COVID-19 Forecasting with Social Distancing Metrics

Gin Chairuangsang
Department of Computer Science
The University of Texas at Austin
ginchai@utexas.edu

Daniel Noble Hernandez
Department of Computer Science
The University of Texas at Austin
dnoble@cs.utexas.edu

Risto Miikkulainen
Department of Computer Science
The Universty of Texas at Austin
risto@cs.utexas.edu

ABSTRACT

The COVID-19 pandemic has killed over 1.5 million people globally as of this writing, and the virus continues to spread. Despite an approved vaccine set for mass immunizations, the COVID-19 pandemic case and death counts are at an all-time high; there remains a crucial need for accurate forecasting of cases into 2021 so as to inform policy and healthcare decisions relating to the crisis. The USA currently leads the world significantly in both case and death counts, and informed decision-making by American policymakers could reduce these counts to more tolerable levels, saving many lives over the next few months. A number of predictive models have been developed by distinguished research groups around the world since the beginning of the pandemic, with focuses on projections by country or - for American models - projections by state. This paper explores a predictive model that incorporates social distancing data from SafeGraph across census block groups in the USA as the primary features for the time-series forecasting of future case counts. Under the suspicions that accurate forecasting can be developed by considering the nuances in social distancing data on a smaller scale, and that local government can tailor decision-making based on specialized predictions, we aggregate social distancing metrics at the county level, rather than state- or nation-wide. Our model uses these metrics in a long short-term memory network architecture, which we then compare to a baseline model to judge the predictions, particularly for the surge that occurred in early November. Our experiments give promising results in predicting the November surge at the county level compared to those that do not consider social distancing metrics, suggesting that this data could be well-served to inform local decision-making going forward.

KEYWORDS: COVID-19, social distancing, time series, LSTM

1 INTRODUCTION

Over the course of human history, pandemics have ravaged the globe at different points in time - amongst the most prevalent in the common mind are the Black Death in medieval Europe, the Spanish flu of 1918, and now, COVID-19. Characteristic of these epidemics, and perhaps what has made them so dangerous, are their high rates of transmission. What distinguishes our current outbreak from these historical ones however, is the mortality; the Black Death killed an estimated hundreds of millions of people, effectively wiping out one quarter of the world, and the Spanish flu death toll numbered in the tens of millions. At the height of the COVID-19 pandemic and with inoculations beginning, the death toll is significantly lower; aside from the enormous strides we've made in medicine in recent history, part of this improved outcome comes from an in-depth understanding of how transmission occurs, and the necessary steps to reduce it. Given the astronomical amounts of data available to us in the modern day, predictive models for cases have been crucial to grasping the severity of the pandemic in advance, providing opportunities to take preventative measures; whether or not that has been achieved effectively is another discussion.

There have been many different predictive models built for COVID-19 since the beginning of the pandemic, projecting cases and deaths, seeking to provide insight into what measures should be taken, and even yielding projections dependent on these potential measures. In short, predictive models - prevalent examples of which have come out of many major institutions - have used a huge variety of data, spanned different geographic levels, and yielded variable results; we will briefly discuss some of the existing work further in the paper. However, in our survey of the literature, we have not come across concerted efforts to build models based solely on case counts and social distancing metrics at county levels in the US. There is ample reason to believe that these specific data would directly capture a representation of transmission rates, and propose a model trained on this data to directly project cases based on the public's location data.

In order to test this hypothesis, we built a predictive model using a long short term memory architecture, one of the foremost architectures for time series data. We initially ran the model on a univariate time series consisting of solely COVID-19 case data to serve as a baseline. We were then able to run the model on a multivariate time series, consisting of the same case data as well as social distancing metrics, and were able to compare it to the univariate benchmark to evaluate the usefulness of the added data. Results from the forecast that incorporated the social distancing metrics significantly outperformed the forecast that was trained solely on COVID-19 case data; we saw a marked reduction in error for variable selections of groups of counties over the time period that we chose. Because it is benchmarked on what is essentially a surrogate model that we created for this purpose, this opens an avenue to explore in the future, whereby we could replicate an established model and incorporate the social distancing metrics into it. This would be a more accurate reflection of the potential improvement than the current model we have independently developed. Nonetheless, the results are certainly indicators that there are subtleties to be gleaned from social distancing metrics that were not accurately represented by exclusively COVID-19 case data, at least in this case. If these results are maintained on a well-established model, then this could be a significant improvement to current predictors, both to be used in what is left of the pandemic and to contribute to scientific knowledge of epidemiological modeling.

2 BACKGROUND/RELATED WORK

Before diving into our experimental set-up and the results we generated, we will lay out some of the existing work and highlight models in the literature we believe are a robust sample of what has

been done. We'll first take a broad approach and discuss a framework proposed by a group at The University of Washington that aims to model possible trajectories assuming different conditions. After that, we'll examine a more localized proposed model of transmission that estimates the impact of social distancing measures on healthcare demand in Austin, TX.

2.1 Nationwide Projections

A logical focus of many models has been on projecting statewide case counts across the nation; to this end, we bring up and briefly analyze a paper that describes a deterministic SEIR framework that predicts cases assuming various social distancing mandates and mask use levels. An SEIR framework is a compartmental models used in epidemiology, whereby the population is assigned to compartments with labels - for SEIR, those labels are Susceptible, Exposed, Infectious, and Recovered. Each of these variables - S, E, I, R - represent a fraction of the population and are thus normalized to sum to 1[1]; determining the derivatives of each of these groups then allowed the researchers to define an equation for the basic reproductive number of transmission, which in turn led to projections of case counts [8]. To assess the efficacy of different measures on projected cases, the researchers established three boundary scenarios. The first forecast the expected cases if states continued to remove social distancing measures at the then current pace - the time of the writing was September. The second, considered the "plausible reference" scenario, the researchers modeled projections based on the assumption that states would shut down sufficiently to threshold the death rate whilst maintaining economic activity, characterized by social distancing measures for 6 weeks. The third "universal" scenario represented an ideal situation, wherein 95% of people would wear masks[10]. The results of this work - as expected, universal lead to the best results, and relaxing measures led to worse projections - were used to provide likely lower and upper bounds on the projected cases for each state in the hope of providing further insight into what variable measures we take would lead to. This end goal matches our own; though it differs in the specific data used and the scale at which projections are carried out, it is serves as a useful sample.

2.2 Impact of Measures on Healthcare Demand

Rather than focus on coming up with accurate forecasts of case counts going forward, some researchers have opted to model the spread and control of the disease. Here, we discuss a paper exploring a compartmental model that utilizes data on both contact rate and age-specific high risk group proportions to estimate the effects that two important interventions - social distancing measures school closures - may have on reducing transmission. This is measured indirectly by using these interventions to project the numbers of ventilator needs, hospitalizations, and ICU visits, on top of the cases and deaths[12]. These projections focused on the city of Austin, TX, but the authors argued that the model was widely applicable to other large metropolitan areas as well.

Based on the model results, the researchers concluded that school closures alone would do little to prevent the transmission of COVID-19; the curve would be flattened slightly, but the most effective outcome would be reached by some combination with social distancing[12]. Perhaps more importantly, valuable insight may be gleaned from this paper in that it not only provides some idea of projected cases and deaths, but also provides a measure of severity in terms of ventilator and ICU needs, and total hospitalizations. Although the authors showed that social distancing measures paired with school closures would lead to a reduction of case risks, this did not necessarily account for high-risk populations whose hospitalization would not be affected by such minimal social distancing measures. In the paper, the researchers go on to show that only at the 75% and 90% contact reduction scenarios were hospitalizations, ICU care, and ventilator needs brought below the estimated capacity for the city[12]. Given the ventilator shortage that has been and continues to be experienced in countries around the world - with doctors having to make difficult decisions of who will be put on a potentially life-saving ventilator - the pandemic has clearly created a vital need for officials to carefully evaluate projections for these and related needs[6]. In so doing, they might take measures that will minimize preventable deaths.

3 METHODS

This section describes some principles of developing a forecasting model that will be used for time-series prediction tasks. We discuss some crucial considerations and nuances in selecting proper strategies for a model in regard to the prediction problem. We follow this by briefly illustrating the evaluation metrics that the models will be measured against.

3.1 Predicting Cases with LSTM

Long Short-Term Memory (LSTM) networks are known to be one of the most effective sequence models used in practical applications [4]. These models include the long short-term memory and gated recurrent units; hence, they are a special type of gated Recurrent Neural Networks (RNNs). The core contribution of LSTM was in introducing self-loops that allow information from previous intervals to be remembered within the cell. Therefore, LSTM networks are well-suited to predict time series given time lags of unknown duration. They have been shown to be successful in some time series prediction tasks [3, 7].

We created a single LSTM model for the entire panel data consisting of time-series observations from multiple observational units (counties). This enabled us to avoid the problem of low sample sizes, simply because publicly available case data for COVID-19 in the US has been recorded on a daily basis. As a result, there are at most 286 observations per U.S. county from January 20, 2020, reported of the first confirmed case of COVID-19 in the United States [5], until November 1, 2020 - the cut-off date of the data used for our models. Stacking time-series from multiple observational units in turns provides the model the ability to learn from and generalize amongst data points from different regions. In this work, we limit our dataset to the top 250 counties where the top n counties refers to n counties with the greatest number of cumulative cases as of November 1, 2020. Figure 1 shows the sample of the top 10 counties from our dataset.

In the context of time series forecasting, one-step ahead prediction is usually the first technique to come to mind, given its early

Cases	County
309190	Los Angeles, California
193102	Cook, Illinois
186808	Miami-Dade, Florida
162807	Harris, Texas
159781	Maricopa, Arizona
103392	Dallas, Texas
86961	Broward, Florida
82966	Clark, Nevada
68233	Tarrant, Texas
68178	Riverside, California

Figure 1: Top-10 County with Total Cases as of November-01-2020

impact and popularity in the field of finance. However, forecasting COVID-19 cases is only useful in a multi-step setting, where a number of strategies can be considered. For instance, a recursive multi-step forecast runs a one-step ahead forecasting model multiple times, and then uses the prediction from the previous time steps as an input to the prediction for the next step. In this approach, prediction errors can quickly accumulate, resulting in degradation of accuracy as the prediction time horizon increases [2]. Instead, we used a multiple output strategy where the model predicts the entire sequence in a one-shot manner. Given a time series $X = \{\mathbf{x}^{(t-1)}, \mathbf{x}^{(t-2)}, ..., \mathbf{x}^{(t-n)}\}$ where each $\mathbf{x}^{(t)} \in R^m$ is an m-dimensional vector at time step t, a multiple output model generates a vector of predicted values $\{\hat{y}n^t, \hat{y}^{(t+1)}, ..., \hat{y}^{(t+p)}\}$ in one prediction, where $\hat{y}^{(t)}$ is a single scalar prediction at time step t.

Naturally, multivariate time series contain features with varying ranges of values. When fitting a model on unscaled data, it is possible for the data to slow down the learning and convergence of the network, hence, making the training process inefficient. We used min-max normalization technique by scaling each feature into a range between 0 and 1. Subsequently, this requires the use of an output activation function for LSTM networks that encompass the same range.

3.2 Evaluation Metrics

Accuracy of the forecasting models can be measured by comparing predicted values and observed values. In this work, our models were evaluated using three measurements for comparison. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error are expressed by the following equations, respectively.

$$MAE = \frac{\sum_{t=1}^{n} |y^{(t)} - \hat{y}^{(t)}|}{n} \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y^{(t)} - \hat{y}^{(t)})^2} \tag{2}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|y^{(t)} - \hat{y}^{(t)}|}{y^{(t)}} \tag{3}$$

where $y^{(t)}$ is an observation at time step t.

4 EXPERIMENTS

When discussing the experiments we carried out, we outline data preparation, including preprocessing techniques that were applied prior to running our experiments. We describe how forecasting models were formulated including considerations of network architectures and parameters, model training and evaluating processes are reported and ultimately, we present our results, comparing forecasts between different models.

4.1 Setup

Data Preparation. We used publicly available county-level COVID-19 data published by The New York Times [11]. The dataset contains historical daily cases for each county in the U.S., beginning with the first reported case in Snohomish county in Washington State on January 1, 2020. We discarded observations whose county was recorded as "Unknown".

In conjunction with the COVID-19 dataset, we used the Social Distancing Metrics dataset provided by SafeGraph, a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places. To enhance privacy, SafeGraph excludes census block group information if fewer than five devices visited an establishment in a month from a given census block group [9]. Because the observational unit (census block group) of this dataset is more granular than that of the COVID-19 dataset, aggregating the metrics into a county-level was required prior to merging the two datasets together.

We applied a 7-day moving average to the dataset to smooth-out the varying seasonality components, which we found to have caused significant performance degradation in our models. Lastly, a daily new cases variable was calculated and added to our dataset.

We filtered our dataset to only include the top 250 counties with the highest number of total cases as of November 1, 2020 and selected the observations whose date ranges from May 1, 2020 to October 30, 2020 for our experiment. To evaluate our forecasting models, we reserve observations from November 1, 2020 to November 28, 2020 (inclusive) as our main "forecasting" period.

LSTM with Univariate Time Series. In order to fairly evaluate prediction power of Social Distancing Metrics, we first separately trained an LSTM network with a univariate time series using only the COVID-19 cases dataset. The variable of interest was daily new cases. Number of previous observations for an input sequence was 28, and the number of predicted values for an output sequence was set at 28.

LSTM with Multivariate Time Series. For this model, in addition to COVID-19 cases, we incorporated social distancing metrics and trained a second LSTM network using this multivariate time series. SafeGraph provided over 17 additional features. However, with limiting resources, we simply could not train our model with all the possible combinations. We first handpicked 5 features, and after iterating over different combinations, the set of additional features that yielded the most stable and lowest training loss were

"device count", "percentage of devices completely home", and "median home dwell time". The predicted variable, number of lagged observation, and predicted values all remained the same.

Table 1: Model parameters

input features	1 (Univariate), 4 (Multivariate)
time step	28
hidden layers	1
units in hidden layers	50
batch size	56
sequence length	56
prediction length	28
learning rate	0.001
training epochs	1800 (Univariate), 2600 (Multivariate)
optimizer	Adam

Network Architecture and Training Parameters. For each of the training experiments, we performed a hyper-parameter search amongst different network structures (e.g. shallow, deep), activation functions (e.g. *tanh*, *relu*, *sigmoid*), learning rates, batch sizes, numbers of lagged observations, and numbers of predicted values. We found that lowering the number of predicted values naturally improves models performance. However, we kept the number at 28, primarily because decision-makers often seek to make changes that have a longer impact than 7 or 14 days, especially in the context of the COVID-19 outbreak. We let each model train for 3000 epochs whilst checkpointing the weights and losses improvement every 100 epochs. We picked the final weights for a model using those from the earliest epoch such that its subsequent checkpoint showed no improvement in training loss. It is worth noting that we were able to achieve significant improvement in the stability of training and validation losses for both models after reducing the learning rate from 0.01 to 0.001. The summary of our models' architecture and parameters is shown in Table 1

4.2 Results

After forecasting new cases of COVID-19 from November 1, 2020 to November 28, 2020, we calculated the errors and took averages across different numbers of counties as shown in Table 2. Averaging across the top 5 counties - and judging based on MAE - the prediction errors from the model with Social Distancing Metrics are roughly 55% lower than those of the model without. We see similar drops in error when comparing the RMSEs and MAPEs of the univariate and multivariate group. The reduction in forecasting MAE trails down to 50% and 35% when looking at the top 10 and top 20 counties respectively. Once again, we observe the same pattern followed for both RMSE and MAPE. Although we noticed substantial reductions in error when grouping up to 20 counties, there was no significant difference in errors when averaged across all 250 counties.

Sample forecast plots for different counties are illustrated in Figure 2. It is visually clear that the model trained on multivariate data with both historical case data and social distancing metrics was able to perform significantly better than the model that was trained solely on the historical data. On the left hand side, for each of the

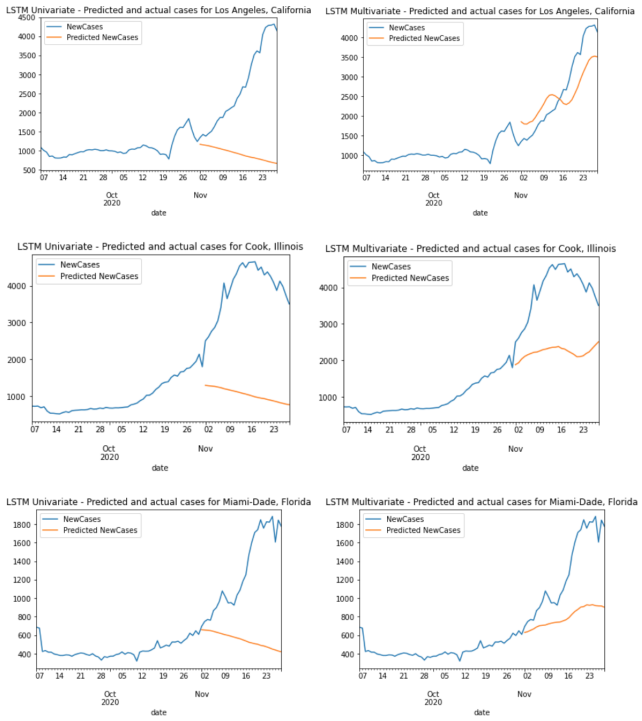

Figure 2: Forecasts for Top-3 Counties. Left: LSTM Univariate. Right: LSTM with Social Distancing Metrics

top 3 counties, the univariate model failed to predict the second wave surges of COVID-19 that we experienced in November. In contrast, the right side of the figure shows that the multivariate model continues to forecast an upward trend in new cases.

Figure 3 shows a number of resulting forecasts that reflect increases and decreases in social distancing in Tarrant, Harris, and Dallas counties in Texas, all of which find a place in the top 10 counties by case count. These plots - with the left of the figure showing forecasts for cases with 50% more stay home devices and

Table 2: Evaluation Metrics for New Cases of COVID-19 with Different Numbers of County

Top-n County	Model	MAE	RMSE	MAPE%
5	LSTM Univariate	1328	1505	53
	LSTM Multivariate	**602**	**660**	**24**
10	LSTM Univariate	883	1017	48
	LSTM Multivariate	**454**	**512**	**27**
20	LSTM Univariate	595	696	47
	LSTM Multivariate	**391**	**446**	**36**
250 (all)	LSTM Univariate	152	177	42
	LSTM Multivariate	**142**	**163**	42

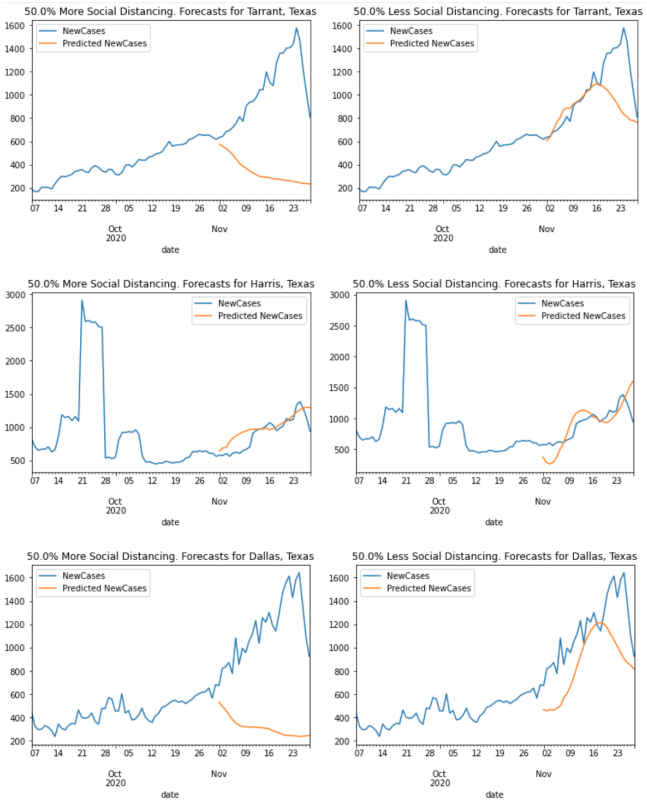

Figure 3: Forecasts for Texas Counties. Left: 50% More Stay-Home Devices. Right: 50% Less Stay-Home Devices

the right showing forecasts for cases with 50% less stay home devices - suggest that the model was able to learn useful implications of stay-home versus non-stay-home behavior.

5 DISCUSSION AND FUTURE WORK

In this section, we discuss the significance of these results, how they contribute to the existing work in the field, and then discuss the implications of the work and what future extensions may arise from it naturally.

5.1 Evaluation of Results and Contribution

In evaluating what we have established, we first turn to the significant reduction of errors from the top 5, top 10, and top 20 counties; in these groupings, the hypothesis that was put forth at the beginning of this paper - namely, that social distancing metrics will make a significant contribution to the predictive power of COVID-19 case forecasting models - is heavily supported. Compared to baseline univariate models that consider only case counts, it is clear that our multivariate model significantly outperformed it, across the 3 groupings mentioned above, as shown in the plots and error table. Considering these results qualitatively, they are results that follow naturally; with a disease that is virulently contagious and whose elimination could be directly tied to the severing of the chain of transmission, we turn to data that reflects those human connections directly. Although case counts are meaningful data to project future cases, social distancing metrics accurately capture individuals' location data and most fundamentally represent future transmission of the virus and the resultant cases. There does, however, remain to be addressed the fourth entry into Table 2; when considering the top 250 counties, the model's predictive power does not improve upon incorporation of social distancing metrics. To us, this suggests that COVID-19 cases in the counties faced with severe outbreaks find themselves more subject and sensitive to social distancing behaviour. In particular, social distancing as measured by the proportion of the public staying at home plays a significant role in these hard-hit regions compared to those with a relatively minute number of cases.

A question now arises on how this research in particular fits in with and contributes to the current wealth of knowledge and work on the COVID-19 pandemic. Though this can be considered yet another model seeking to accurately forecast case counts in the US, what sets it apart lies in the data it uses. Our model is primarily trained directly on social distancing metrics, along with case counts, in an effort to capture nuances in data that we may have otherwise missed with other data. Further, it aggregates census block groups into counties and yields promising results at this geographic level in the most hard-hit counties that need it the most.

5.2 Future Work

The results we generated lead us to an open avenue for further work. This paper delineates experimentation with a univariate forecasting model that is used to establish a baseline, and then goes on to compare this with a multivariate forecasting model that additionally incorporates the data of interest. Essentially, what we have done is create a surrogate to serve as a benchmark, and our results show that compared to this constructed benchmark, the model of interest performs well. This is, in and of itself, a desirable and promising outcome, but by itself, it is not enough to allow us to categorically state that it will hold up in actuality. A logical next step would be to take an established and already widely-cited predictive model and modify it to incorporate social distancing metrics. If there is a reduction in error at that stage, then we can safely say that our results are impactful in consideration of future decision-making. Nonetheless, the results we generated lend considerable support to the efficacy of this framework.

6 CONCLUSION

Existing research on the COVID-19 pandemic has spanned all manner of applications - from projecting healthcare demands to numbers of cases and deaths on a national or global scale - and has varied widely in terms of the frameworks and data used, including machine learning, epidemiological, and statistical models. Now, this work occupies its own niche in the existing database of prior work - a forecasting LSTM model trained on social distancing data that aims to capture nuanced features at a county level. Stepping out to a high level, this research has the potential to positively impact localized COVID-19 pandemic responses going forward. Although a vaccine has been developed and the end is perhaps in sight, there is no room for complacency in the remainder of the epidemic. In the months that lie ahead whilst vaccines are administered, lax

measures resulting from a false sense of security could prove to be devastating, and continue to claim many lives. Further extensions of this social distancing metric-based predictive model, if proven as effective as our work suggests, could stand to provide valuable insights into the future severity of the pandemic in individual counties. These counties, particularly those of densely-packed cities, will have very different outcomes half a year from now, depending on the decisions that their local leadership makes; we hope this will contribute to public-centric, responsible decision-making.

REFERENCES

[1] UC Santa Barbara. [n.d.]. *An SEIR model.* https://sites.me.ucsb.edu/~moehlis/APC514/tutorials/tutorial_seasonal/node4.html.

[2] Jason Brownlee. 2020-12-03. *Deep Learning with Time Series Forecasting.* https://machinelearningmastery.com

[3] Yue-Shan Chang, Hsin-Ta Chiao, Satheesh Abimannan, Yo-Ping Huang, Yi-Ting Tsai, and Kuan-Ming Lin. 2020. An LSTM-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research* 11, 8 (2020), 1451 – 1463. https://doi.org/10.1016/j.apr.2020.05.015

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press. http://www.deeplearningbook.org.

[5] Michelle L. Holshue, Chas DeBolt, Scott Lindquist, Kathy H. Lofy, John Wiesman, Hollianne Bruce, Christopher Spitters, Keith Ericson, Sara Wilkerson, Ahmet Tural, George Diaz, Amanda Cohn, Le Anne Fox, Anita Patel, Susan I. Gerber, Lindsay Kim, Suxiang Tong, Xiaoyan Lu, Steve Lindstrom, Mark A. Pallansch, William C. Weldon, Holly M. Biggs, Timothy M. Uyeki, and Satish K. Pillai. 2020. First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine* 382, 10 (5 March 2020). https://doi.org/10.1056/NEJMoa2001191

[6] Karthikeyan Iyengar, Shashi Bahi, Raju Vaishya, and Abhishek Vaish. 2020. Challenges and solutions in meeting up the urgent requirement of ventilators for COVID-19 patients. *Diabetes Metab Syndr* 14, 4 (5 may 2020). https://doi.org/10.1016/j.dsx.2020.04.048

[7] J. Kumar, Rimsha Goomer, and Ashutosh Kumar Singh. 2018. Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters. *Procedia Computer Science* 125 (2018), 676–682.

[8] Samuel Mwalili, Mark Kimathi, Viona Ojiambo, Duncan Gathungu, and Rachel Mbogo. 2020. SEIR model for COVID-19 dynamics incorporating the environment and social distancing. *BMC Res Notes* 13, 352 (23 jul 2020). https://doi.org/10.1186/s13104-020-05192-1

[9] SafeGraph. 2020-11-01. . https://www.safegraph.com/.

[10] IHME COVID-19 Forecasting Team, Robert C. Reiner, and Ryan M. Barber. 2020. Modeling COVID-19 scenarios for the United States. *Nat Med* (23 oct 2020). https://doi.org/10.1038/s41591-020-1132-9

[11] The New York Times. 2020-11-01. *Coronavirus (Covid-19) Data in the United States.* https://github.com/nytimes/covid-19-data.

[12] Xutong Wang, Remy F. Pasco, Zhanwei Du, Michaela Petty, Spencer J. Fox, Alison P. Galvani, Michael Pignone, S. Claiborne Johnston, and Lauren Ancel Meyers. 2020. Impact of Social Distancing Measures on Coronavirus Disease Healthcare Demand, Central Texas, USA. *Emerging Infectious Diseases* 26, 10 (oct 2020). https://doi.org/10.3201/eid2610.201702