic is an ongoing health threat to people both globally and in Canada. Reliable prediction on the trend of the disease can be much helpful in governors' decision making. In this paper, we demonstrate the forecasting of the number of confirmed cases and deaths with the aid of the mobility of given provinces in Canada. a mobility-assisted long short-term memory (LSTM) model and a multilayer perceptron (MLP) model are designed and implemented. Their performances and predictions are compared with a basic SEIR model to demonstrate their significance.

Keywords: COVID-19 · LSTM · MLP · SEIR · forecasting.

1 Introduction

Since December 2019, the global propagation of the COVID-19 virus has caused great casualties and financial loss. COVID-19 has a long incubation period of 14 days [10], during which time the patient can be contagious. It makes the number of potential carriers hard to track and thus requiring the assistance of mathematical models to estimate the outbreak potential in certain area, which is quite accurate at the start of the epidemic. However, the real-world intervention such as lock-down and quarantine cause the true data to deviate further and further from the predictions of mathematical models. Machine learning is known to be able to find out hidden patterns in data and give predictions, which catches many attentions during the pandemic.

Recently, deep learning researchers have proposed various algorithms to predict the countrywide trend of daily new cases, whose results are reasonably accurate [1–3, 5]. The studies we looked into are all country-level researches, built on reliable statistics such as population crossing the border. The accurate values cut the percentage error and ease the training of the neural network. The large population of a country can also tolerate big numbers of errors, which guarantees reliability. The predictions are adopted and the following Canada-wise lock-down measures successfully brought the spread of the virus in Canada under control.

Aside from country-wise measures, the province-wise predictions are also urgent but often neglected in the literature. It is because the data such as the number of people crossing the border would be difficult to access if the algorithms are to be applied to analyzing a province alone. Moreover, the relatively low population restricts the room for ambiguity, meaning the error by untested patients has a greater impact on the model than the country case. The date of re-opening can be different from east to west, and the date as early as possible can minimize the already-done financial damage. Therefore, a robust model capable of predicting COVID-19 for a small population becomes essential.

My team employs two custom neural networks including long short-term memory (LSTM) model and multi-layer perceptron (MLP) model to forecast the trend of the number of COVID-19 confirmed cases and deaths among provinces in Canada and compare their performance. SEIR model is chosen as base line method to evaluate the efficiency of the proposed two models. This study is closely related to our current daily life and can have a positive impact on the decision making of local government.

2 Literature Review

One classic mathematical model used in disease propagation study is the SEIR model with the assumption that every two people have equal chances to meet each other. It has a long history of receiving modification to fit real-world data [4, 7, 8], majoring incorporating the role of intervention. However, since it still requires pre-consideration of potential parameters before building the model, it is still rigid and difficult to tune. Neural networks (NNs) are known to be perfect black-box methods for time series prediction and function regression. NNs do not require pre-definition of parameters, which saves many headaches and often gives better results than SEIR. Therefore, we treat SEIR model as the baseline forecasting method in this report.

The machine learning (ML) community gives enthusiastic responses to the trend prediction of this epidemic and multiple papers have been published analyzing the factors influencing the spread of the virus [2, 3, 5, 9]. According to sutdies [2, 5, 7], lock-down is a major factor that slows down the propagation and considering that in NNs would increase forecasting accuracy. Dandekar [3] reported the important role played by the quarantine and isolation measures in the global COVID-19 spread. His research focuses on four locales: Wuhan, South Korea, the United States of America, and Italy, successfully simulating the outbreak and slowdown of the disease in these countries using NNs.

Another recent paper [1] presents a comprehensive analysis on multiple machine learning models to predict the COVID-19 outbreak. Two models stands out, which showed promising results, one being multilayered perceptron (MLP) and the other being the adaptive network-based fuzzy inference system. The two ML models hold a high generalization ability for long-term prediction. Therefore, this study suggests machine learning as an effective tool to model the outbreak.

Our LSTM model and MLP model can find similar applications in literature [1, 9] but they have been modified to achieve best performances in COVID-19 prediction inside Canada. We also incorporated the google community mobility data [11] in public area to enhance accuracy, which has not been systemically studied so far. We also attached multi-variant regression to the SEIR model, which help producing smooth fit.

3 Algorithms

3.1 SEIR Model

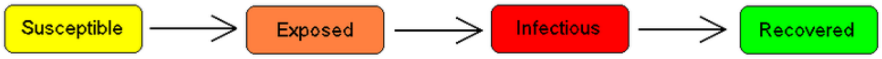

Fig. 1. SEIR Model

SEIR is one of the basic compartmental models used by mathematicians to simulate the spread of infectious diseases in a closed environment. The population is assigned to four groups including, susceptible (S), exposed (E), Infectious (I), recovered (R), corresponding to different stages of the disease as shown in **Figure.2**.

The SEIR model is a batch of ordinary differential equations (ODE), which is best shown as follows,

$$\begin{aligned}\frac{dS}{dt} &= \Lambda - \mu S - \frac{\beta I S}{N}\\ \frac{dE}{dt} &= \frac{\beta I S}{N} - (\mu + a)E\\ \frac{dI}{dt} &= aE - (\gamma + \mu)I\\ \frac{dR}{dt} &= \gamma I - \mu R.\end{aligned}$$

assuming that $S+E+I+R=N$ and the average incubation period is a^{-1}, and also assuming the presence of vital dynamics with birth rate Λ equal to death rate μ.

SEIR model is highly customizable and researchers have added parameters such as age ratio and gender ratio to improve accuracy in estimating disease spread and show how different public health precautions may affect the outcome of the epidemic. Here, we first use multi-variant regression to fit the daily mobility data in public area of each province, which gives an estimation of virus exposure. The data is then used as one optional reproduction number, aside from Hill decay function and preset constant number, in SEIR solving. The SEIR model is essentially optimization of parameters in ODEs to minimize the loss. The loss is given by the mean square error (MSE) between true and estimated number of infectious patients.

Though many adjustments are attached to the model, it is still a rigid mathematical model, difficult to be applied to real-life events. It assumes complete mixture of population and therefore cannot handle population isolation caused by lock-down and quarantine. It serves as a benchmark to evaluate our custom models.

3.2 LSTM

The LSTM is a variation of RNN, which uses gate and memory cells for sequence prediction. It estimates dependencies of different time scales and is suitable in the COVID-19 forecasting task. Fundamentally, an LSTM handles the sequence by having a recurrent hidden state whose activation is dependent on that of the previous state.

The structure of our LSTM model is illustrated in **Figure.1**. It is a custom deep neural network with one large LSTM layer and one small LSTM layer, which is then followed by two fully connected layers. Each dense layer has a dropout layer after them to prevent gradient vanishing. Lastly, one output neuron to predict values (total cases, deaths) of given date. Rectified linear unit (ReLU) function is used as activation function in all layers to avoid gradient explosion and vanishing.

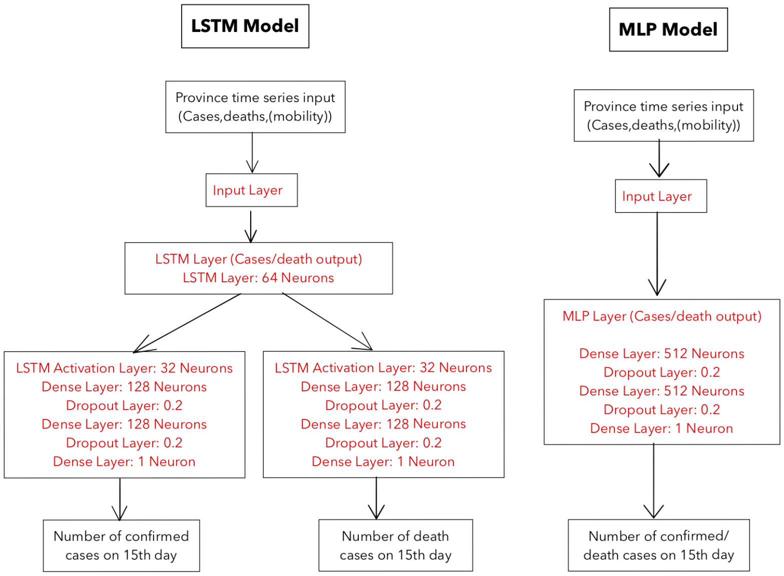

Fig. 2. The structure of LSTM and MLP model

3.3 MLP

The graphical representation of our custom MLP is shown in **Figure.1**. MLP is a class of feedforward artificial neural network (ANN) composed of multiple layers of full connected perceptrons [**?**]. MLP can be used in regression and give polynomial fit of the input graph. It updates the weights through back propagation during training. It can also have a regularization term added to the loss function that shrinks model parameters to prevent overfitting.

We observe that the existing trend of COVID-19 can be summarized with one curve and no sudden change exists in the data. A regressor is capable of performing certain degrees of prediction given that no impactful events happen in future. Here we use a MLP model with two dense layers with 512 neurons, each followed by a dropout layer. Finally, one output neuron to predict values (total cases, deaths) of given date. Similar to the case of LSTM, ReLU activation functions and dropout layers are applied to prevent gradient issues in back propagation.

4 Model Implementation

4.1 Dataset

The COVID-19 Canada provincinal dataset from website of Government of Canada contains data from the beginning of March to late July. The COVID-19 dataset contains several time series including confirmed cases and death cases [10].

The daily mobility data measures relatively how often people stays in certain public areas like parks, working places, restaurants during the day. The data is accessible from Goggle Community Mobility Reports [11].

4.2 Data Preprocessing

For this study, we only focus on the four provinces with the highest population (Ontario, BC, Alberta, and Quebec). First we generate a summary of the whole dataset to check the total confirmed cases and deaths cases around Canada. The result is shown as Table 1.

Data cleaning The redundant entries are dropped them and the date column is parsed as date-time datatype to make preparation for time series forecasting. For the mobility dataset, we fill all the empty values with 0 and similarly, parse the date column as date-time data type.

Table 1. Summary of Canada COVID-19

province	confirmed	deaths	active	recover
British Columbia	2916.0	174.0	2590.0	152.0
Alberta	8108.0	154.0	7407.0	547.0
Ontario	35068.0	2672.0	30344.0	2052.0
Quebec	55458.0	5503.0	24798.0	25157.0
Canada	118546.0	8966.0	67594.0	28008.0

Normalization For cases data, we applied separate standard scalers on the cases data of different province and store the scalers in a list for re-transformation in result analysis part.

For mobility data, we first use a min-max scaler to normalize all the mobility columns together and take the average to generate a mobility_cof columns, which is a coefficient in [0,1] representing the influence of mobility data.

Input Generation We train all three models with provincial data from March to June 2020. For LSTM model, we create a trend dataframe containing time series columns like "confirmed trend", "death trend" "mobility trend" and two more label columns "expected cases" and "expected deaths". The length of a time series was set as 14 days since it is the average incubation time. The confirmed and death data of the 15th day will perform as the label in training part. For MLP and SEIR model, the original time-series data is used directly.

We generated two separate input sets - one with mobility data and one without. And we will do training, validation and prediction on both sets to analysis the influence of mobility data about the COVID-19 trend.

4.3 Model Construction and Validation

SEIR model is built from scratch. The ODEs are solved iteratively using *scipy* package *solve_ivp* function. There are 3 modes to varying the reproduction number - by constant reproduction number, by a hill decayed reproduction number, and by coefficient obtained by regression on mobility data. Then to fit the model, the last 7 days of the whole training set are chosen as the validation set. and we fit the model with real data by minimizing the MSLE. Then the best reproduction decay function was chosen automatically by comparing the loss function in the generated solution in three modes. Finally, the future trend is predicted using the best-fit parameters in the chosen SEIR model and MSLE is calculated.

LSTM and MLP models are built with TensorFlow keras package. The loss functions for both models are both Mean Square Error and the optimizer is always Adam. The entire training set is shuffled and split into two parts - training set (80%) and validation set (20%). Then the model is implemented on the validation to see if it is performing well. If not, the parameters are modified to get the best performance. The RMSE is calculated to evaluate and compare performances among models. Future predictions are performed on July 2020 and is compared with actual number of infected and deaths.

5 Result and Analysis

5.1 Baseline SEIR Model Result Analysis

SEIR model is sensitive to variance in reproduction number, thus often requiring careful choices of reproduction functions. **Figure.3** and **Figure.4** are generated using the best reproduction function among the provided three functions with minimized loss. It can be observed from **Figure.3** that, although the model tries its best to fit the training data, there are still big gaps between the fit and actual line. It is because SEIR model is a ideal model by nature and is destined to differ from real data.

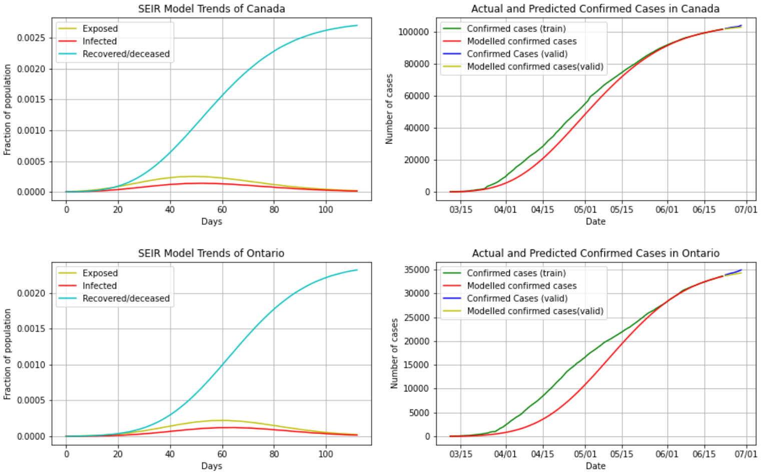

Fig. 3. The best SEIR model (left) and its corresponding fitting line (right) on data of Canada (up) and Ontario (down).

On the other hand, the COVID-19 is a global epidemic which causes great intervention from the whole society. A high level of intervention leads to strong perturbation to the ideal SEIR model, making its results unreliable. The predictions is shown in **Figure.4** and the root mean square error (RMSE) values are listed in **Table.2**, which is obviously far from being satisfactory.

5.2 Training Performance of LSTM and MLP

It is noted that SEIR is the slowest in training, LSTM is the second slowest, while MLP is the fastest. MLP model has the simplest design and least parameters

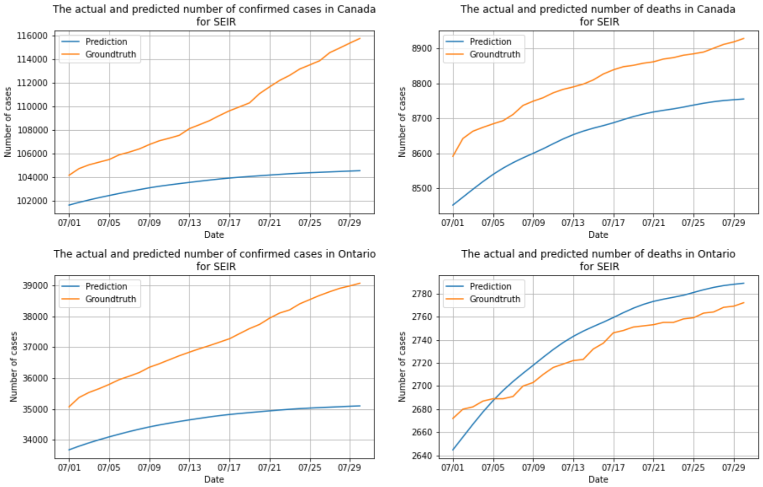

Fig. 4. SEIR prediction on data of Canada (up) and Ontario(down).

Table 2. RMSE of Predictions on Provinces using SEIR

Province	**RMSE**	
	Confirmed	**Deaths**
British Columbia	331	8
Alberta	1193	8
Ontario	2672	18
Quebec	2150	198
Canada	6541	149

so it is fast. The loss evaluation of SEIR is inefficient, it updates 4 parameters mostly by trail-and-error, which makes it the slowest.

For LSTM and MLP on all provinces, the training loss steadily goes downwards and eventually plateaus after around 5 epochs. The **Figure.5** and **Figure.6** are two representative figures of losses of LSTM and MLP respectively. The presence of an extra column of mobility data does not slow down the convergence. It is also noted that the loss of provinces with large population tend to reach plateau faster than that of provinces with small population. It indicates that large population leads to robustness of the model, having high tolerance to deviant points.

5.3 LSTM Result Analysis

Table.3 shows the RMSE values of several representative provinces using Mobility data or not. Generally, the LSTM model outperform the basic SEIR model a lot. From the table, it is noticeable that for most regions, the model with mobility data performs better than the without mobility on both confirmed cases and

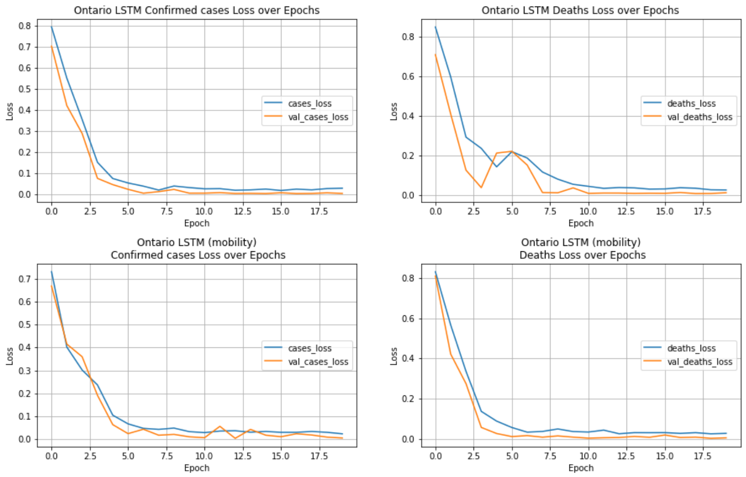

Fig. 5. LSTM performance with and without mobility data on Ontario

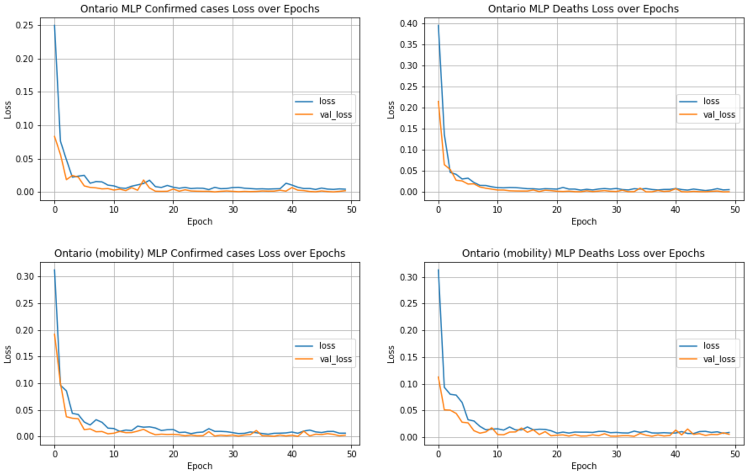

Fig. 6. MLP performance with and without mobility data for Ontario.

deaths, indicating that mobility has impact on the spread of the virus. Since the mobility of a given area is directly impacted by Canada's lock-down regulation from April to June, the slowdown of the increase in new cases is expected.

Table 3. RMSE of Provinces using LSTM

Province	Without Mobility		With Mobility	
	Confirmed	**Deaths**	**Confirmed**	**Deaths**
British Columbia	112	4	105	4
Alberta	471	7	390	6
Ontario	556	20	550	16
Quebec	542	27	542	52
Canada	1688	57	1689	72

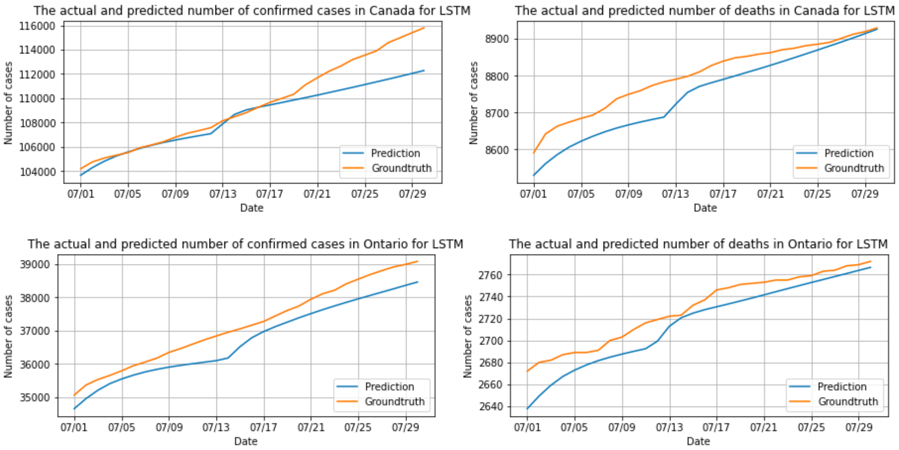

Fig. 7. LSTM prediction on number of confirmed cases (left) and deaths (right) without mobility data for Canada and Ontario.

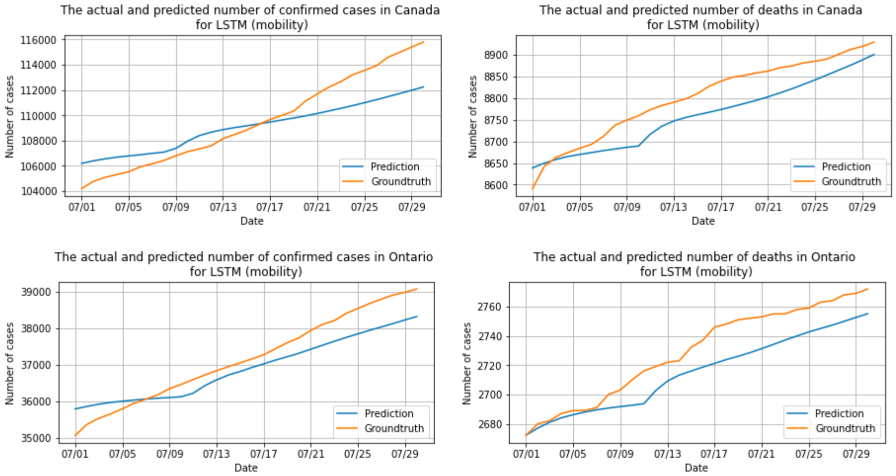

Fig. 8. LSTM prediction on number of confirmed cases (left) and deaths (right) with mobility data for Canada and Ontario.

For Quebec and entire Canada, the deaths RMSE is lower without mobility, which is against the pattern we just concluded. It could be that the spread of virus does not immediately results in casualty. Therefore, the relation between mobility and death cases are weak and could be removed for better accuracy. Two representative prediction results on entire Canada and Ontario can be seen if **Figure.7** and **Figure.8**.

5.4 MLP Result Analysis

From the **Table.4**, it is worth mentioning that the MLP model performs much better than LSTM, for every representative province and without mobility data. But if mobility data is added, the performance is similar to or even worse the previous model with mobility. Since it is a simple function regression, introducing extra dimension (mobility) could cause difficult convergence. When we remove the impact of mobility data, the RMSE shows a sharp decrease for every province and entire Canada.

Table 4. RMSE of Provinces using MLP

Province	Without Mobility		With Mobility	
	Confirmed	**Deaths**	**Confirmed**	**Deaths**
British Columbia	58.95943	2.4728203	150.7925	4.389088
Alberta	230	4	464	7
Ontario	282	7	831	28
Quebec	2001	15	764	46
Canada	791	16	2288	80

Similar trends are demonstrated in **Figure.9** and **Figure.10**. The MLP curve fits perfectly almost every date in July without mobility. But although the trend is similar to the true label, the curve with mobility seems to have a bias which leads to a larger value of RMSE in validation set. It is expected since MLP regression assumes smooth no unexpected change in the regressed function, while in real-life, many parameters vary with time and cause deviation. In this case, the development is steady enough that the MLP works well. Overall, the MLP model without consideration of mobility is the fastest and gives the most accurate prediction.

It can be learnt from the above table and plots that although LSTM is more acknowledged for time series predictions, it is worth trying some simple models like regular MLP networks sometime.

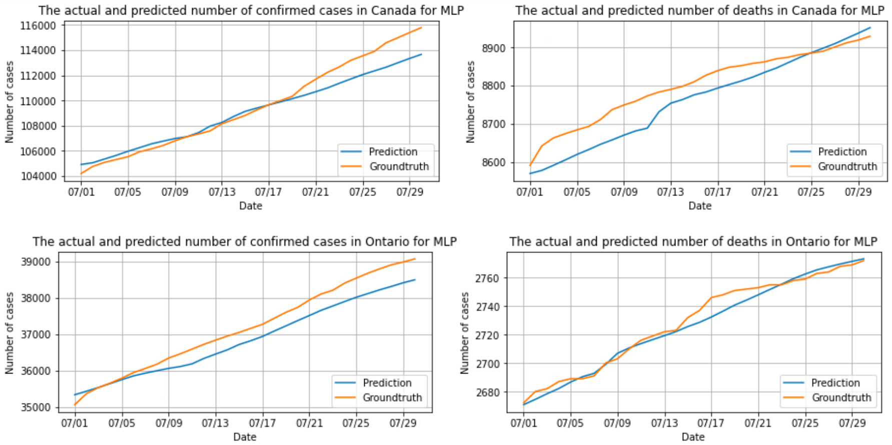

Fig. 9. MLP prediction on number of confirmed cases (left) and deaths (right) without mobility data for Canada and Ontario.

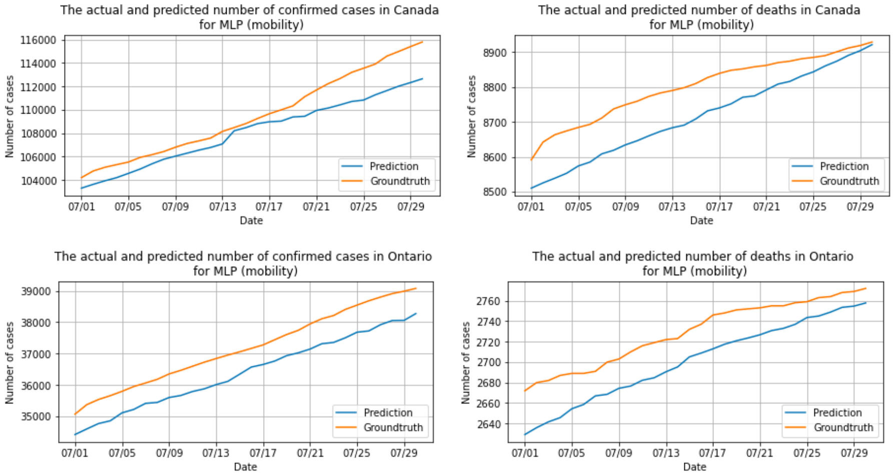

Fig. 10. MLP prediction on number of confirmed cases (left) and deaths (right) with mobility data for Canada and Ontario.

6 Conclusion and Future Work

For domestic COVID-19 prediction (Canada), both custom LSTM and MLP models outperform the SEIR mathematical model. The custom MLP model without mobility impact performs best among all LSTM, MLP and SEIR model, which can help the government to overcome the COVID-19 crisis. It is also observed that mobility data, or measurements of public gathering, would increase the accuracy of LSTM-based models.

Although we have developed a prediction model with rather high performance, there is still a long way to go if we plan to analyze the factors impacting the spread of COVID-19 in Canada better. A possibly extension is GAN(Generative adversarial networks) and ARIMA model that are generally used to do regression and prediction. Therefore, our future work mainly focuses on more advanced time series forecasting methods.

References

1. Ardabili, S.F. et al.: COVID-19 Outbreak Prediction with Machine Learning. (2020).
2. Chinazzi, M. et al.: The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. (2020).
3. Dandekar, R. et al.: A machine learning aided global diagnostic and comparative tool to assess effect of quarantine control in Covid-19 spread. (2020).
4. Dukic, V. et al.: Tracking Epidemics With Google Flu Trends Data and a State-Space SEIR Model. Journal of the American Statistical Association. 107, 500, 1410–1426 (2012).
5. Huang, C.-J. et al.: Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China. (2020).
6. Kraemer, M.U. et al.: The effect of human mobility and control measures on the COVID-19 epidemic in China. (2020).
7. Li, G. et al.: Global stability of an SEIR epidemic model with constant immigration. Chaos, Solitons & Fractals. 30, 4, 1012–1019 (2006).
8. Liu, L. et al.: Global stability of an SEIR epidemic model with age-dependent latency and relapse. Nonlinear Analysis: Real World Applications. 24, 18–35 (2015).
9. Rizk M. R. et al.: COVID-19 forecasting based on an improved interior search algorithm and multi-layer feed forward neural network. (2020).
10. Canada official statistics on COVID-19, https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html. Last accessed 8 Aug 2020
11. Google Community Mobility Reports, https://www.google.com/covid19/mobility/. Last accessed 14 Aug 2020