COVINet: A deep learning-based and interpretable prediction model for the county-wise trajectories of COVID-19 in the United States

Ting Tian*, Yukang Jiang*, Yuting Zhang*, Zhongfei Li§, Xueqin Wang†, and Heping Zhang‡
*School of Mathematics, Sun Yat-sen University, Guangzhou, China †School of Management, University of Science and Technology of China, Hefei, Anhui, China ‡School of Public Health, Yale University, New Haven, CT, USA §Business School, Southern University of Science and Technology, Shenzhen, Guangdong, China

Abstract—The cases of COVID-19 have been reported in the United States since January 2020. We propose a COVINet by combining the architecture of both Long Short-Term Memory and Gated Recurrent Unit. First, we use the 10-fold cross-validation to train and assess different prediction models for which all counties serve alternatively as the training and test counties. Then, we focus on the prediction for the 10 severest counties. We employ the Mean Relative Errors (MREs) to measure the performance of the COVINet in predicting confirmed cases and deaths. Two COVINet models with 26 and 19 input variables, respectively, are trained. We estimate their respective MREs in the last 30 days before January 23, 2021, by the 10-fold CV, which are 0.0898 and 0.1068 for the number of confirmed cases, and 0.0694 and 0.0724 for the number of deaths. The MREs are also small for all predictions of the events in the last 7 or 30 days before January 23, 2021. The COVINet uses features including workforce driving alone to work, traffic volume, income inequality, and longitude and latitude of infected counties to predict the trajectories of COVID-19 in counties of the United States. The increasing awareness of how predictors affect the pandemic helps policymakers develop plans to mitigate the spread of COVID-19.

Index Terms—COVINet model, Deep learning, Geographical signals, Income inequality, Traffic volume, Transportation choices

1 Introduction

According to the New York Times [1], the early confirmed cases of COVID-19 were reported on January 21, 2020, in the United States. In March [2], the outbreak of COVID-19 was proclaimed as a "pandemic" by the World Health Organization. Since then, the United States has had the largest number of confirmed cases and deaths globally [3], where the confirmed cases and deaths were 25,047,893 and 417,390, respectively, as of January 23, 2021. A vast majority of states in the United States issued a "stay at home" order to reduce the transmission of COVID-19 since March 2020 [4]. As the states are reopening to achieve normalcy, it is essential to predict the trajectories of COVID-19 based on the actionable factors to provide the decision-makers with a quantitative and dynamic assessment. Here, we define the actionable factors as those that may be routinely surveilled and collected by the local and national authorities, such as the level of air pollution [5]. Among them, environmental factors affect the spread of infectious diseases. For instance, the hospitalization rate of H1N1 2009 had a disproportionate impact on high-poverty areas in New York City [6] and on the small population of racial/ethnic groups in Wisconsin [7]. Consequently, we consider county health ranking and roadmaps programs [8]. The details about the database are available from "https://www.countyhealthrankings.org/reports/county-health-rankings-reports." We focus on the adverse health factors related to physical and social environments, as summarized in Table 1.

There are many studies dedicated to forecasting the spread of COVID-19. The epidemic models are prevalent tools to predict the infection trajectories [9], [10], [11]. Instead of relying on disease resumption, some authors proposed neural networks to precisely estimate the epidemic [12], [13]. These data-driven approaches had superior performance in predicting the dynamics of COVID-19. Yang et al. [13] proposed a Long Short-Term Memory (LSTM) [14] based model, and Bandyopadhyay and Dutta [15] compared three models, including LSTM, Gated Recurrent Unit (GRU) [16], and LSTM combined with GRU in predicting COVID-19. The LSTM combined with GRU had been proven to generate a high accuracy rate [15]. However, a deep learning-based model is generally complex and not useful in making informed decisions. Therefore, our primary goal is to build deep learning models that can help decision-making for the epidemic. We include historical epidemic data and adverse health factors from all infected counties in the United States to build a county-wise prediction model combining LSTM and GRU. Given the weights of the adverse health factors in our proposed model, actionable interventions could be implemented to slow the epidemic's spread.

- *T. Tian, Y. Jiang, and Y. Zhang contributed equally to this article.*

TABLE 1
The list of 10 health adverse risks related to the physical and social environment and their sources.

Category	Factors	Meanings	Sources
Physical	Driving alone to work	Percentage of the workforce that drives alone to work	American Community Survey, 5-year estimates
	Traffic volume	Average traffic volume per meter of major roadways in the county	Environmental Justice Screening and Mapping Tool
	Air pollution particulate matter	Average daily density of fine particulate matter in micrograms per cubic meter ($PM_{2.5}$)	Environmental Public Health Tracking Network
	Severe housing problems	Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, lack of kitchen facilities, or lack of plumbing facilities	Comprehensive Housing Affordability Strategy (CHAS) data
	Long commute driving alone	Among works who commute in their car alone, the percentage that commute more than 30 minutes	American Community Survey, 5-year estimates
	Homeownership	Percentage of occupied housing units that are owned	American Community Survey, 5-year estimates
Social	Income inequality	Ratio of household income at the 80^{th} percentile to income at the 20^{th} percentile	American Community Survey, 5-year estimates
	Some college	Percentage of adults ages 25-44 with some post-secondary education	American Community Survey, 5-year estimates
	Unemployment	Percentage of population ages 16 and older unemployed but seeking work	Bureau of Labor Statistics
	Social associations	Number of membership associations per 10,000 population	County Business Patterns

To evaluate the performance of the proposed model, we compare our model with four competing models: LSTM, GRU, the method proposed by Yang et al. [13], and random forest [17]. Moreover, we randomly split all counties into training counties and test counties in the 10-fold cross-validation (CV) to evaluate the performance of our model, where the ratio of the training counties to the test counties is 9:1. Finally, after predicting the COVID-19 pandemic for all counties, we present our predictive model for the 10 severest counties in light of their greatest public health interest. Our work is to obtain accurate predictions in the projected trajectories of COVID-19 in the hot-spot areas and directly provide measurable and actionable responses to reduce the spread of COVID-19.

2 METHODS

2.1 Data Sources

We collect the daily numbers of cumulative confirmed cases and deaths from January 21, 2020, to January 23, 2021, for infected counties in the United States from the New York Times [1]. The daily cumulative confirmed cases and deaths are collected from health departments and U.S. Centers for Disease Control and Prevention (CDC), where patients are identified as "confirmed" based on the positive laboratory tests and clinical symptoms and exposure [1]. All risk factors are compiled from 2020 annual data on the County Health Rankings and Roadmaps program's official website [8]. In addition, the longitude and latitude of each infected county are collected from Census TIGER 2000 [18]. Data analysis is conducted in version 3.7 Python with TensorFlow-GPU 1.140 and Keras 2.3.0.

2.2 The selection of the features

The input data are divided into two parts. The first part consists of the cumulative confirmed cases and deaths in the past 7 days:

$$\mathbf{X}_{\cdot,k}^{(\text{main})} = \begin{pmatrix} x_{1,k}^{(\text{cases})} & \cdots & x_{7,k}^{(\text{cases})} & x_{1,k}^{(\text{deaths})} & \cdots & x_{7,k}^{\text{deaths})} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ x_{N-7,k}^{(\text{cases})} & \cdots & x_{N-1,k}^{(\text{cases})} & x_{N-7,k}^{(\text{deaths})} & \cdots & x_{N-1,k}^{(\text{deaths})} \end{pmatrix}_{T\times 14},$$

$$\mathbf{Y}_k = \begin{pmatrix} x_{8,k}^{(\text{cases})} & x_{8,k}^{(\text{deaths})} \\ \vdots & \vdots \\ x_{N,k}^{(\text{cases})} & x_{N,k}^{(\text{deaths})} \end{pmatrix}_{T\times 2}, k = 1, 2, \ldots, K,$$

where N is the length of the training period, $T = N - 7$, and K is the total number of counties. $x_{i,k}^{(\text{cases})}$ are the cumulative confirmed cases and $x_{i,k}^{(\text{deaths})}$ are the total deaths at the corresponding date. For example, $i = 1$ corresponds to the first day when the confirmed cases were officially reported. These cumulative confirmed cases and total deaths give rise to 14 historical epidemic variables as the first part of the input data. The other part of the inputs includes J county features, $\mathbf{X}_k^{(\text{cov})} = [x_{1k}^{(\text{cov})}, ..., x_{Jk}^{(\text{cov})}]^T$. These features are 10 actionable factors (Table 1) in addition to the longitude and latitude of infected counties. Thus, J is 12 for the second part of our input data. Although the longitude and latitude of infected counties are not actionable variables, we incorporated them in our model because of their established importance in prediction [19], [20]. Altogether, as presented in Figure 1, the 26 input features are included in constructing the initial proposed model. Note that the input data are not predicted from the model.

Our next step is to find out whether there is a parsimonious model with a smaller number of features and comparable accuracy to the initial proposed model. To this end, we used the random forest to screen the 10 actionable features. In a random forest, a common practice is to select the features with the largest variances [17]. This approach selects the following three features: percentage of the workforce who drive alone to work, average traffic volume per meter of major roadways, and income inequality (the ratio

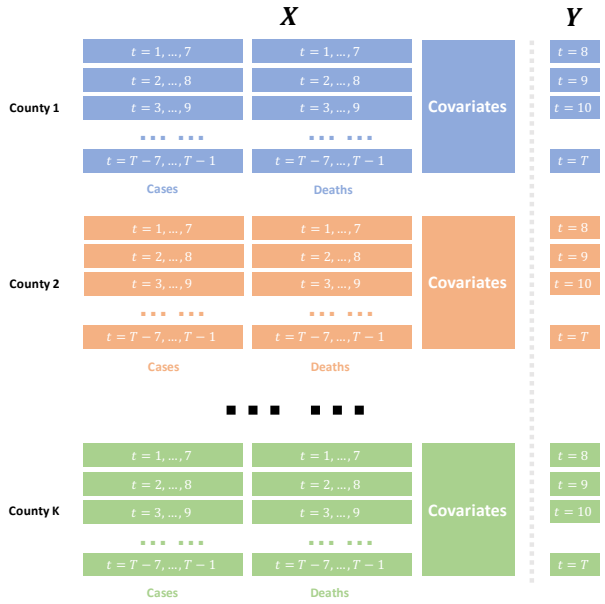

Fig. 1. The structure of data usage in the models.

of household income at the 80th percentile to that at the 20th percentile). Therefore, our smaller proposed model uses 19 features as the input data: 14 historical epidemic variables and 5 county features (3 selected actionable features, longitude, and latitude). To compare the performance of all competing models, these same 19 features are used.

2.3 COVINet

2.3.1 Model architecture

Our proposed model incorporates an LSTM layer, a GRU layer [14], [15], [16], and a fully connected layer, formulated as:

$$f\left(\mathbf{X}_{..k}^{(\text{main})}, \mathbf{X}_{k}^{(\text{cov})}\right) = g^{(\text{dense})}\Big(g^{(\text{LSTM})}\left(\mathbf{X}_{..k}^{(\text{main})}\right), g^{(\text{GRU})}\left(\mathbf{X}_{..k}^{(\text{main})}\right), \mathbf{X}_{k}^{(\text{cov})}\Big), \tag{1}$$

where $g^{(\text{dense})}$ is a fully connected layer, $g^{(\text{LSTM})}$ is an LSTM layer, and $g^{(\text{GRU})}$ is a GRU layer. The time series of historical epidemic data $\mathbf{X}_{..k}^{(\text{main})}$ are the inputs of LSTM and GRU layers, which are typically used in time series analysis for the deep learning process. We then concatenate the outputs of these two layers and the time-invariant county features $\mathbf{X}_{k}^{(\text{cov})}$ in a fully connected layer.

An LSTM layer ($g^{(\text{LSTM})}$) contains the input gate in_t, the forget gate f_t, the output gate o_t , the cell state c_t (i.e., the hidden status), the candidate value $\tilde{c}_t$, and the hidden state vector/final output h_t. $\mathbf{X}_{t.k}^{(\text{main})}$ is a t^{th} row of $\mathbf{X}_{..k}^{(\text{main})}$ used as the input vector of the LSTM layer, then the iterative formula for each item is shown as follows:

$$\begin{aligned}
in_t &= \sigma\left(W_i \mathbf{X}_{t.k}^{(\text{main})} + U_i h_{t-1}^{(\text{LSTM})} + b_i\right)\\
f_t &= \sigma\left(W_f \mathbf{X}_{t.k}^{(\text{main})} + U_f h_{t-1}^{(\text{LSTM})} + b_f\right)\\
o_t &= \sigma\left(W_o \mathbf{X}_{t.k}^{(\text{main})} + U_o h_{t-1}^{(\text{LSTM})} + b_o\right)\\
\tilde{C}_t &= \tanh\left(W_c \mathbf{X}_{t.k}^{(\text{main})} + U_c h_{t-1}^{(\text{LSTM})} + b_c\right)\\
C_t &= f_t \otimes C_{t-1} \oplus int \otimes \tilde{C}_t\\
h_t^{(\text{LSTM})} &= o_t \otimes \tanh\left(C_t\right)
\end{aligned}$$

Comparatively, a GRU layer ($g^{(\text{GRU})}$) streamlines the operation. The layer removes the cell state C_t, the information transmits in the hidden state (h_t), input gate in_t and forget gate f_t emerge to form an updated gate z_t, a reset gate r_t adds, and removes the final output gate. Thus, the corresponding update functions are:

$$\begin{aligned}
r_t &= \sigma\left(W_r \mathbf{X}_{t.k}^{(\text{main})} + U_r h_{t-1}^{(\text{GRU})} + b_r\right)\\
z_t &= \sigma\left(W_z \mathbf{X}_{t.k}^{(\text{main})} + U_z h_{t-1}^{(\text{GRU})} + b_z\right)\\
\tilde{h}_t &= \tanh\left(W_h \mathbf{X}_{t.k}^{(\text{main})} + U_h\left(r_t \otimes h_{t-1}^{(\text{GRU})}\right) + b_h\right)\\
h_t^{(\text{GRU})} &= (1 - z_t) \otimes h_{t-1}^{(\text{GRU})} \oplus z_t \otimes \tilde{h}_t
\end{aligned}$$

where matrices W_i,W_f,W_o,W_c,W_z,W_r,W_h,U_i,U_f,U_o,U_c,U_r, U_h,U_h and vectors b_i,b_f,b_o,b_c,b_z,b_r,b_h are model parameters. σ is a sigmoid function,$\otimes$ and $\oplus$ are pointwise multiplication, pointwise addition, respectively.

For a fully connected layer ($g^{(\text{dense})}$), we apply a dropout step to limit the dimensions of the outputs, referred to as nodes in the deep learning literature, generated from LSTM and GRU layers and prevent overfitting. The outputs are dropped randomly at a rate to be specified by the users, which we discuss in Section 2.3.3. The number of nodes and the dropout rates for LSTM and GRU layers are tuned as the hyperparameters in the network configurations. The activation function of the fully connected layer is set as the ReLU function to generate the non-negative cumulative confirmed cases and total deaths. Our proposed model, referred to as COVINet, conducts the deep learning process by incorporating county features. The corresponding COVINet is shown in Figure 2.

All data involved in the model are min-max normalized before being used. This step is found to increase the accuracy of our model and training speed. For unknown data containing the same variables, we use the scales from the training data to transform future epidemic data and then predict the future COVID-19. After obtaining the predicted data, we proportionally restore the predicted cumulative confirmed cases and deaths by reversing the scales.

2.3.2 Training

During the training process, the observed cumulative confirmed cases and deaths every 7 past days in each county of the United States are used to predict the cumulative confirmed cases and deaths in the present day. Our analysis is divided into two parts. The first part is to learn the observed patterns of COVID-19 and then to validate the learned

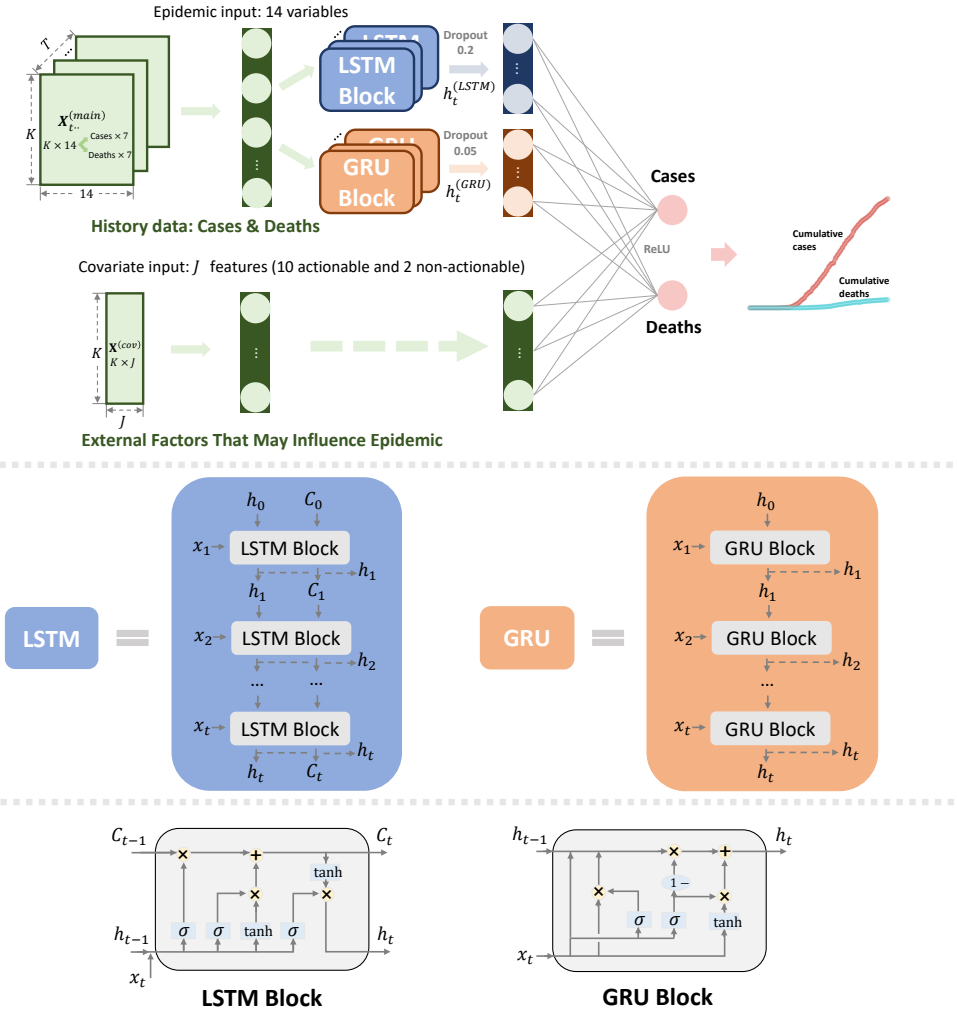

Fig. 2. The COVINet combining Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) using J (12) county features.

patterns, where the accuracy of the models is evaluated by Mean Relative Errors (MRE_t) as validation loss:

$$\text{MRE}_t = \frac{1}{t}\sum_{i=1}^{t}\frac{|\text{ Actual}_i - \text{Predicted}_i\ |}{\text{Actual}_i}, t = 7, 30,$$

where Actual_i are the actual cumulative confirmed cases or total deaths at the i^{th} day and Predicted_i are the predicted ones at the same corresponding date. The weights of an entire network are estimated by backpropagation through minimizing the loss function (MRE).

We assess the performance of all models through both spatial and temporal domains. First, we use a 10-fold CV to evaluate the prediction accuracy from one set of training counties to a different set of test counties in an approximately 9:1 ratio. The trajectories of COVID-19 before January 23, 2021, are used in the 10-fold CV, although we report only the MRE_{30} for the last 30 days before January 23, 2021. Then, we evaluate the prediction accuracy in the last 7 or 30 days (MRE_7 for the last 7 days and MRE_{30} for the last 30 days before January 23, 2021) for the 10 severest counties. The following are two settings of the data: (a) $t = 7$: the training data were from January 21, 2020, to January 16, 2021, and the validation data from January 17 to January 23, 2021. Here, $i = 1$ corresponded to January 17, 2021; (b) $t = 30$: the training data were from January 21 to December 23, 2020, and the validation data from December 24, 2020 to January 23, 2021, $i = 1$ corresponded to December 24, 2020;

2.3.3 *Tuning the hyperparameters*

While building models by LSTM and GRU, we need to tune two hyperparameters to achieve high accuracy. The first one is the number of nodes in LSTM and GRU. We consider 50, 100, and 150 as commonly done [15]. The second one is the dropout rates. We set the range from 0 to 50% with an increment of 5%. The choices of these tuning hyperparameters with the lowest MRE are selected. Specifically, 50 nodes are used for each network in both LSTM and GRU, and the dropout rates are set at 20% and 5% for LSTM and GRU, respectively. We use the Adam optimizer for model training, and following Kingma and Ba [21], we set $\alpha = 0.001$ (step size or learning rate), $\beta_1 = 0.9$, $\beta_2 = 0.999$ (exponential decay rates for the moment estimates), and $\varepsilon = 10^{-7}$ for the Adam optimizer. The batch size, i.e., the number of training samples for each iteration, is set as 32. The COVINet model is trained up to 100 epochs. For the learning rate, if the MRE does not decrease for consecutive 10 epochs, we reduce the learning rate to its 30% until MRE decreases

or the minimum learning rate reaches 0.00001. The training process is stopped if the MRE does not improve over 40 consecutive epochs.

3 RESULTS

3.1 Model validation through different counties

We conduct the 10-fold CV of all counties to evaluate the performances of our proposed models. The averages of the MREs for the COVINet with the initial 26 features and the COVINet with the 19 features (3 of the 10 actionable features) are 0.0898 and 0.1068 for the cumulative confirmed cases, and 0.0694 and 0.0724 for the total deaths, respectively. The corresponding MREs for LSTM alone, GRU alone, the method proposed by Yang et al. are over 0.45 for the cumulative confirmed cases and deaths. The random forest is better than the two COVINet models for the validated cumulative confirmed cases but worse for the validated total deaths, although they are in a similar order of magnitudes. Overall, the 10-fold CV-based MREs in all counties for two COVINet models are either much better than or comparable to those using the existing models.

Table 2 presents the MREs from these 6 models for the subsequent (and last) 7 or 30 days for the cumulative confirmed cases and total deaths in the severest counties. For the COVINet model with the initial 26 features, the MRE_{30} are 0.0261 for the cumulative confirmed cases and 0.0179 for the total deaths. The MRE_{30} for the cumulative cases and total deaths are 0.0210 and 0.0386 for the smaller COVINet model with 19 features. For the 4 competing models, the MRE_{30} for LSTM alone, GRU alone, the method proposed by Yang et al. and random forest for the cumulative confirmed cases are 0.0844, close to 1, 0.0217, and 0.1574, and for the total deaths are 0.1683, 1, 0.0746 and 0.1578, respectively. Thus, the COVINet models have the lowest values of MRE_{30} for the cumulative confirmed cases and total deaths. The conclusion is similar for MRE_7 . Finally, to assess the importance of the 19 features that are included in the smaller COVINet model, we build a COVINet without them. The resulting MREs are close to 1.

3.2 Prediction of future trajectories of COVID-19 in the 10 severest counties

The MRE_{30} and MRE_7 between the observed and projected counts from the day after two training periods to January 23, 2021, are computed to assess the accuracy of the temporal prediction for each of the 10 most severely infected counties, because those hot-hit areas were of the severest public health interest. Table 3 presents individual MRE_{30} and MRE_7 for those 10 counties using the smaller COVINet with 19 features. For example, Orange County, California, has the smallest MRE_{30} for the confirmed cases and the smallest MRE_7 for total deaths. Dallas County, Texas, has the smallest MRE_{30} for the total deaths. Riverside County, California, has the smallest MRE_7 for the confirmed cases. Overall, the MRE_{30} and MRE_7 are relatively small, assuring the accuracy of our COVINet model in predicting future trajectories of COVID-19 for the numbers of confirmed cases and deaths for the severest counties.

The 30-day projected trajectories of the cumulative confirmed cases and deaths using the smaller COVINet from December 24, 2020, to January 23, 2021, are presented in Figure 3. From Figure 3, the predicted cumulative confirmed cases from December 24, 2020, to January 23, 2021, are remarkably close to the actual ones for the 10 severest counties. The situation is similar in predicting the death counts except for San Bernardino County, California, for which the predicted deaths are relatively higher than the actual ones, with MRE_{30} being 0.0861 (Table 3). The projected values of the confirmed cases for the 10 severest counties would increase at a slow rate in the near future.

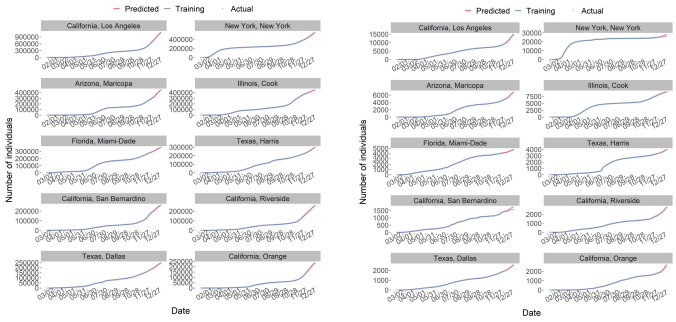

Fig. 3. The trajectories of COVID-19 for 10 severest counties until January 23, 2021, are displayed. The blue curves indicate the observed cumulative confirmed cases (a) and total deaths (b) as training data, while the red curves indicate the projected ones from December 24, 2020 to January 23, 2021. The purple curves represent the actual confirmed cases (a) and deaths (b) at the same period.

3.3 Covariate effect on COVID-19

Our smaller COVINet model incorporates 3 selected adverse health factors, the longitudes and latitudes of the counties. The weights of longitudes and latitudes are learned from the training county data, where their values are 0.0001 and 0.0002 for the confirmed cases and 0.0002 and 0.0002 for the total deaths, respectively. Accordingly, the Northern and Eastern regions have relatively more confirmed cases, and thus there are more deaths in the same regions. The maps of the cumulative confirmed cases and total deaths of COVID-19 on January 23, 2021, are presented in Figure 4 and are consistent with our prediction. There are more infected counties in the Northern and Eastern regions.

The weights of the 3 selected adverse health risk factors are positive for both confirmed cases and deaths. For example, the largest values of weights for confirmed cases are the percentage of workforce driving alone to work at 0.0005 and for deaths are the average traffic volume per meter of major roadways at 0.0004 (Table 4). Specifically, an increase in the percentage of workforce driving alone to work, average traffic volume per meter of major roadways, and income inequality ratio would increase in both the cumulative confirmed cases and deaths.

To offer insight into the prediction dynamics of the smaller COVINet model, we vary the levels of the 3 actional features and present the resulting trajectories of COVID-19 for Los Angeles County, California. Moreover, for better visibility, we draw the projected trajectories of COVID-19 from January 24, 2021, to February 2, 2021, where the observed confirmed cases and total deaths on January 24, 2021, serve as the reference value set at 0 in Figure 5. For

TABLE 2
The average 10-fold cross-validation (CV) results of the COVINet including the initial 26 features, the COVINet model with 3 selected features, existing competing models for all counties and the values of MRE_{30} and MRE_7 for 10 severest counties of COVID-19.

Models	Average 10-fold CV		MRE_7 projection		MRE_{30} projection	
	Confirmed cases	Deaths	Confirmed cases	Deaths	Confirmed cases	Deaths
The COVINet model with the initial 26 features	0.0898	0.0694	0.0080	0.0203	0.0261	0.0179
The COVINet model with 3 selected features	0.1068	0.0724	0.0056	0.0138	0.0210	0.0386
LSTM alone	0.5129	0.5589	0.0330	0.1141	0.0844	0.1683
GRU alone	0.6345	0.5739	1.0000	1.0000	0.9999	1.0000
Deep learning based method of Yang et.al	0.4608	0.4679	0.0990	0.0626	0.0217	0.0746
Random forest	0.0461	0.1494	0.0447	0.0430	0.1574	0.1578

TABLE 3
MRE_7 and MRE_{30} of cumulative confirmed cases and deaths using COVINet model with 3 selected features for each of 10 severest counties of COVID-19.

State, County	MRE_7		MRE_{30}	
	Confirmed cases	Deaths	Confirmed cases	Deaths
California Los Angeles	0.0043	0.0067	0.0253	0.0067
New York New York	0.0063	0.0249	0.0110	0.0769
Arizona Maricopa	0.0021	0.0039	0.0179	0.0199
Illinois Cook	0.0129	0.0167	0.0337	0.0208
Florida Miami-Dade	0.0070	0.0171	0.0245	0.0415
Texas Harris	0.0026	0.0100	0.0226	0.0359
California San Bernardino	0.0039	0.0269	0.0108	0.0861
California Riverside	0.0060	0.0129	0.0188	0.0204
Texas Dallas	0.0048	0.0039	0.0248	0.0136
California Orange	0.0062	0.0150	0.0206	0.0645

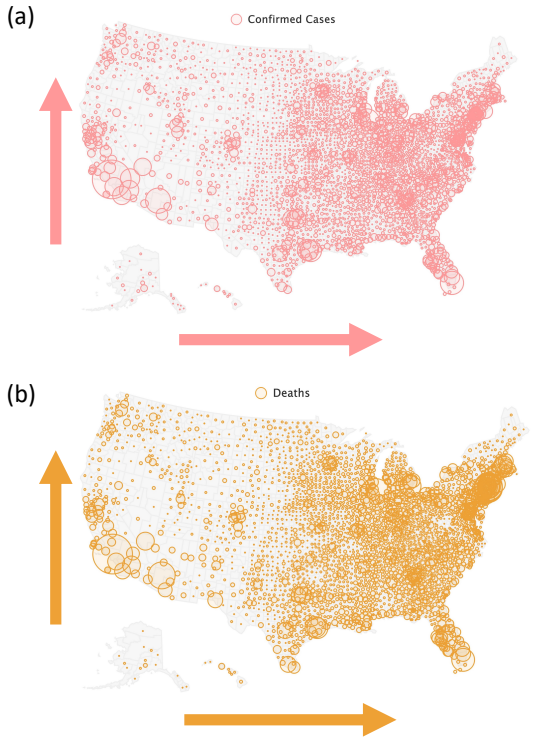

Fig. 4. The map of all infected counties. The circle sizes indicate the number of cumulative confirmed cases (a) and deaths (b) on January 23, 2021. The arrows indicate the trend of change in confirmed cases and deaths over longitudes and latitudes.

example, if the average traffic volume per meter of major

TABLE 4
The weights of five county features

Factors	Weights	
	Confirmed cases	deaths
Driving alone to work	4.5506×10^{-4}	2.0113×10^{-4}
Traffic volume	1.1130×10^{-4}	4.1634×10^{-4}
Income inequality	2.5271×10^{-4}	2.1376×10^{-4}
Longitude	8.9214×10^{-5}	1.8581×10^{-4}
Latitude	1.9427×10^{-4}	2.0398×10^{-4}

roadways increases 4 times on January 24, 2021, the number of cumulative confirmed cases will increase over 7500 on February 2, 2021 (Figure 5). The impact of the 3 actional features on COVID-19 in both the cumulative confirmed cases and deaths is visible, depending on the weights of the features (Table 4). Overall, the numbers of cumulative confirmed cases and deaths are projected to rise slowly in the following month in Los Angeles County, California.

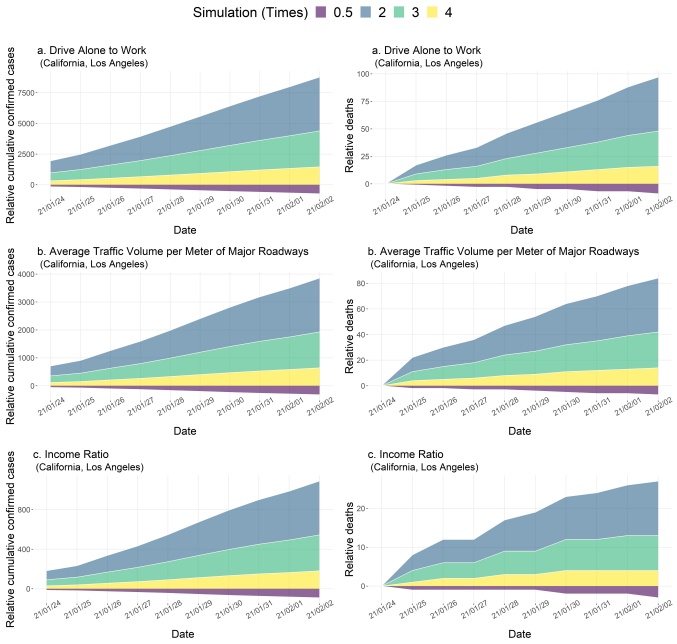

Fig. 5. The projected relative trajectories of COVID-19 for Los Angeles County, California, of cumulative confirmed cases and deaths from January 24 to February 2, 2021. The levels of the 3 risk factors are changed from 0.5 times to 4 times since January 24, 2021. The observed cumulative confirmed cases and deaths are set at 0 on January 24, 2021.

4 Discussion

Our COVINet is built by deep learning and is shown to effectively model, which elegantly predicts the cumulative confirmed cases and deaths in the counties of the United States. The risk factors that are used in the COVINet provide visible evidence on actionable steps that influenced the trajectories of COVID-19. Thus, COVINet takes advantage of deep learning and the interpretability of risk factors. In our study, we find that the higher the workforce driving alone to work, the higher the risk of COVID-19 spread. Those car-only commuters are associated with a higher body fat percentage [22], increasing the risk of infections [23]. Studies indicate that pre-existing cardiovascular disease could increase the severity of the COVID-19 [24], [25], so does the air pollution [26], [27]. The residential proximity to high vehicle traffic at a distance would increase exposure to air pollution and risk of cardiovascular disease (CVD) [28], [29], [30]. Therefore, it seems to be an indirect risk to COVID-19 for people living near heavy traffic. In the end, income inequality within a county accentuates the risk of poor health [31]. Those people have relatively low health status, making them more vulnerable to novel diseases [32] and infected in hospitals. Overall, if the values of those adverse health factors increase, the trajectories of COVID will be increased accordingly. This might be consistent with the fact that those adverse health factors result in poor health and thus have a high likelihood of increasing the trajectories of COVID-19. Therefore, adverse health factors are expected to differ in the COVID-19 trajectories significantly. As a result, of the COVID-19 pandemic, it is a public health matter and an issue of social responsibility.

We also take into account the geographical information of infected regions; there could be a link between geographical signals and COVID-19. Our results indicate that higher latitudes have more cases, consistent with previous studies [19], [20]. As the most severe county in the United States, the Los Angeles County of California is located in the southwest of the United States with the highest number of cases of COVID-19 since 2021. However, for the overall hotspot areas of COVID-19, approaching north (higher values in the latitude) and east (higher values in the longitude) areas of the United States, the more severe counties with higher numbers of cases have been. Accordingly, the same situations apply to the deaths of COVID-19. The majority of severely infected counties are located in the northeast areas of the United States. There might be other factors that we could consider in building the COVINet. However, we chose to use the 3 actionable adverse health factors based on a criterion in the random forest, and they may be controllable by local authorities relatively quickly. Moreover, the smaller COVINet model performed comparably to the initial COVINet model. Thus, for model parsimony, we chose the smaller model as our final model. LSTM combined with GRU was shown to capture more temporal information, consistent with the work proposed by Dutta et al. [15]. The potential structure of the data that can be captured by using GRU or LSTM alone might be relatively simple. We believe each method alone might not effectively capture the information for accurate prediction. By using both network structures, we can have a more prosperous prediction [15]. We also included the longitude and latitude because they help predict COVID-19. For the 10 severest counties, the overall prediction for the cumulative confirmed cases is more accurate than that for the total deaths because there are more observed confirmed cases than deaths, resulting in a larger denominator for the MRE.

Our models produce accurate county-level short-term (7-day) and long-term (30-day) predictions of cumulative confirmed cases and total deaths together. More significantly, they are based on measurements routinely surveilled and collected by the local and national authorities, providing actionable information to reduce the spread of COVID-19. Therefore, it is easy to understand and act by the decision-makers. Also, with its relatively small MREs, the room for further prediction improvement is expected to be too small to make a practical difference. In addition, considering the inputs and outputs in our model, we can have multimodal predictors and multiple outcomes.

5 Conclusion

In summary, we built an interpretable and highly accurate prediction model using deep learning for COVID-19. This developed deep learning model can precisely predict the different periods of cumulative confirmed cases and deaths in infected regions. By incorporating the time-invariant factors in deep learning, the accuracy could improve remarkably to predict the trajectories of COVID-19. By analyzing the spread of COVID-19 and adverse health risk factors related to physical and social environments, we can improve the healthcare system for COVID-19.

Acknowledgments

We would like to thank all individuals who collected epidemiological data of the COVID-19 outbreak and the county health ranking and roadmaps program data. This research was supported by the National Natural Science Foundation of China (No. 71991474; No. 12001554; No. 72171216), the Key Research and Development Program of Guangdong, China (No. 2019B020228001), the Science and Technology Program of Guangzhou, China (No.202002030129), the International Science & Technology cooperation program of Guangdong, China (2016B050502007), the Natural Science Foundation of Guangdong Province, China (No. 2020A1515010617), and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (No. 2021qntd21). The funding agencies had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

References

[1] The New York Times, "Coronavirus in the u.s.: Latest map and case count," https://www.nytimes.com/interactive/2021/us/covid-cases.html, 2021.

[2] WHO Director General, "Who director-general's opening remarks at the media briefing on covid-19 -11 march 2020," https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020, 2020.

[3] National Health Commission of the People's Republic of China, "Distribution of covid-19 cases in the world," http://2019ncov.chinacdc.cn/2019-nCoV/global.html., 2020.

[4] Governor New York State, "Governor cuomo signs the 'new york state on pause' executive order," https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order., 2020.

[5] T. Tian, J. Zhang, L. Hu, Y. Jiang, C. Duan, Z. Li, X. Wang, and H. Zhang, "Risk factors associated with mortality of covid-19 in 3125 counties of the united states," *Infectious diseases of poverty*, vol. 10, no. 1, pp. 1–8, 2021.

[6] S. Balter, L. S. Gupta, S. Lim, J. Fu, S. E. Perlman, N. Y. C. . H. F. I. Team *et al.*, "Pandemic (h1n1) 2009 surveillance for severe illness and response, new york, new york, usa, april–july 2009," *Emerging infectious diseases*, vol. 16, no. 8, p. 1259, 2010.

[7] S. A. Truelove, A. S. Chitnis, R. T. Heffernan, A. E. Karon, T. E. Haupt, and J. P. Davis, "Comparison of patients hospitalized with pandemic 2009 influenza a (h1n1) virus infection during the first two pandemic waves in wisconsin," *Journal of Infectious Diseases*, vol. 203, no. 6, pp. 828–837, 2011.

[8] The County Health Rankings Roadmaps Program, "State ranking data & reports," https://www.countyhealthrankings.org/reports/county-health-rankings-reports., 2020.

[9] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689–697, 2020.

[10] A. Mahajan, N. A. Sivadas, and R. Solanki, "An epidemic model sipherd and its application for prediction of the spread of covid-19 infection in india," *Chaos, Solitons & Fractals*, vol. 140, p. 110156, 2020.

[11] L. Wang, Y. Zhou, J. He, B. Zhu, F. Wang, L. Tang, M. Kleinsasser, D. Barker, M. C. Eisenberg, and P. X. Song, "An epidemiological forecast model and software assessing interventions on the covid-19 epidemic in china," *Journal of Data Science*, vol. 18, no. 3, pp. 409–432, 2020.

[12] Z. Hu, Q. Ge, S. Li, L. Jin, and M. Xiong, "Artificial intelligence forecasting of covid-19 in china," *Working paper*, 2020.

[13] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. 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*, vol. 12, no. 3, p. 165, 2020.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] S. Dutta and S. K. Bandyopadhyay, "Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release," *Iberoamerican journal of medicine*, vol. 2, no. 3, pp. 172–177, 2020.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Working paper*, 2014.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] U.S. Counties, "National weather service," https://www.weather.gov/gis/Counties., 2020.

[19] M. Shokouhi, F. Miralles-Wilhelm, M. A. Amoroso, and M. M. Sajadi, "Temperature, humidity, and latitude analysis to predict potential spread and seasonality for covid-19.","," *Working paper*, 2020.

[20] M. Sarmadi, N. Marufi, and V. K. Moghaddam, "Association of covid-19 global distribution and environmental and demographic factors: An updated three-month study," *Environmental Research*, vol. 188, p. 109748, 2020.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Working paper*, 2014.

[22] E. Flint and S. Cummins, "Active commuting and obesity in mid-life: cross-sectional, observational evidence from uk biobank," *The lancet Diabetes & endocrinology*, vol. 4, no. 5, pp. 420–435, 2016.

[23] F. Gao, K. I. Zheng, X.-B. Wang, Q.-F. Sun, K.-H. Pan, T.-Y. Wang, Y.-P. Chen, G. Targher, C. D. Byrne, J. George *et al.*, "Obesity is a risk factor for greater covid-19 severity," *Diabetes care*, vol. 43, no. 7, pp. e72–e74, 2020.

[24] E. Driggin, M. V. Madhavan, B. Bikdeli, T. Chuich, J. Laracy, G. Biondi-Zoccai, T. S. Brown, C. Der Nigoghossian, D. A. Zidar, J. Haythe *et al.*, "Cardiovascular considerations for patients, health care workers, and health systems during the covid-19 pandemic," *Journal of the American College of Cardiology*, vol. 75, no. 18, pp. 2352–2371, 2020.

[25] Y.-Y. Zheng, Y.-T. Ma, J.-Y. Zhang, and X. Xie, "Covid-19 and the cardiovascular system," *Nature reviews cardiology*, vol. 17, no. 5, pp. 259–260, 2020.

[26] D. Contini and F. Costabile, "Does air pollution influence covid-19 outbreaks?" p. 377, 2020.

[27] E. Conticini, B. Frediani, and D. Caro, "Can atmospheric pollution be considered a co-factor in extremely high level of sars-cov-2 lethality in northern italy?" *Environmental pollution*, vol. 261, p. 114465, 2020.

[28] R. Beelen, G. Hoek, P. A. van Den Brandt, R. A. Goldbohm, P. Fischer, L. J. Schouten, M. Jerrett, E. Hughes, B. Armstrong, and B. Brunekreef, "Long-term effects of traffic-related air pollution on mortality in a dutch cohort (nlcs-air study)," *Environmental health perspectives*, vol. 116, no. 2, pp. 196–202, 2008.

[29] L. M. Baumann, C. L. Robinson, J. M. Combe, A. Gomez, K. Romero, R. H. Gilman, L. Cabrera, N. N. Hansel, R. A. Wise, P. N. Breysse *et al.*, "Effects of distance from a heavily transited avenue on asthma and atopy in a periurban shantytown in lima, peru," *Journal of Allergy and Clinical Immunology*, vol. 127, no. 4, pp. 875–882, 2011.

[30] B. Brunekreef, R. Beelen, G. Hoek, L. Schouten, S. Bausch-Goldbohm, P. Fischer, B. Armstrong, E. Hughes, M. Jerrett, and P. van den Brandt, "Effects of long-term exposure to traffic-related air pollution on respiratory and cardiovascular mortality in the netherlands: the nlcs-air study." *Research report (Health Effects Institute)*, no. 139, pp. 5–71, 2009.

[31] K. E. Pickett and R. G. Wilkinson, "Income inequality and health: a causal review," *Social science & medicine*, vol. 128, pp. 316–326, 2015.

[32] R. Hatch, D. Young, V. Barber, J. Griffiths, D. A. Harrison, and P. Watkinson, "Anxiety, depression and post traumatic stress disorder after critical illness: a uk-wide prospective cohort study," *Critical care*, vol. 22, no. 1, pp. 1–13, 2018.