A Deep Recurrent Neural Network to Support Guidelines and Decision Making of Social Distancing

Mohammed Aledhari*, Rehma Razzak*, Reza M. Parizi*, Ali Dehghantanha†,
*College of Computing and Software Engineering, Kennesaw State University, GA, USA
Emails: maledhar@kennesaw.edu, rrazzak@students.kennesaw.edu, rparizi1@kennesaw.edu
†Cyber Science Lab, School of Computer Science, University of Guelph, Ontario, Canada
Email: adehghan@uoguelph.ca

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377800

***Abstract*—The recent Covid-19 pandemic instigated many changes in our way of life within the United States, and slowly but surely we are working towards mitigating the virus. Due to Covid-19, there are higher demands for models to accurately forecast the number of Covid-19 cases that factor mandated guidelines such as social-distancing. Many scholarly and corporate research entities are investigating ways to achieve this goal preemptively; Unfortunately, current models are not yet able to accurately model future Covid-19 cases that factor in various guidelines; What is lacking with these models is an understanding of crucial factors affecting spread, accuracy, availability of reported cases on a small scale, and quantifiable metrics for how social distancing and quarantine efforts mitigate the spread. Therefore, the goal of this study is to produce a mathematical model to directly aid policy decisions by comparing predicted models of various decisions and social distancing protocols. This model can be applied on top of existing models to factor in more imminent data and produce predictive curves, indicating troughs and peaks of new daily Covid-19 cases with comparatively high accuracy, which can aid in analysis. These predictive curves can, therefore, be generated using data corresponding to projected responses to proposed guidelines and compared to each other to choose the optimal solution for "flattening the curve" of the Covid19 infection rate. We use an LSTM-RNN model with ANN Regression in an attempt to predict future Covid-19 cases. Our model achieved comparable results, but further improvements could be implemented for more optimal results.**

***Index Terms*—Machine Learning, Epidemiology, Medical Field, Covid-19, Healthcare, Policy Making, Pandemic Response**

I. Introduction

As COVID-19, the disease caused by the SARS-CoV-2 virus, changes the landscape of business, education, socialization, government policy, and life in general [1], it is important to keep track of daily figures regarding new cases as they appear. For purposes of planning, many institutions have also implemented mathematical models to attempt to predict these figures [2]. Prior studies, such as one published in Chaos, Solitons, and Fractals, have attempted to do the same within the context of academia. These models are often implemented via machine learning, with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) being particularly important; The reason for this is because both RNNs and CNNs enable us to do time-series analysis. [3]. Other prior studies, such as the aforementioned attempt by Chimmula et al [3], have demonstrated success in utilizing RNNs alone or in combination with Long-Short Term Memory (LSTM) Layers [4] to predict the trend in the number of new COVID-19 cases.

Currently, most of the models in circulation are not complex enough to factor in plans for social distancing and other logistics. This is because the current models rely purely on past COVID-19 case data to predict future COVID-19 case data. Additionally, the predictions obtained from the models and their credibility are only as good as the model that is used, the extent and accuracy of the assumptions made, and the quality of the data that gets used for the models. To enable policymakers to utilize mathematical models more effectively in their decision making process, these models must be expanded to include external factors such as the day of the week and social distancing indices in their calculation process. A multi-network approach is the appropriate course of action, as they have been proven by multiple studies, including one written by N. Ueda and R. Nakano published in 1996, the early days of machine learning technology [5].

In regards to data, the amount of data on COVID-19 infection, deaths, tests plus other factors can also be limited due to undeterdetection or inconsistent detection of cases, poor documentation, which greatly affects the model's output quality. As a result, assumptions have been based on limited data, which in turn can hinder future planning and policymaking efforts for the prevention of Covid-19. Models also need to consider environmental factors as well, such as location. Implementing predictive models for larger and more developed countries such as the U.S. are even more challenging due to heterogeneous characteristics in local areas, individual factors such as age, and population distribution. Therefore, the motivation for our study is to advance public health and policymaking in the context of pandemic response by providing a means to better accurately predict large outbreaks of COVID-19 and other pandemic diseases that may arise in the future, whether in the U.S or other countries. The model present in the study can simulate the effects proposed social distancing measures will have on the spread of COVID-19. This could lead to better evaluation of proposed measures and a decrease in the spread of the COVID-19.

As such, the article offers the following contributions:

- Strong insight on how consideration of environmental factors affects the number of tests performed or effects on the rate of contagion through the use of social distancing measures. In doing so, our study attempts to be the first to do so.
- Qualitative insights on our model's results and future improvements we could implement to it.
- Exploration of the effects of pandemic response poliies on citizens' health, which will greatly benefit policymakers because then they will be able to form stronger response plans to the pandemic.

Our overall goal of this study is to facilitate improvement in policy making by providing an informed model that predicts new and daily COVID-19 cases. This is accomplished via the use of indirectly related data as inputs for the COVID-19 predictive model, and this model will improve upon existing models based on LSTM Time-Series analysis. In doing so, we can document the design and implementation of the model, evaluate its quantitative results, and speculate its possible usage as a policy-making tool. By incorporating the LSTM Time-Series analysis model, we are hopeful that it will inspire more models, whether simple or complicated, to be used for predicting Covid-19 cases. Although ML has long been used for modeling other aspects, its application in modeling outbreaks is still in the early stages, and more sophisticated methods are yet to be explored. As such, our model could provide a decent guideline for more implementing more sophisticated models that allow us to derive more information than previously, and ultimately lead to more informed planning and policymaking.

The model produced in this study could lead to a useful new way to inform policy decisions regarding pandemic and epidemic response. This is important for use with diseases that spread quickly and for which a vaccine is not yet available. If new measures are adopted based on this model, the metaphorical road map to disease dormancy may become much clearer. This would improve public health and likely other aspects of well-being as well. As the world adjusts to what has been deemed the new normal by many, perhaps this model could become the new normal for proactive decision making regarding social distancing and scheduling guidelines during times of viral or bacterial outbreak.

The remainder of this paper is structured as follows. Section II contains literature reviews of works relevant to the contents of this study. Section III describes the techniques and dataset used for this experiment. Section IV describes the method by which the results of data analysis were evaluated. Section V briefly summarizes the results and gives suggestions for further research. Section VI concludes the paper.

II. Related works

Many works have deleved into the impacts of covid-19 and efforts to mitigate the diseases. Works, such as [1] mentions the negative economic and social impacts of Covid-19. The fact that people are falling ill with Covid-19 has caused a decrease in the demand for many special products such as hotel stays, agricultural products, and oil. This has resulted in large-scale economic downturn. The study also mentions that there is a potential for negative impacts in the realm of education both at a lower level and at colleges and universities. One example is that children requiring free or reduced lunches at school are often having to go without, as many schools are unable to provide meals at present. Furthermore, places of higher education that focus on research are being forced to suspend non-covid-related research until the outbreak of the virus is contained or a covid-19 vaccine developed. These factors are relevant to the motivation of this study, as the goal of social distancing measures is to curb the spread of the disease, and this study attempts to improve the ability to plan and implement these measures and show exactly how effective they are.

Another interesting find regarding covid-19 impacts come from [6], where the authors examine how even political beliefs can affect compliance with social-distancing orders. According to the authors, willingness to ignore social distancing guidelines can change depending where the individuals lie on the political spectrum. The article notes that residents in republican counties are less likely to abide by stay-at-home orders while democrats are more likely to switch to remote spending. The authors utilized geolcation data and transaction data to test their findings. While strange, the article does provide insight on how one's political beliefs can affect their compliance with social distancing orders from covid-19. The article could aid policy makers in designing protocols that hopefully will account for one's political beliefs. Additionally, we hope the introduction of more data and differing methodologies in how that data is found will reduce this trend of defiance.

Like the previous work, the article [7] discusses social distancing, and mentions that social distancing has been shown to be effective on other viruses such as influenza. The authors note that social distancing measure in only one area were able to do the equivalent of flattening the curve by atleast 23%. Such results support the notion of social distancing being extremely effective. However, with covid-19, there is a much higher transmission rate than influenza according to the fact sheet published by the US center for disease control and prevention (CDC) titled "Similarities and Differences between Flu and COVID-19" [8]. Furthermore, the article notes that more research is needed to assess the effectiveness of social distancing measures in a variety of industries and workplace scenarios.

Additionally, the work by [9] expressed that enforcement of social distancing guidelines could result in delays of infection up to 50 days when applied to all age groups. This means that the current level of infection as of the study had reached only the level it was predicted to reach 50 days prior. The authors utilized a susceptible-exposed-infected-removed (SEIR) model that specified the effects of social distancing on various age categories. SEIR models take into account four statistics regarding a given population or, in this case, age group. Upon implementation and testing, the authors found interventions started earlier in the epidemic actually delay the epidemic curve and interventions started later flatten the curve. The epidemic rebounded when interventions ended. The authors' model suggested that social distancing can provide crucial time to increase healthcare capacity but must occur along with testing and contact tracing of all suspected cases to mitigate virus transmission. This study is relevant because it was able to quantitatively express the impact of social distancing measures using mathematical models common in the realm of epidemiology. Showing what role social distancing measures can have is more crucial than ever, as many states have eased restrictions and the rules that are in place are not being followed as stringently as they were earlier in the pandemic. In fact, the rules are being ignored by almost 40 percent of respondants to a survey performed by Stanford University News in March [4].

While the authors of [9] implemented a SEIR model for predicting impact of social distancing protocols, the authors of [10] perform time-series analysis and use LSTM (Long Short Term Memory) models with RNNs (Recurrent Neural Networks). LSTM layers are composed of LSTM units. According to an article by R. Gall, An LSTM unit typically consists of an input gate, a forget gate, and an output gate. However, this study is less concerned with how they work, as they will be implemented using keras, a high level API for use with the tensorflow or theano libraries as a backend. As for how these layers relate to time series analysis and can help with covid-19, a study published in "Chaos, Solitons, and Fractals" succeeded in predicting covid data using a complex RNN containing LSTMs [3]. The dataset used by the study is a similar one to the "Our world in data" set used by this study. This study does use LSTM layers; however, the way in which they are arranged within the neural network is simpler, as the goal is merely to compare the accuracy of the neural network containing the LSTMs when working alone and when working in concert with a feed forward neural network designed for simple regression.

Another study by [11] also uses time-series analysis with LSTMs, but they modify their approach. Here, they integrate epidemiological concepts, and also use the SEIR model. Their model was able to generate more accurate results. The data produced by this study still lacks the troughs and peaks that are common within the actual covid-19 case data.

In the work of [12], the authors performed a post-implementation evaluation of the lock down procedure mandated by the Italian government. They did this using observed data reported by various Italian provinces as well as mobile phone tracking data. They utilized a time series model mined by tracking mobile phone data via their SIM cards and a time series of data representative of covid-19 cases in each province and analyzed them using statistical analysis. They found that greater compliance with the mobility restrictions was associated with a swifter and more marked decrease in SARS-CoV-2 positive tests. The goal of [12] is similar to the goal in this study in that it is aimed at linking data and policy making, however it is applied post-plan implementation as an evaluative measure as opposed to pre-plan implementation as an evaluative and decisive measure. As the data was already available, the methodology in the study of Italy does not make use of any machine learning technology. It is merely an example of how statistics can be utilized to evaluate and inform decisions on policy regarding covid-19.

The article [13] uses statistical analysis similar to [12] to estimate effect of local, regional and national policies regarding growth and infection rate of covid-19. The authors of [13] use data from China, South Korea, Iran, Italy, France and the United States. This study found large correlations between increased adoption of policy changes and decreased covid-19 infection rates. One example of that was given is that a region in France adopted all french federal guidelines regarding covid-19 and brought its infection rate from 0.33 to 0.16. This shows that policy making plays a large role in slowing the infection rate.

Finally, an article by [14] highlighted an important caveat about policy-making, especially in regards to covid-19. Policy is often being driven by emotive arguments and rash judgements rather than hard facts. The model in this study is a partial solution to this problem, as it will enable policy makers to set aside emotion in favor of emperical data.

III. Techniques and dataset used in the study

In this study, an LSTM time series analysis model was used and trained on new daily COVID-19 case data ranging from January 13, 2020, to April 21, 2020. The model was then used to predict new daily COVID-19 cases ranging from April 22 to June 26 of the same year. This data was then aggregated into a pandas data frame with data representing the day of the week and indices from apple representing social distancing levels of people driving, people making use of public transportation, and people traveling on foot. This data frame was split into training and testing data. The training data consisted of data representing the dates from April 22, 2020, to May 31, 2020, and the testing data consisted of data representing the dates from June 1, 2020, to June 26, 20. June 26 is the end of the dates covered by the dataset, as the earliest dataset was downloaded on June 27, 2020, and could only contain data up to the previous day. A regression model using Keras "dense" layers was trained using the training data in normalized form. Afterward, it was then used to predict the number of new COVID cases on the dates ranging from June 1, 2020, to June 26, 2020, using the testing data.

The data for new daily COVID-19 cases used in this study was collected from a pre-existing CSV file. The data was downloaded from the COVID-19 page on "Our World in Data" on June 27, 2020. We preprocessed it to only include daily new case data from the USA for dates between January

13 2020 and June 26, 2020 [15]. It was also scaled during processing so that the values were all between zero and one, which is commonplace in machine learning models. The data for the day of the week was generated via a utility function that generates a list of a given length that cycles through the numbers zero through six starting from a predetermined number. In the case of the regression training data, this number was three, since April 22, 2020, was a Wednesday. In the case of the regression testing data, it was one, since June 1, 2020, was a Monday. The social distancing data used in this study was also collected from a pre-existing CSV file. The data was mined by apple via location data gathered from consumer devices. The data was downloaded from the Apple Mobility Index web page on June 28, 2020 [15].

The exact construction of the LSTM time series analysis model was inspired by an article by U. Malik [16]. The model is based on the Keras "Sequential" class as it is a recurrent neural network. There ten total layers in the model. A single-unit Keras dense layer is used as the input layer. This is followed by eight hidden layers that alternate between fifty-unit LSTM layers with recursive training and dropout layers with a twenty percent (0.2) chance of dropout. The output layer is another single-unit dense layer. The model is compiled with a mean squared error loss function and the adam optimizer. It is trained for two-hundred epochs with a batch size of four.

The construction of the Regression model contains only Keras dense layers. The total number of layers is five. The input layer is a five-unit dense layer. It is five units because there are five input factors: the predicted value from the LSTM time series model, the day of the week, the apple mobility driving index, the apple mobility transit index, and the apple mobility walking index. The hidden layers are thirty, fifty, and one-hundred units respectively. The activation function used for the first hidden layer is linear, while the activation functions for the second and third layers respectively are both rectified linear unit (Relu) function. The output layer is a single-unit dense layer. The model is compiled with a mean squared error loss function and the adam optimizer. It is trained for fifty epochs with an automatically generated batch size of two.

The reason for utilizing an LSTM time series analysis model first is that the novel approach of this study is meant to add to, but not replace, time series analysis models in order to add realistic peaks and troughs into the predicted curves. An LSTM time series analysis model was a good baseline to use since LSTM models are one of the most frequently used types of time series analysis models. Ten layers were used as that was the format utilized by [16]. The regression model was used simply because regression was the action being performed. The approach to combine the LSTM and regression models was a subject of the contest in the study. An alternative to directly feeding the predictions from the time series analysis model into the regression model was to compute the mean of estimates based on the regression models and LSTM models separately. However, that approach would have required a weighted average, as it is unlikely that the LSTM predictions would contribute to the results at exactly fifty percent. The weight of the LSTM predictions would then have had to be manually computed or put through a third regression function in such a case. This seemed like a waste of computing power compared to simply putting all of the relevant data into a single regression model.

IV. Proposed method

A. Experimental Setup

We developed the model in Python (version 3.8.2) within a Jupyter notebook. Jupyter notebooks are the successor to ipython notebooks and are used to organize source code and display its output in a visually pleasing format. Jupyter notebooks are a standard tool used by academic researchers in the disciplines of machine learning, data science, and artificial intelligence (AI). The python 3.8.2 source code was run inside of a jupyter notebook on a Dell XPS 15 7590 with an intel i9-9980HK with 32 GB of RAM running Ubuntu 20.0.4. Only CPU processing was used; There was no GPU involvement.

B. Architecture

Figure 1 illustrates the architecture of the model, with the addition to traditional models being the regression ANN. The observed data from the first one-hundred days is passed into an LSTM time series analysis model. This means that the training data for the first model consists of 60 arrays, with a size of 100. The data predicted by time series analysis is then passed into the Regression ANN for training and testing, with the test data generating a final output. The training set is forty entries of each of the five parameters. These entries correspond to the dates between April 22, 2020, and May 31, 2020. The testing set is twenty-six entries of the same. The testing set corresponds to the month of June 2020 through the twenty-sixth.

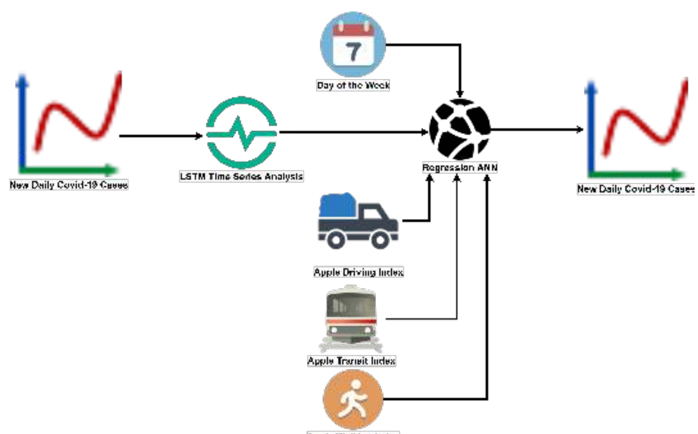

Fig. 1: Proposed Architecture.

The LSTM RNN consists of ten layers total and is trained to compute the number of new daily COVID-19 cases on a given day between April 22, 2020, and June 26, 2020, using the previous one-hundred days of observed data. The layers used include a single-unit dense layer for input, four fifty-unit LSTM layers with recursive training enabled, four dropout layers with a twenty percent (0.2) chance of dropout, and a final one-unit dense layer for output. The LSTM layers use the default activation function of hyperbolic tangent (tanh), the dense layers are set to the default activation linear activation function, and the dropout layers have no activation function. The learning rate for each layer is the Keras default of 0.01 with zero applied momentum. The model, shown in Figure 2, is compiled with the mean squared error loss function and the Adam optimizer. We trained it for two-hundred epochs with a batch size of four.

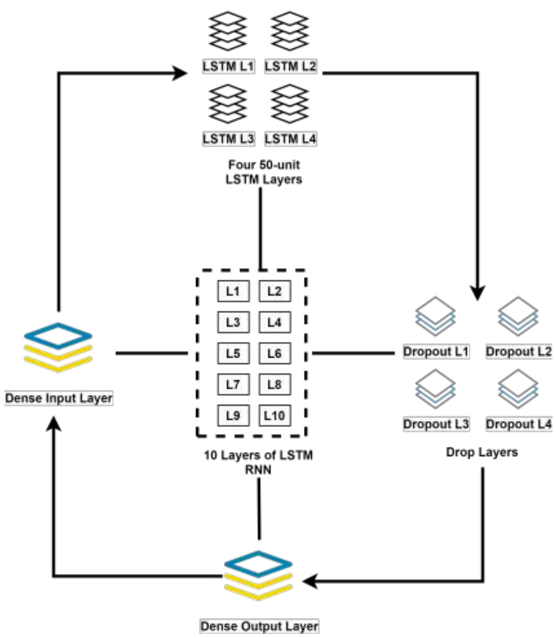

Fig. 2: LSTM Model of the ten layers (denoted as L1, L2, etc.) From left to right, there is the dense input layer, the four 50-unit LSTM layers, then the dropout layers, and finally, the dense output layer.

Meanwhile, the Regression ANN is comprised of five dense layers using the linear activation function. The input layer has five units due to the fact that the data has five parameters. The hidden layers have thirty, fifty, and one-hundred units respectively. The output layer has only one output. The learning rate for each layer is the keras default of 0.01 with zero applied momentum. The model, shown in Figure 3 is compiled using the mean squared error loss function and Adam optimizer. It is trained for fifty epochs with an automatically generated batch size of two. The regression model is shown below:

The goal of the additional ANN is to improve prediction of peaks and troughs while retaining comparative accuracy to the original model. For purposes of this study, the combined model must maintain average accuracy (expressed as a percentage) within 3 percentage points of the original model and must produce troughs and peaks that roughly match the timing of troughs and peaks in the observed data to be deemed a success. The accuracy is computed quantitatively using floating point numbers inside of the jupyter notebook source code. The troughs and peaks are judged by visual comparison to the observed data. The accuracy data is organized in a table with the following categories:

- Value Predicted by The LSTM Model
- Value Predicted by The LSTM-Regression Model
- Value Observed in the Original Data
- Error incurred by the LSTM Model
- Error incurred by the LSTM-Regression Model
- Error incurred by the LSTM Model as a Percentage
- Error incurred by the LSTM-Regression Model as a Percentage

The error data is then averaged and expressed in a table to compare the overall accuracy of the LSTM model alone with that of the LSTM-Regression Model. The table contains columns that express the model used, the average error as absolute distance and the average error as a percentage.

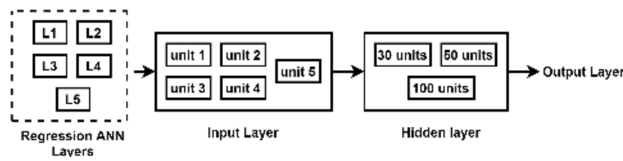

Fig. 3: Regresion ANN Model.

V. Results and Discussions

Before results are discussed, it is best to discuss the data itself, starting with Figure 4. Figure 4 shows data for all one-hundred-sixty-six days.

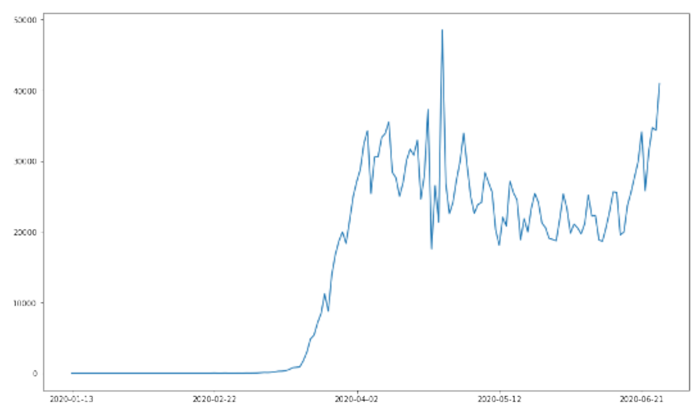

Fig. 4: Representation of the raw data for all one-hundred sixty-six days.

By examining Figure 4 above, it can be seen that up until roughly early march, there is very little covid-19 activity. There is a large peak in early to mid-April that is somewhat of an outlier as well. Some regularity can be seen, so it was deemed appropriate to use days of the week as an input factor for the second part of the model. Meanwhile, Figure 5 presents output from the LSTM time series analyzer, and Figure 6 illustrates a comparison of actual daily new cases detected versus predicted new cases detected.

Judging by Figure 6, the estimated trend line did follow the general path and was quite accurate, but it was free of the

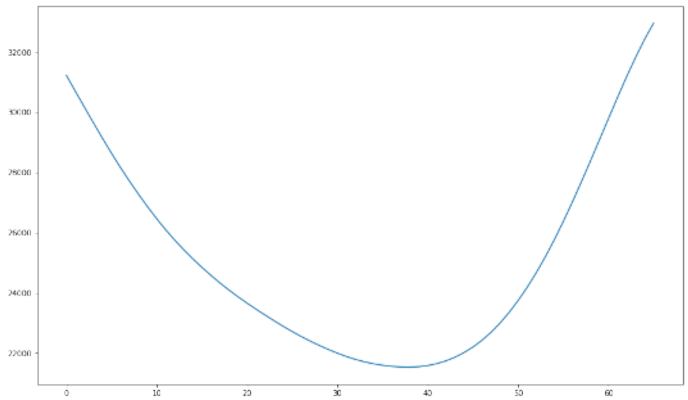

Fig. 5: Raw LSTM prediction output.

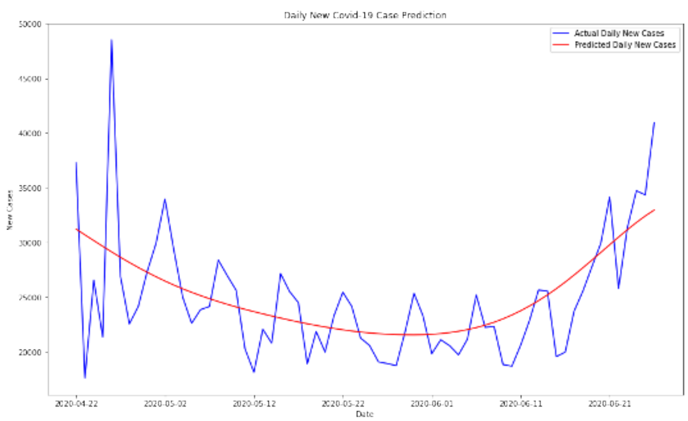

Fig. 6: Actual daily new cases versus predicted daily new cases. The predicted daily new cases are shown in red while the actual daily new cases are shown in blue.

peaks and troughs observed in the actual data on the same days.

To improve upon this data, a set of inputs were put together to incorporate into a regression algorithm. The data came from the dates from April 22, 2020 to May 31, 2020. Table 1(shown below) shows the first ten data points from within the data set for training. The table lists, from left to right, the predicted number of new cases that day as output by the LSTM model, the day of the week, the driving social distancing index, the public transportation social distancing index, and the walking social distancing index. All of these were used to train the regression model. Then a second set (Table 2) was created using the same data for the month of June. The first ten points of data are shown in Table 2. Table 2 was used as test input for the regression model. Please note that due to the hardware used (Intel core i9 with no GPU), the data produced by the test could be slightly different each time the program is run. The output of the test is shown below as a graph in Figure 7.

The data shown in Figure 7 does have some troughs and peaks. To determine whether the data was truly improved however, it must be compared to the actual data. As such,

TABLE I: Predicted number of new cases on specific days (training set).

	Predicted Cases	Day	Driving	Transit	Walking
0	31216.230469	3	94.50	32.65	79.45
1	30685.804688	4	99.48	32.74	84.13
2	30159.291016	5	102.44	33.20	86.94
3	29639.000000	6	107.95	33.67	90.55
4	29126.039062	0	125.32	36.12	104.37
5	28629.283203	1	115.84	33.43	101.57
6	28153.341797	2	95.79	31.38	85.35
7	27698.808594	3	92.54	33.08	78.61
8	27264.914062	4	106.47	37.14	91.83
9	26851.412109	5	108.78	36.98	93.44
10	26459.140625	6	111.45	36.37	94.30

TABLE II: Predicted number of new cases on specific days (testing set).

	Predicted Cases	Day	Driving	Transit	Walking
0	21587.632812	1	107.63	36.92	94.60
1	21650.578125	2	109.14	37.39	96.94
2	21741.576172	3	111.61	37.37	96.33
3	21862.599609	4	117.88	38.99	100.52
4	22015.187500	5	133.68	40.66	113.19
5	22201.367188	6	131.07	41.16	119.76
6	22425.031250	0	105.40	37.39	91.85
7	22689.521484	1	115.92	43.32	99.64
8	22997.515625	2	118.49	43.35	101.36
9	23349.574219	3	121.12	43.35	102.98
10	23744.931641	4	126.30	42.92	106.25

Figure 8 compares the two sets of data.

Looking at Figure 8, the estimated data (shown in red) does, in fact, come closer to the observed data (shown in blue) in terms of where troughs and peaks appear. As such, the experiment can be said to have succeeded.

One drawback, however, is that the new method produces slightly less accurate results. The LSTM-regression model has an average error that is roughly 400-700 cases further away from the original data when compared to the LSTM model. The individual accuracy ratings as a percentage disregarding outliers are comparable however. The total difference is typically less than one-point-five percent and is therefore an acceptable loss of accuracy. Table 4 expresses error values for each piece of data for June of 2020 between the LSTM and LSTM regression model while the comparisons for the average error readings are shown in Table 3.

TABLE III: Average Error readings for average LSTM and LSTM Regression

Error	Cases	Percentage
Average LSTM Error	2969	11.6
Average LSTM-Regression Error	3470	12.5

VI. Conclusions

To conclude, the results of this study are indicative of the fact that planning through predictive analysis can be improved in order to facilitate the adoption of more effective social

TABLE IV: Performance comparison of LSTM and LSTM Regression

Value Predicted: LSTM	Value Predicted: LSTM-Regression	Value Observed: Original Data	Error Incurred: LSTM	Error Incurred: LSTM-Regression	LSTM Error: Percentage	LSTM-Regression Error: Percentage
22732	19188	19807	2925	619	14.8	3.1
22674	20020	21086	1588	1068	7.5	5.1
22632	20673	20544	2088	129	10.2	0.6
22607	22263	19699	2908	2564	14.8	13.0
22600	23098	21140	1460	1958	6.9	9.3
22613	24535	25178	2565	643	10.2	2.6
22648	18215	22223	425	4008	1.9	18.0
22708	19081	22302	406	3221	1.8	14.4
22794	19783	18822	3972	961	21.1	5.1
22911	21095	18665	4246	2430	22.7	13.0
23057	22095	20614	2443	1481	11.9	7.2
23238	22134	22883	355	749	1.6	3.3
23455	23824	25639	2184	1815	8.5	7.1
23712	19964	25540	1828	5576	7.2	21.8
24015	21747	19543	4472	2204	22.9	11.3
24366	20602	19957	4409	645	22.1	3.2
24767	21323	23705	1062	2382	4.5	10.0
25223	22581	25559	336	2978	1.3	11.7
25733	22121	27762	2029	5641	7.3	20.3
26301	23860	29909	3608	6049	12.1	20.2
26931	26686	34158	7227	7472	21.2	21.9
27624	32120	25793	1831	6327	7.1	24.5
28377	29930	31390	3013	1460	9.6	4.7
29185	28333	34720	5535	6387	15.9	18.4
30049	27097	34339	4290	7242	12.5	21.1
30966	26743	40949	9983	14206	24.4	34.7

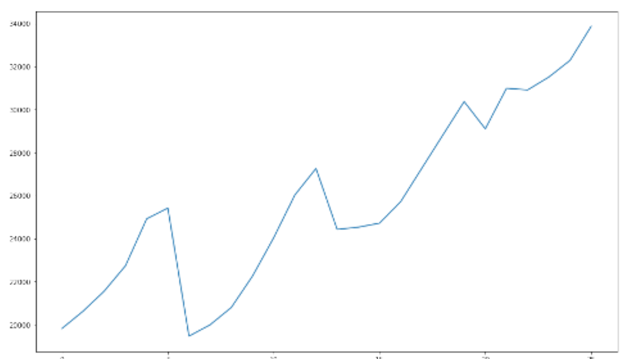

Fig. 7: Raw Regression prediction test output.

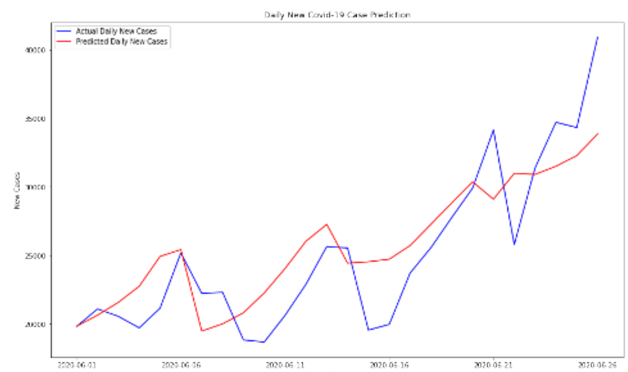

Fig. 8: Comparison of actual data and predicted data for predicting new daily coviv-19 cases.

distancing guidelines in times of pandemic disease outbreak. Natural, periodic peaks and troughs in data can be simulated well and establish accuracy losses well within the range of acceptability stated by this study. In the future, models such as these may be used to aid in policy making decisions in the context of pandemic response.

Other possible uses include informed economic policy making due to the fact that the stock market is often tracked using LSTM models similar to the ones used to track COVID-19 case data. The economic implications of this model are many. Feats such as the improved tracking of stock prices using public opinion scores and turnover rates of high-level executives, the prediction of minor recessions, and the monitoring of particularly volatile stocks are possible. The model could also be adapted for use in promoting influenza vaccination in historically underrepresented communities. Lastly, the model's accuracy could be further improved to aid the performance of all of the above and more.

References

[1] M. Nicola, Z. Alsafi, C. Sohrabi, A. Kerwan, A. Al-Jabir, C. Iosifidis, M. Agha, and R. Agha, "The socio-economic implications of the coronavirus pandemic (covid-19): A review," *International Journal of Surgery*, vol. 78, pp. 185–193, 2020.

[2] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, U. Reuter, T. Rabczuk, and P. M. Atkinson, "Covid-19 outbreak prediction with machine learning," *SSRN Electronic Journal*, 2020.

[3] V. K. R. Chimmula and L. Zhang, "Time series forecasting of covid-19 transmission in canada using lstm networks," *Chaos, Solitons, and Fractals*, vol. 135, p. 109864, 2020.

[4] S. University, "Why people didn't social distance," Apr 2020. [Online]. Available: https://news.stanford.edu/2020/04/14/people-didnt-social-distance/

[5] N. Ueda and R. Nakano, "Generalization error of ensemble estimators," *Proceedings of International Conference on Neural Networks (ICNN96)*, 1996.

[6] M. Painter and T. Qiu, "Political beliefs affect compliance with covid-19 social distancing orders," *SSRN Electronic Journal*, 2020.

[7] F. Ahmed, N. Zviedrite, and A. Uzicanin, "Effectiveness of workplace

social distancing measures in reducing influenza transmission: a systematic review," *BMC Public Health*, vol. 18, no. 1, 2018.

[8] "Similarities and differences between flu and covid-19," Jul 2020. [Online]. Available: https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm

[9] L. Matrajt and T. Leung, "Evaluating the effectiveness of social distancing interventions to delay or flatten the epidemic curve of coronavirus disease," *Emerging Infectious Diseases*, vol. 26, no. 8, 2020.

[10] R. Gall, "What is lstm?" Apr 2018. [Online]. Available: https://hub.packtpub.com/what-is-lstm/

[11] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, and et al., "Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, pp. 165–174, 2020.

[12] M. Vinceti, T. Filippini, K. J. Rothman, F. Ferrari, A. Goffi, G. Maffeis, and N. Orsini, "Lockdown timing and efficacy in controlling covid-19 using mobile phone tracking," *EClinicalMedicine*, p. 100457, 2020.

[13] S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, L. Y. Huang, A. Hultgren, E. Krasovich, and et al., "The effect of large-scale anti-contagion policies on the covid-19 pandemic," *Nature*, 2020.

[14] C. M. Weible, D. Nohrstedt, P. Cairney, D. P. Carter, D. A. Crow, A. P. Durnova, T. Heikkila, K. Ingold, A. Mcconnell, D. Stone, and et al., "Covid-19 and the policy sciences: initial reactions and perspectives," *Policy Sciences*, vol. 53, no. 2, pp. 225–241, 2020.

[15] "Covid-19 - mobility trends reports." [Online]. Available: https://www.apple.com/covid19/mobility

[16] U. Malik, "Time series analysis with lstm using python's keras library." [Online]. Available: https://stackabuse.com/time-series-analysis-with-lstm-using-pythons-keras-library/