Forecasting Ontario’s Daily Hospitalization Due to COVID-19

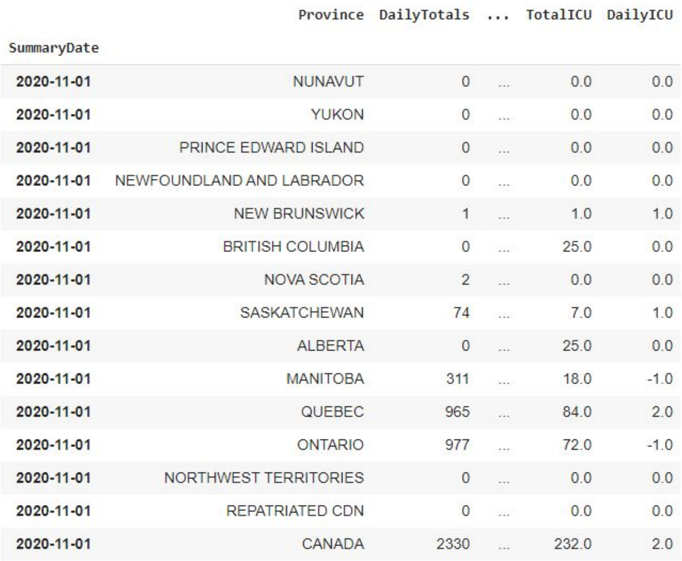
Introduction

The data mining task attempted was to forecast the daily hospitalization count for the province of Ontario, Canada. This has significant implications in allowing hospital management to foresee the influx of patients in a certain area. If data was available on a more granular level, it would be even more beneficial as Ontario is a large province with numerous cities and hospitals. The dataset used was taken from the government of Canada’s published resources pertaining to COVID-19 data per province and Canada Wide [1]. The data extracted included the date range from 2020-01-25 to 2020-12-02.

A multivariate, multi-step Long-Short-Term-Memory (LSTM) model was chosen to model the data. This type of LSTM model allows for multi-step prediction which is useful for predicting multiple days in advance at a time. The mean absolute percentage error (MAPE) is regarded as one of the most widely used metrics for forecast accuracy and is used to evaluate the proposed model as it is both scale-independent and is easy to interpret. Existing COVID-19 prediction algorithms that were treated as benchmarks also used the same metric.

Preprocessing Steps

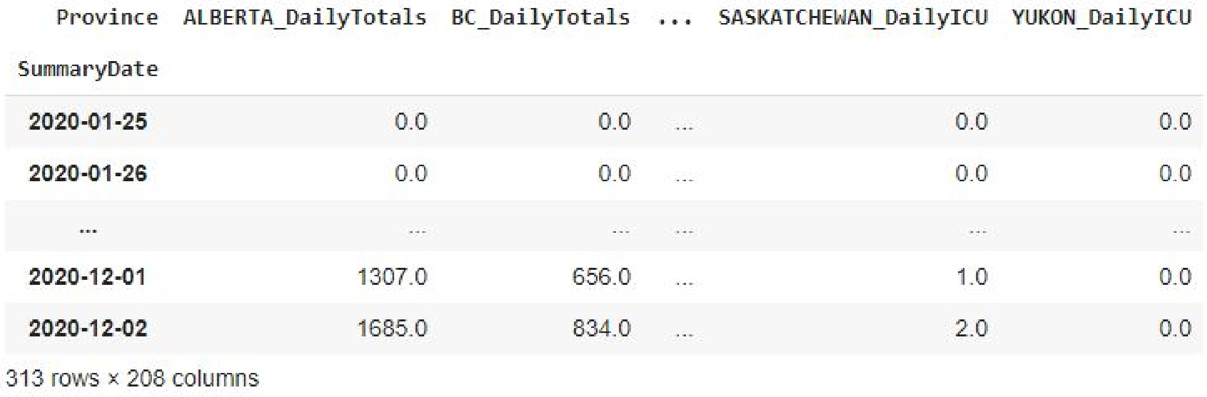
The data includes rows of daily information for the ten provinces and three territories of Canada. Additionally, the summation of all these values per day is an added row. A row called ‘Repariated CDN’ is also present to account for any discrepancies in the data that was later corrected. An example of how the data looks like on one day is shown in Figure 1. below, however not all columns are shown for simplicity.



	Province	DailyTotals	...	TotalICU	DailyICU
SummaryDate					
2020-11-01	NUNAVUT	0	...	0.0	0.0
2020-11-01	YUKON	0	...	0.0	0.0
2020-11-01	PRINCE EDWARD ISLAND	0	...	0.0	0.0
2020-11-01	NEWFOUNDLAND AND LABRADOR	0	...	0.0	0.0
2020-11-01	NEW BRUNSWICK	1	...	1.0	1.0
2020-11-01	BRITISH COLUMBIA	0	...	25.0	0.0
2020-11-01	NOVA SCOTIA	2	...	0.0	0.0
2020-11-01	SASKATCHEWAN	74	...	7.0	1.0
2020-11-01	ALBERTA	0	...	25.0	0.0
2020-11-01	MANITOBA	311	...	18.0	-1.0
2020-11-01	QUEBEC	965	...	84.0	2.0
2020-11-01	ONTARIO	977	...	72.0	-1.0
2020-11-01	NORTHWEST TERRITORIES	0	...	0.0	0.0
2020-11-01	REPATRIATED CDN	0	...	0.0	0.0
2020-11-01	CANADA	2330	...	232.0	2.0

Figure 1. Partial data for 2020-11-01

As this task was treated as a time series prediction, the dataset had to be transposed so that each row corresponded to the data on one day. Figure 2. Depicts this transposed dataset, that resulted in a total of 313 days of data.



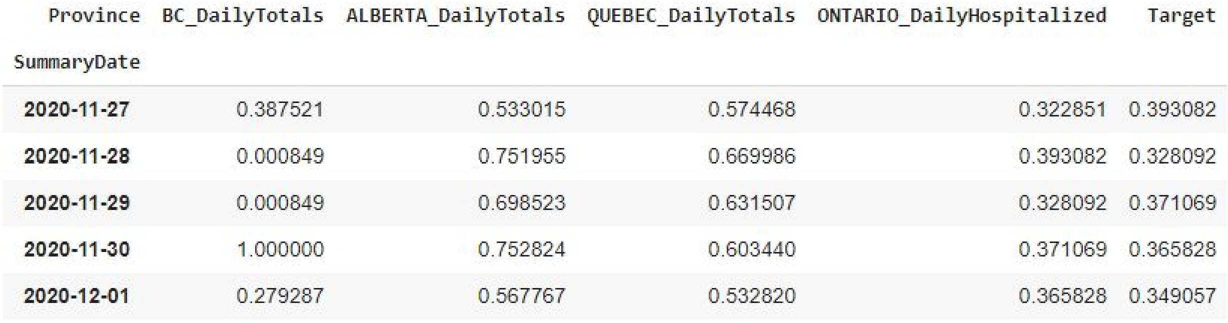
Province	ALBERTA_DailyTotals	BC_DailyTotals	...	SASKATCHEWAN_DailyICU	YUKON_DailyICU
SummaryDate					
2020-01-25	0.0	0.0	...	0.0	0.0
2020-01-26	0.0	0.0	...	0.0	0.0
...	...	...	...	...	...
2020-12-01	1307.0	656.0	...	1.0	0.0
2020-12-02	1685.0	834.0	...	2.0	0.0

313 rows × 208 columns

Figure 2. Transposed dataset for time series

Columns with zero variance were removed from the newly transposed dataset as they did not add any value to the prediction. The data was also normalized between 0 and 1 so that all features were placed on a common scale. Different types of feature selection methods were tested, and feature selection by way of correlation analysis was found to have the best overall result in terms of the prediction error of the model. For all feature sets, the target variable is the one day ahead value of ‘Ontario_DailyHospitalized.’

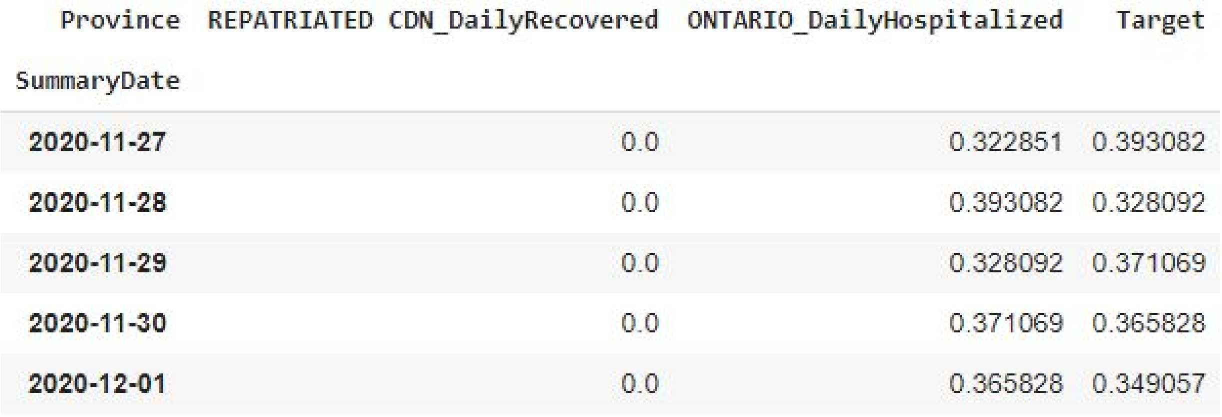
In Linear Regression Coefficient Ranking, all features were used to predict the target variable (Ontario Daily Hospitalized) using a linear regression algorithm. The resulting coefficients were then ordered, with the larger coefficients indicating a higher impact on the target variable. Only features with positive coefficients were kept, and the top four features as shown in Figure 3., were used as input features to the LSTM.



Province	BC_DailyTotals	ALBERTA_DailyTotals	QUEBEC_DailyTotals	ONTARIO_DailyHospitalized	Target
SummaryDate					
2020-11-27	0.387521	0.533015	0.574468	0.322851	0.393082
2020-11-28	0.000849	0.751955	0.669986	0.393082	0.328092
2020-11-29	0.000849	0.698523	0.631507	0.328092	0.371069
2020-11-30	1.000000	0.752824	0.603440	0.371069	0.365828
2020-12-01	0.279287	0.567767	0.532820	0.365828	0.349057

Figure 3. Dataset from Linear Regression Coefficient Ranking

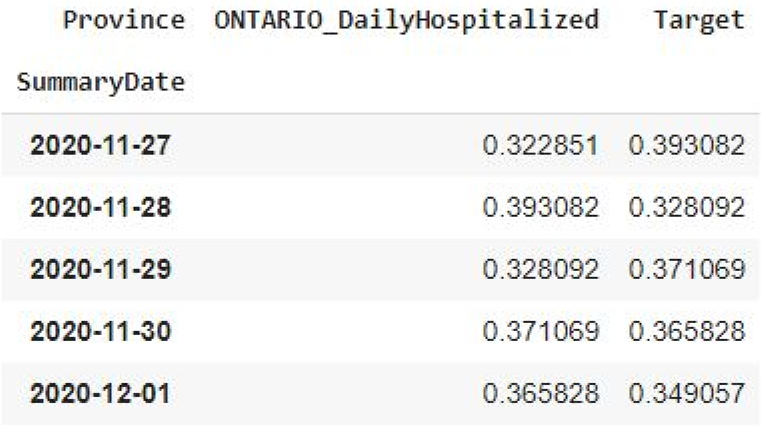
Feature selection via correlation analysis, and the resulting model prediction error was also obtained. Features that had a correlation coefficient lower than 0.7 to other features were kept in the dataset, so that highly correlated features, resulting in redundant features, would be removed. Additionally, only features that had a correlation coefficient greater than 0.7 with the target variable were kept in the dataset. Two features remained in the dataset after using this algorithm as shown in Figure 4.



Province	REPATRIATED CDN_DailyRecovered	ONTARIO_DailyHospitalized	Target
SummaryDate			
2020-11-27	0.0	0.322851	0.393082
2020-11-28	0.0	0.393082	0.328092
2020-11-29	0.0	0.328092	0.371069
2020-11-30	0.0	0.371069	0.365828
2020-12-01	0.0	0.365828	0.349057

Figure 4. Dataset from Correlation Analysis

Finally, a model that only used the shifted version of the target variable as a feature was also tested. A dataset excerpt corresponding to this model can be seen in Figure 5. This model had the lowest computational power usage in comparison to the other feature sets, however the tradeoff between computational power and accuracy was not a big concern for this particular project as training each epoch of the model took on average less than 50 ms.



Province	ONTARIO_DailyHospitalized	Target
SummaryDate		
2020-11-27	0.322851	0.393082
2020-11-28	0.393082	0.328092
2020-11-29	0.328092	0.371069
2020-11-30	0.371069	0.365828
2020-12-01	0.365828	0.349057

Figure 5. Univariate Dataset

Model Description and Data Preparation

A multivariate multi-step LSTM (MMLSTM) model was chosen for this analysis. Other RNNs are accustomed to the vanishing gradient problem and the exploding gradient problem due to back propagation that is conducted over numerous layers. An LSTM however, uses its different gates to have selective memory and therefore avoids both these problems. The MMLSTM proposed in particular uses a windowing approach to predict future values. Data from time t-29 to time t is used to predict the target variable from time t to time t+6. The model consists of one LSTM layer with 64 units, followed by a dropout layer and a dense layer. The dropout layer was included to introduce randomization and increase the generalizability of the model. The dropout factor was kept at 0.4, meaning that 40% of the nodes are randomly dropped during each epoch.

The first 60% of the data was used for training, the next 20% was used for validation, and the last 20% were reserved for testing. As it is a time series problem the data was not shuffled. The same train-validation-test splits were used for all the tested models.

Sensitivity and Comparative Analysis

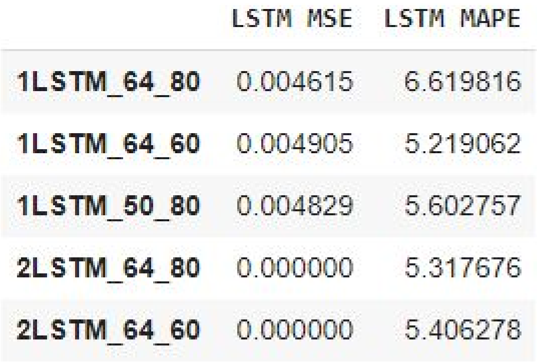
Sensitivity testing was performed by varying the number of layers and units in the model, as well as the epochs. The results corresponding to this analysis are discussed in the Results section.

Comparative analysis was performed by building a multivariate multi-step CNN. This CNN took in the same features as the LSTM. The CNN was built with one one-dimensional convolutional layer, followed by a max pooling layer, a flattening layer and then two dense layers. An autoregressive integrated moving average (ARIMA) statistical model was also tested during the comparative analysis phase. Unlike both the LSTM and the CNN, this model only predicted data for the next day, and so was expected to have a lower overall prediction error. The specifications of the ARIMA model used was a lag order of 4, a degree of differencing of 1, and a moving average order of 0. The two deep learning models were run five times and the median MAPE, and MSE values were stored and averaged. The ARIMA model was only run once as its results are consistent.

Results

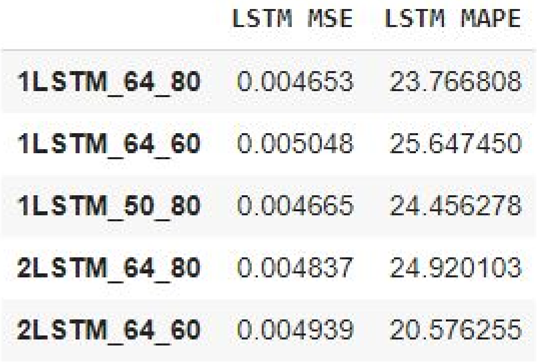
The results corresponding to the sensitivity analysis and comparative analysis are included in Tables 1 to 4 below. It was found that using the feature set created after correlation analysis resulted in the lowest errors among all the different LSTM models tested as indicated in Table 1. These results are, however, only marginally better than the univariate model errors found in Table 3. In Table 1, the LSTM model with the lowest MAPE is the proposed model with one layer, 64 units and 60 epochs, and so this model was selected as the final model. When this model was compared with the CNN model, the performance was marginally better. The average MAPE over five runs was 5.68 for the LSTM versus 5.78 for the CNN. It should be noted that there were instances where the CNN had a lower MAPE, for example in the last row of Table 4. Therefore, with further testing, there may be instances where the CNN outperforms the LSTM. The ARIMA model performed better than both the LSTM and the CNN as expected, due to the fact that the ARIMA model only predicts one day in advance, whereas both the deep learning models tested predicted a week in advance.

Table 1. Sensitivity Analysis - Correlation Features



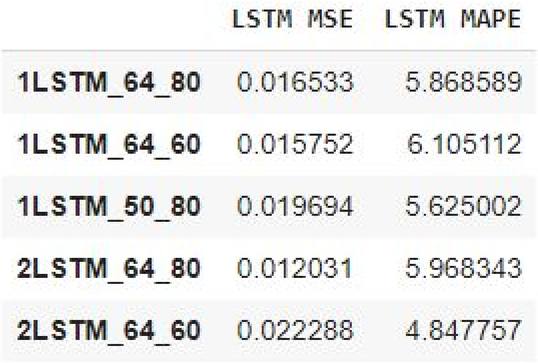
	LSTM MSE	LSTM MAPE
1LSTM_64_80	0.004615	6.619816
1LSTM_64_60	0.004905	5.219062
1LSTM_50_80	0.004829	5.602757
2LSTM_64_80	0.000000	5.317676
2LSTM_64_60	0.000000	5.406278

Table 2. Sensitivity Analysis - LR Coefficient Ranking Features



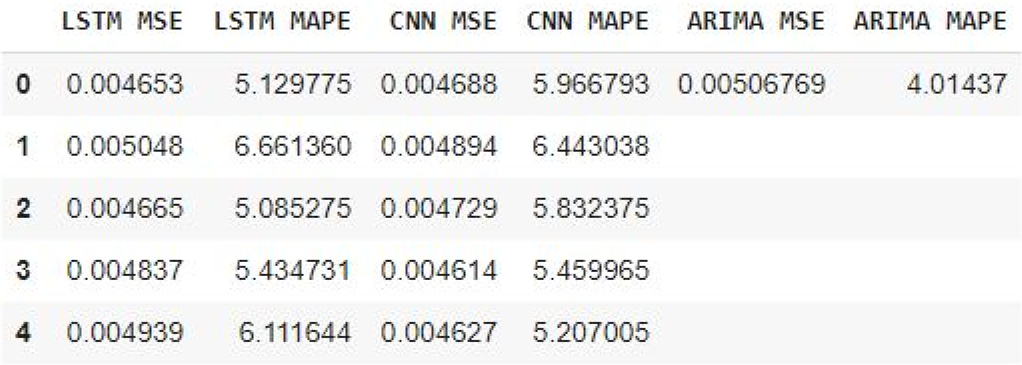
	LSTM MSE	LSTM MAPE
1LSTM_64_80	0.004653	23.766808
1LSTM_64_60	0.005048	25.647450
1LSTM_50_80	0.004665	24.456278
2LSTM_64_80	0.004837	24.920103
2LSTM_64_60	0.004939	20.576255

Table 3. Sensitivity Analysis - Univariate



	LSTM MSE	LSTM MAPE
1LSTM_64_80	0.016533	5.868589
1LSTM_64_60	0.015752	6.105112
1LSTM_50_80	0.019694	5.625002
2LSTM_64_80	0.012031	5.968343
2LSTM_64_60	0.022288	4.847757

Table 4. Comparative Analysis



	LSTM MSE	LSTM MAPE	CNN MSE	CNN MAPE	ARIMA MSE	ARIMA MAPE
0	0.004653	5.129775	0.004688	5.966793	0.00506769	4.01437
1	0.005048	6.661360	0.004894	6.443038		
2	0.004665	5.085275	0.004729	5.832375		
3	0.004837	5.434731	0.004614	5.459965		
4	0.004939	6.111644	0.004627	5.207005		

The validation error and training error curves are shown in Figure 6. As both curves converge, it can be stated that this model is not severely overfit or underfit. Underfitting is described as when a model does not comprehensively learn patterns from the training data. Overfitting occurs when a network is overtrained which can be identified as the curve behaviour past the convergence point, where the train and validation error move in opposite directions, with the validation error continually increasing. [2]

The model is slightly underfit due to the low amount of training data. 80% of the 313 day dataset is used for training and validation in this case, amounting to about 250 days. When the configurations are changed to have the model train and validate on 90% of the data, the model is evidently less underfit, as can be seen in Figure 7. It can therefore be assumed that using this model when more data is available will produce more robust results.

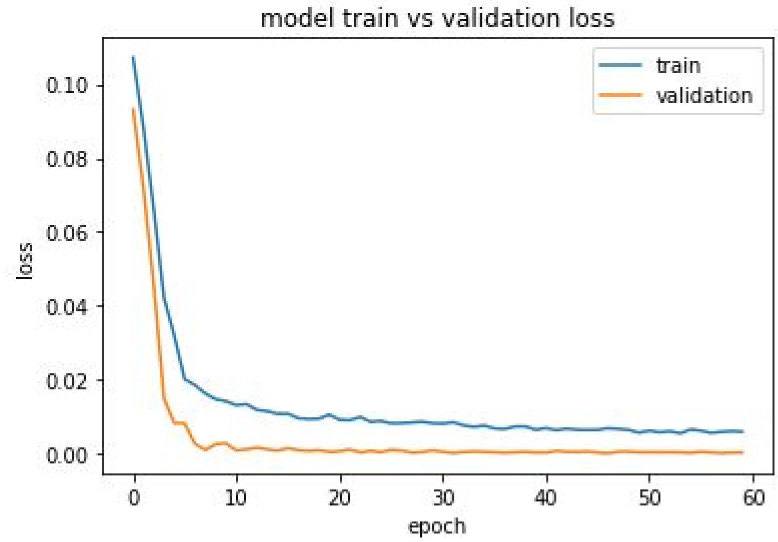


Figure 6. Model History: LSTM, 1 layer, 64 units, 60 epochs

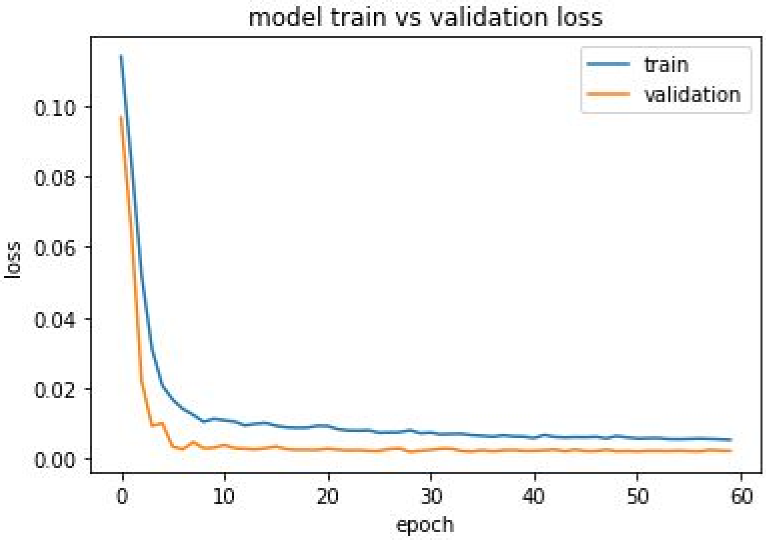


Figure 7. Proposed LSTM trained and validated with 90% of dataset

The results of this project were compared that of researchers in [3] who forecasted the total number of confirmed COVID-19 cases for a two week period in many countries with both an LSTM and an ARIMA model. They found that although the LSTM was a consistently good predictor, there were countries that were better predicted using the ARIMA model. The MAPE values obtained by the LSTM ranged from 0.1640 to 2.5025. Comparatively, the results of the proposed MMLSTM discussed in this report had a higher MAPE of 5.68 on average. The dataset used by the authors of [3] differed from what was used in this project, and the feature predicted was different as well. Additionally, the hyperparameters for the model were not mentioned by the authors and may account for discrepancies in MAPE.

Conclusion and Future Work

Resource allocation during the time of a global pandemic is something of utmost importance. Projects such as this MMLSTM could also aid decision makers in advocating for or against extended lockdown measures in a data driven manner. Extending the prediction period to more than one week as well as testing the model on other provinces and cities would be the next steps for this project. It would serve as a good indicator of the generalizability of the model. The ARIMA model tested in this experiment could be modified to be a multi step prediction algorithm as well, so that it is more comparable to the other two models discussed. Due to the nature of the small dataset that was used, simpler statistical models like the ARIMA model may be a better alternative to machine learning models that work well with big data. Sensitivity analysis could also be conducted on the CNN model that was tested, as there were some instances where it performed slightly better than the MMLSTM. The mean absolute percent error was used to compare the models, however, there are some instances where the MAPE trended to infinity due to division by a zero in the true value of the test set. Due to this, the mean MAPE could not be used but instead, the median MAPE was used. Alternatively, the Mean Arctangent Absolute Percent Error (MAAPE) that uses bounded influences for outliers by considering the ratios as an angle rather than a slope could be used. This preserves the properties of the MAPE, while being able to overcome the problem with MAPE of the division by zero [4].

References

[1] https://resources-covid19canada.hub.arcgis.com/datasets/provincial-daily-totals/data?page=469

[2] W. Khan, S. Chung, M. Awan and X. Wen, "Machine learning facilitated business intelligence (Part II)", Industrial Management & Data Systems, vol. 120, no. 1, pp. 128-163, 2019.

[3] İ. Kırbaş et al, "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches," Chaos, Solitons and Fractals, vol. 138, pp. 110015-110015, 2020.

[4] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts", International Journal of Forecasting, vol. 32, no. 3, pp. 669-679, 2016.