Using Mobility Data to Understand and Forecast COVID19 Dynamics

Lijing Wang[1*] , **Xue Ben**[2*] , **Aniruddha Adiga**[1*] , **Adam Sadilek**[2] , **Ashish Tendulkar**[2] ,
Srinivasan Venkatramanan[1] , **Anil Vullikanti**[1] , **Gaurav Aggarwal**[2] , **Alok Talekar**[2] ,
Jiangzhuo Chen[1] , **Bryan Lewis**[1] , **Samarth Swarup**[1] , **Amol Kapoor**[2] ,
Milind Tambe[2,3] , **Madhav Marathe**[1]
[1]University of Virginia, [2]Google, [3]Harvard University
Email: {lw8bn,aa5dw,srini,vsakumar,chenj,brylew,swarup,marathe}@virgina.edu,
{sherryben,sadilekadam,ashishvt,gauravaggarwal,atalekar,ajkapoor}@google.com,
{milind_tambe}@harvard.edu

Abstract

Disease dynamics, human mobility, and public policies co-evolve during a pandemic such as COVID-19. Understanding dynamic human mobility changes and spatial interaction patterns are crucial for understanding and forecasting COVID-19 dynamics. We introduce a novel graph-based neural network(GNN) to incorporate global aggregated mobility flows for a better understanding of the impact of human mobility on COVID-19 dynamics as well as better forecasting of disease dynamics. We propose a recurrent message passing graph neural network that embeds spatio-temporal disease dynamics and human mobility dynamics for daily state-level new confirmed cases forecasting. This work represents one of the early papers on the use of GNNs to forecast COVID-19 incidence dynamics and our methods are competitive to existing methods. We show that the spatial and temporal dynamic mobility graph leveraged by the graph neural network enables better long-term forecasting performance compared to baselines.

1 Introduction

The COVID-19 pandemic has affected almost every country in the world and has resulted in an unprecedented response by governments across the world to control its spread. The social distancing measures are one of the most effective non-pharmaceutical interventions at this stage. The social distancing measures have led to significant change in human mobility behaviour while the mobility change has also affected the disease dynamics inevitably. To better understand COVID-19 dynamics and help to control the disease spread, it is crucial and challenging to provide accurate and timely spatio-temporal forecasting of epidemic dynamics. As machine learning and artificial intelligence(AI) has been successful in many domains, there is an urge to investigate how we can leverage AI-based technologies for infectious disease understanding, modeling, forecasting, and controlling. In this work, we focus on applying AI-based techniques to solve

the above challenge by incorporating a new large-scale aggregated spatio-temporal mobility data into graph-based neural networks. Using aggregate mobility data to understand COVID-19 dynamics has received wide interests recently. There have been a number of recent studies along these lines, for example, in China using Baidu data [Chinazzi *et al.*, 2020], in the US using mobility data [Kraemer *et al.*, 2020; Adiga *et al.*, 2020b], and at a global scale using airline traffic [Adiga *et al.*, 2020a]. On the other side, lots of COVID-19 forecasting methods are proposed since the initial outbreak early this year, such as mechanistic methods [Yang *et al.*, 2020; Anastassopoulou *et al.*, 2020; Kai *et al.*, 2020], and time series methods using statistical regression models [Ribeiro *et al.*, 2020] or deep learning model [Ramchandani *et al.*, 2020]. However few of the existing works investigate spatio-temporal forecasting using graph neural network (GNN) with integrated real-time mobility data. In our work, we introduce a novel method to incorporate global population mobility flows into graph-based spatial-temporal neural networks for COVID-19 dynamic forecasting. Our major contributions are:

- We analyze the joint effects of social-distancing guidelines and mobility patterns at US state levels using an integrated map of mobility flows (MF), COVID-19 Surveillance data and data on social distancing guidelines;

- We design a dynamic mobility informed GNN that considers both temporal dynamics and cross-location co-evolution dynamics using a recurrent message passing (RMP) module to recurrently embed information from a node's neighbors;

- We also design multiple variants of the proposed model which use static mobility graph, geographical adjacency graph, and attention-based trainable graph;

- We evaluate the proposed model on forecasting the US state level daily new cases, and demonstrate that the dynamic spatial and temporal mobility informed GNN allows for better forecasting performance compared with its variants as well as several existing classic and state-of-the-art time series methods.
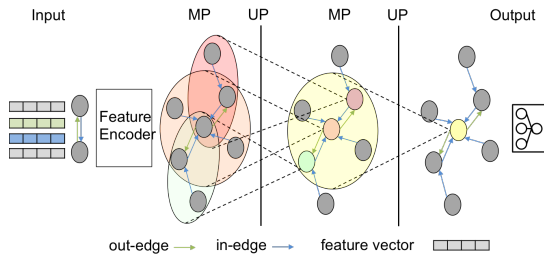
---

*Equal contribution

Figure 1: An example of two-hop RMP architecture. Temporal node feature and edge feature vectors are encoded using the feature encoder module. A two-hop RMP module is used to further embed spatio-temporal information to hidden representations. The output module makes the final predictions.

2 Methods

2.1 Problem formulation

Assuming we have N regions in total. Each region is associated with time series of multiple observed features, e.g. surveillance cases, in a time window T, where T is the observation duration. It could be of weekly or daily granularity depending on the data resolution. We define a dynamic graph of N regions as $G(\mathcal{V}, \mathcal{E}, \mathcal{T})$, where $\mathcal{V}$ is the set of N nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, and $\mathcal{T}$ is the set of T time points. A node v_i at time t is attributed with $\boldsymbol{d}_{v_i,t} \in \mathbb{R}^{D_v}$ where D_v is the node feature numbers. An edge $e_{ij} \in \mathcal{E}$ at time t connecting nodes v_i and v_j is weighted by either adjacency matrix or mobility flow matrix, is attributed with $\boldsymbol{d}_{e_{ij},t} \in \mathbb{R}^{D_e}$ where D_e is the edge feature numbers.

2.2 Mobility informed graph neural networks

Constructing the graph. We construct a dynamic mobility graph $G(\mathcal{V}, \mathcal{E}, \mathcal{T})$, where each node feature $\boldsymbol{d}_v$ includes a sequence of dynamic observations regarding the the region in a history window $H \leq T$. We include daily new case count, new death count and intra-region mobility flow ($f_{ii}(t)$ which represents the MF from v_i to v_i during time t) as the node features, i.e. $\boldsymbol{d}_{v_i,t} \in \mathbb{R}^{H \times 3}$. The graph edge features are derived from the inter-region mobility by aggregating Google mobility data to the state or county level. At a certain time point t, if there is any human movement from region j to region i in the past H days, we add a directed edge e_{ji} that connects region j and i, and associate it with the inter-region mobility flow $f_{ji}(t)$ and flow of active cases from source region (defined as $f_{ji}^{active}(t) = \frac{N_j^{active}(t)}{N_j^{popu}} * f_{ji}(t)$) as the edge feature i.e. $\boldsymbol{d}_{e_{ji},t} \in \mathbb{R}^{H \times 2}$ where $N_j^{active}(t)$ is the number of active cases (cumulative cases minus recovered cases and deaths) and N_j^{popu} is the population of region j.

Feature encoding. In the graph, node feature vectors and edge feature vectors include temporal information from the past. We encode the vectors using a Long Short-Term Memory (LSTM) module. At time step $t \in \mathcal{T}$, for each feature vector $\boldsymbol{d}_{v_i,t}$ or $\boldsymbol{d}_{e_{ij},t}$, the LSTM module encodes the vector into hidden representations as:

$$
\begin{aligned}
\boldsymbol{h}_{i,t} &= p(\boldsymbol{d}_{v_i,t}, \theta^v) \in \mathbb{R}^{H^n}, \\
\boldsymbol{a}_{ji,t} &= p(\boldsymbol{d}_{e_{ji},t}, \theta^e) \in \mathbb{R}^{H^e}
\end{aligned}
\tag{1}
$$

where p denotes LSTM cell computation, H^n and H^e are hidden dimension of node feature and edge feature, θ^v and θ^e are parameters to be learned.

Spatio-temporal message passing. A region's COVID-19 dynamics can potentially be affected by regions where frequent travels occur between them. This resembles the core insight behind graph neural network models, i.e. the transformation of the input node's signal can be coupled with the propagation of information from a node's neighbors in order to better inform the future hidden state of the original input. This is most evident in the unified message-passing framework proposed by [Gilmer *et al.*, 2017]. In our model, we design an *Recurrent Message Passing* (RMP) module to recurrently pass the hidden representations from a node's neighbors to the current node. As shown in Figure 1, the RMP module has two phases: the message passing (MP) phase and the update phase (UP). It runs for L rounds, so any node in the graph is taking into account of neighbors that are L hops away. To be more specific, at time t, given a node v_i at a certain round $(l + 1)$: In the MP phase, for each node pair (v_i, v_j) that $v_j \in \mathcal{N}(i)$ where $\mathcal{N}i$ denotes neighbors of v_i, we first combine the node hidden states $\boldsymbol{h}_{i,t}, \boldsymbol{h}_{j,t}$ and in-edge hidden representations $\boldsymbol{a}_{ji,t}$ from previous round l using a message passing function f to get a hidden state $\boldsymbol{a}_{ji,t}$ at current round $l + 1$. It will later be aggregated (we use mean operation but can be sum, max, etc.) together over all pairs to obtain a message $\boldsymbol{m}_i$ for node v_i. In UP phase, we use a node upate function g to update the the node hidden states. The hidden states of the node v_i at the $(l + 1)$th round $\boldsymbol{h}_{i,t}^{l+1}$ are updated in RMP module as :

$$
\begin{aligned}
\boldsymbol{a}_{ji,t}^{(l+1)} &= f\big(\boldsymbol{h}_{i,t}^{(l)}, \boldsymbol{h}_{j,t}^{(l)}, \boldsymbol{a}_{ji,t}^{(l)}, \theta^f\big) \in \mathbb{R}^{H^e}, \\
\boldsymbol{m}_{i,t}^{(l+1)} &= \sum_{j \in \mathcal{N}(i)} \boldsymbol{a}_{ji,t}^{(l+1)} \in \mathbb{R}^{H^m}, \\
\boldsymbol{h}_{i,t}^{(l+1)} &= g\big(\boldsymbol{h}_{i,t}^{(l)}, \boldsymbol{m}_{i,t}^{(l+1)}, \theta^g\big) \in \mathbb{R}^{H^n}
\end{aligned}
\tag{2}
$$

where θ^f and θ^g are module parameters to be learned, $\mathcal{N}(i)$ denotes the neighbors of v_i where there exists e_{ji}, f is the message passing function that uses a multilayer perceptron (MLP) and g is the node update function that uses Gated Recurrent Unit (GRU), $\boldsymbol{m}^{(l+1)}$ is the messages passed between nodes.

Output layer. We feed the hidden representations to the output layer for the final predictions:

$$
\hat{y}_{i,t} = o\big(\boldsymbol{h}_{i,t}^{(L)}, \theta^o\big) \in \mathbb{R}
\tag{3}
$$

where θ^o is the parameters to be learned, o is the output function which is a multilayer perceptron (MLP) in our model.

Forward passing process. As shown in Figure 1, we first feed the sequences of temporal node and edge features through the feature encoding module to obtain node and edge

---

Algorithm 1: MF informed GNN forward passing

Input: Time series of COVID-19 surveillance data for N regions, Google mobility flow data among N regions.

1 **for** *each time step t* **do**
2   **for** *each region i* **do**
3      $h_{i,t} \leftarrow \text{FtrEncode}(d_{v_i,t})$
4   **for** *each region pair (i,j)* **do**
5      $a_{ji,t} \leftarrow \text{FtrEncode}(d_{e_{ji},t})$
     ▷ Simultaneous calculations for all regions
6   **for** *each region i* **do**
7   **for** *l in $0, \dots, L-1$* **do**
8        $h_{i,t}^{l+1} \leftarrow \text{RMP}(h_{i,t}^{(l)}, h_{j,t}^{(l)}, a_{ji,t}^{(l)})$
9      $\hat{y}_{i,t} \leftarrow \text{Output}(h_{i,t}^{(H)})$

---

embedding, which are utilized as the initial node and edge hidden representations for the RMP module. Then we perform MP and UP computations for L rounds. This is the core step to allow one region to leverage information from its neighbors and their connectivity in between. The output module will output the final predictions.

Proposed models. The proposed model aims to examine the effect of dynamic mobility on understanding and forecasting COVID-19 dynamics. Thus we design several variants of the proposed model using dynamic mobility graph denoted as GNN-dmob , using a static mobility graph denoted as GNN-smob , using a geographical adjacency graph denoted as GNN-adj , and using an attention-based matrix denoted as GNN-att . The details are listed below.

- **GNN-dmob** the proposed model with dynamic mobility graph.

- **GNN-adj** uses the same graph structure with GNN-dmob but remove intra-region mobility flow from node features i.e. $d_{v_i,t} \in \mathbb{R}^{H \times 2}$, and construct an adjacency matrix by adding an edge e_{ij} if region j is a neighbor of region i with edge weight 1. The adjacency matrix is normalized by row summation. It is static across time steps, thus the edge feature is $d_{e_{ij},t} \in \mathbb{R}$.

- **GNN-smob** is similar to the GNN-adj but obtained by replacing the adjacency matrix with a static mobility graph which is an average of mobility graphs from March 1st, 2020 to August 2nd, 2020.

- **GNN-att** is inspired by cola-GNN proposed in [Deng *et al.*, 2019]. Instead of using a physical matrix in our model, we implement an attention-based model that allows the model to learn an attention matrix of all the regions. In MP phase, we update the message between nodes as:

$$m_{i,t}^{(l+1)} = \eta\Big(\sum_{j \in \mathcal{N}} a_{ji,t}^{(l)} h_{j,t}^{(l)}, \theta^\eta\Big) \tag{4}$$

where θ^η is the trainable parameters, a_{ji} is the attention coefficient defined as: $a_{ji,t}^{(l)} = v^T \sigma(W^i h_{i,t}^{(l)} + W^j h_{j,t}^{(l)} +$

$b) \in \mathbb{R}$ where $v \in \mathbb{R}^{H^a}$, $W \in \mathbb{R}^{H^a \times H^f}$, $b \in \mathbb{R}^{H^a}$ are trainable parameters, σ is Rectified Linear Units (ReLU) applied at element-wise.

3 Experiments

3.1 Datasets

Google COVID-19 Aggregated Mobility Research Dataset, which contains the anonymized relative MF aggregated weekly over users within a 5 km^2 cell. **COVID-19 surveillance data (CSD)** via the UVA COVID-19 surveillance dashboard [UVA, 2020 accessed August 29 2020]. It contains daily confirmed cases and death count worldwide. The data is available at the level of a county and state in the US. Daily case counts and death counts are further aggregated to weekly counts.

3.2 Metrics

The metrics used to evaluate the forecasting performance are: *root mean squared error (RMSE)* and *Pearson correlation (PCORR)*.

3.3 Baselines

We implemented several classic and state-of-the-art time series models as the comparison methods. They are Naive (uses the observed value of the most recent time point as the future prediction) , Autoregressive (AR), Autoregressive Moving Average (ARMA), Long Short-term Memory (LSTM), CNNRNN [Wu *et al.*, 2018], and cola-GNN [Deng *et al.*, 2019].

3.4 Settings and implementation details

The weekly mobility graph is expanded to daily by dividing the weekly values by 7. The training data set is from March 1st to August 1st (125 days), the testing set is from August 2nd to August 29th (28 days). We make 2, 7, 14, 21, and 28 days ahead forecasting for each data point in the testing set. For all models, the historical window $H = 28$. For GNN-dmob , we use a single layer LSTM for feature encoding with 16 units, a two-layer MLP in MP phase with 32 and 16 units, and a single layer GRU in UP with 16 units. The same settings are used for GNN-smob , GNN-att , and GNN-adj . AR and ARMA use AR order 28 and ARMA uses MA order 2. CNNRNN and cola-GNN set with their best parameter settings in the original paper. We set batch size as 32, epoch number as 1000. MSE loss and Adam Optimizer with default settings and early stopping with patience of 100 epochs are used for all model training. All results are average of 5 random runs.

3.5 Research findings

In this section, we present key findings and results.

Spatio-temporal analysis of MF patterns and COVID-19 dynamics in the US

In order to analyze the human mobility and COVID-19 dynamics during different phases of the pandemic, we use a 4-week window and moving one week ahead each time to compute Pearson correlation between new confirmed cases
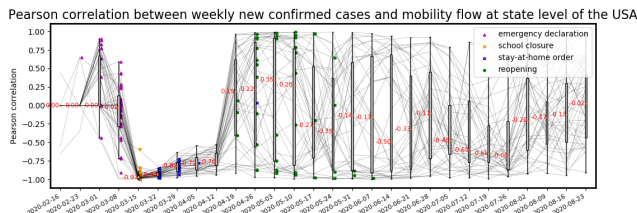
Figure 2: Pearson correlation between MF and new confirmed cases, together with state level social distancing mandates including emergency declaration (purple), school closure (orange), and stay-at-home (blue) are marked. A boxplot is used to display variation in samples of 53 states each week. The median value is shown along with the median line.
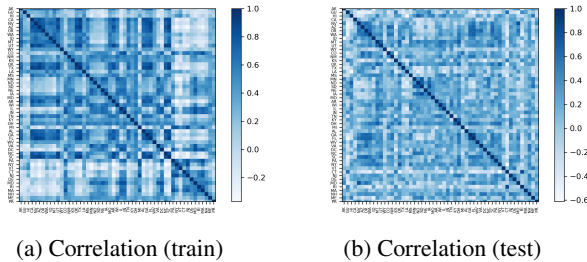

(a) Correlation (train) (b) Correlation (test)

Figure 3: Heatmap of Pearson correlation matrix of state level time series of new confirmed cases. (3a) training data and (3b) testing data. We can observe that the pattern of correlations between states changed dramatically from training dataset to testing dataset.

Table 1: RMSE and PCORR performance of different methods on the US state dataset with horizon = 2, 7, 14, 21, and 28. The values are average of 5 runs. Bold face indicates the best results of each column.

	US state				
RMSE(↓)	2	7	14	21	28
Naive	411	389	445	496	525
AR	376	634	866	975	978
ARMA	400	637	944	1107	1186
LSTM	368	504	517	567	605
CNNRNN	416	432	512	606	659
cola-GNN	320	502	451	530	714
GNN-adj	**310**	411	407	412	513
GNN-att	319	479	985	457	745
GNN-smob	313	405	410	**406**	510
GNN-dmob	313	**330**	**350**	445	**465**
PCORR(↑)	2	7	14	21	28
Naive	0.106	0.361	0.310	0.261	0.157
AR	0.283	0.282	0.044	-0.017	-0.097
ARMA	0.281	0.260	0.071	-0.078	-0.133
LSTM	0.227	0.305	0.287	0.226	0.204
CNNRNN	0.211	0.250	0.265	0.248	0.029
cola-GNN	0.366	0.341	0.247	0.175	0.176
GNN-adj	0.293	**0.395**	0.319	0.315	**0.280**
GNN-att	**0.376**	0.256	0.087	0.136	0.238
GNN-smob	0.321	0.382	0.297	**0.294**	0.226
GNN-dmob	0.298	0.344	**0.320**	0.287	0.216

and MF within the window. Figure 2 shows the Pearson correlation along the weeks. We observe that mobility flow and new confirmed cases show very high negative correlation (median -0.97) for almost all states during March when most of states mandated school closure and stay-at-home orders. This indicates that COIVD-19 dynamics and human mobility are highly correlated. Starting from the mid April when the states started to reopen to some degree, there is a large variation in correlation values, which indicates that COVID-19 dynamics varies a lot due to that it is affected by multiple complicated factors like local population size, individual behaviours (e.g. wearing a mask or not in public location), and government reopening guidelines.

Forecasting performance and discussion
The proposed GNN-dmob model is evaluated w.r.t. forecasting daily new confirmed cases at US state level for 2, 7, 14, 21 and 28 days ahead. Table 1 presents the RMSE, and PCORR performance averaging across 53 states and 28 days. In general, we observe that GNN-dmob has better RMSE performance than the comparisons for long-term forecasting. The best performances on PCORR are evenly distributed among the proposed models. The results indicate that our proposed methods can capture the disease dynamic in both short-term and long-term. Naive baseline outperforms the other baselines for 7, 14, 21, 28 days ahead forecasting. This is not surprising because the testing data is from August during when the time series of all states are showing downward trends with small changing rates. The Naive assumes certain level of

regularity in the time series leading to good forecasting performance. DNN-based models perform better than AR-based models especially on long-term forecasting which indicates that DNN-based models have better generalization capability for forecasting unseen data. In our experiments, attention-based models cola-GNN and GNN-att are not outstanding for both short and long-term forecasting. A possible reason is that the learned attention coefficients are outdated due to the fast evolution in COVID-19 dynamics between training time period and testing time period (see Figure 3), which leads to false attention by regions while predicting. The proposed method GNN-dmob explicitly projects the most recent mobility patterns to the future potentially leading to better performance.

4   Conclusion

The paper introduces a novel GNN framework to incorporate aggregated mobility flows for better understanding the impact of human mobility on COVID-19 dynamics as well as better forecasting of disease dynamics. We propose a recurrent message passing GNN to embed spatio-temporal disease dynamics (COVID-19 surveillance data) and human mobility dynamics (MF data) while making forecasting. The experiment results of forecasting daily COVID-19 new cases for each state in the US demonstrate the additional improvements obtained by using the mobility data. The use of GNN for COVID-19 is just beginning and our results are some of the first results in this area to yield good performance.

Acknowledgments

References

[Adiga *et al.*, 2020a] Aniruddha Adiga, Srinivasan Venkatramanan, James Schlitt, Akhil Peddireddy, Allan Dickerman, Andrei Bura, Andrew Warren, Brian D Klahn, Chunhong Mao, Dawen Xie, Dustin Machi, Erin Raymond, Fanchao Meng, Golda Barrow, Henning Mortveit, Jiangzhuo Chen, Jim Walke, Joshua Goldstein, Mandy L Wilson, Mark Orr, Przemyslaw Porebski, Pyrros A Telionis, Richard Beckman, Stefan Hoops, Stephen Eubank, Young Yun Baek, Bryan Lewis, Madhav Marathe, and Chris Barrett. Evaluating the impact of international airline suspensions on the early global spread of COVID-19. *medRxiv*, 2020.

[Adiga *et al.*, 2020b] Aniruddha Adiga, Lijing Wang, Adam Sadilek, Ashish Tendulkar, Srinivasan Venkatramanan, Anil Vullikanti, Gaurav Aggarwal, Alok Talekar, Xue Ben, Jiangzhuo Chen, et al. Interplay of global multi-scale human mobility, social distancing, government interventions, and covid-19 dynamics. *medRxiv*, 2020.

[Anastassopoulou *et al.*, 2020] Cleo Anastassopoulou, Lucia Russo, Athanasios Tsakris, and Constantinos Siettos. Data-based analysis, modelling and forecasting of the covid-19 outbreak. *PloS one*, 15(3):e0230405, 2020.

[Chinazzi *et al.*, 2020] Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489):395–400, 2020.

[Deng *et al.*, 2019] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Graph message passing with cross-location attentions for long-term ili prediction. *arXiv preprint arXiv:1912.10202*, 2019.

[Gilmer *et al.*, 2017] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*, 2017.

[Kai *et al.*, 2020] De Kai, Guy-Philippe Goldstein, Alexey Morgunov, Vishal Nangalia, and Anna Rotkirch. Universal masking is urgent in the covid-19 pandemic: Seir and agent based models, empirical validation, policy recommendations. *arXiv preprint arXiv:2004.13553*, 2020.

[Kraemer *et al.*, 2020] Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[Ramchandani *et al.*, 2020] Ankit Ramchandani, Chao Fan, and Ali Mostafavi. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. *arXiv preprint arXiv:2008.00115*, 2020.

[Ribeiro *et al.*, 2020] Matheus Henrique Dal Molin Ribeiro, Ramon Gomes da Silva, Viviana Cocco Mariani, and Leandro dos Santos Coelho. Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil. *Chaos, Solitons & Fractals*, page 109853, 2020.

[UVA, 2020 accessed August 29 2020] UVA. *UVA COVID-19 Surveillance Dashboard*, 2020 (accessed August 29, 2020).

[Wu *et al.*, 2018] Yuexin Wu, Yiming Yang, Hiroshi Nishiura, and Masaya Saitoh. Deep learning for epidemiological predictions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1085–1088. ACM, 2018.

[Yang *et al.*, 2020] Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook-San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, et al. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of Thoracic Disease*, 12(3):165, 2020.