medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

ARTICLE TEMPLATE

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

Yukang Jiang[a], Ting Tian[a], Wenting Zhou[a], Yuting Zhang[a], Zhongfei Li[d], Xueqin Wang[b,*], Heping Zhang[c,*]

[a]School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong, China; [b]School of Management, University of Science and Technology of China, Hefei, Anhui, China; [c]School of Public Health, Yale University, New Haven, CT, USA; [d]Business School, Southern University of Science and Technology, Shenzhen, Guangdong, China

ARTICLE HISTORY
Compiled December 18, 2023

ABSTRACT
The cases of COVID-19 have been reported in the United States since January 2020. There were over 103 million confirmed cases and over one million deaths as of March 23, 2023. We propose a COVINet by combining the architecture of both Long Short-Term Memory and Gated Recurrent Unit and incorporating actionable covariates to offer high-accuracy prediction and explainable response. First, we train COVINet models for confirmed cases and total deaths with five input features, compare their Mean Absolute Errors (MAEs) and Mean Relative Errors (MREs) and benchmark COVINet against ten competing models from the United States CDC in the last four weeks before April 26, 2021. The results show that COVINet outperforms all competing models for MAEs and MREs when predicting total deaths. Then, we focus on the prediction for the most severe county in each of the top 10 hot-spot states using COVINet. The MREs are small for all predictions made in the last 7 or 30 days before March 23, 2023. Beyond predictive accuracy, COVINet offers high interpretability, enhancing the understanding of pandemic dynamics. This dual capability positions COVINet as a powerful tool for informing effective strategies in pandemic prevention and governmental decision-making.

KEYWORDS
COVINet, Interpretable deep learning, Geographical signals, Air pollution, Traffic volume, Severe housing problems

1. Introduction

According to the New York Times [33], the early confirmed cases of COVID-19 were reported on January 21, 2020, in the United States. In March [40], 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 [24], where the confirmed cases and deaths were 103,910,087 and 1,135,344, respectively, as of March 23, 2023.

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 [18]. As the states are reopening

Y. Jiang, T. Tian, W. Zhou, and Y. Zhang contributed equally to this article.
* Corresponding author

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

to achieve normalcy, it is essential to predict the trajectories of COVID-19 based on 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 [34]. 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 [4] and on the small population of racial/ethnic groups in Wisconsin [36]. Consequently, we consider county health ranking and roadmaps programs [32]. The details about the database are available from `https://www.countyhealthrankings.org/reports/county-health-rankings-reports`. We focus on health factors related to physical and social environments as well as demographics, which are selected based on variable importance ranking of the random forest, 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 [23, 38, 41]. For example, the United States (US) COVID-19 Forecast Hub[14] is a data repository that collects and aggregates the predictions of various epidemic models for the US COVID data. Instead of relying on disease resumption, some authors proposed neural networks to precisely estimate the epidemic [20, 43]. These data-driven approaches had superior performance in predicting the dynamics of COVID-19. Yang et al. [43] proposed a Long Short-Term Memory (LSTM) [19] based model, and Bandyopadhyay and Dutta [5] compared three models, including LSTM, Gated Recurrent Unit (GRU) [11], and LSTM combined with GRU in predicting COVID-19. The LSTM combined with GRU had been proven to generate a high accuracy rate [8]. 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 propose COVINet, a model that utilizes LSTM and GRU networks to forecast disease dynamics at the county level. By incorporating three actionable features reflecting community health risk, as well as longitude and latitude data for each county, COVINet captures local impacts of the disease, identifies high-risk and low-risk factors, and provides valuable and actionable information for public health. To evaluate the performance of COVINet, we align our county-level results with the state-level predictions of ten competing models from the US Centers for Disease Control and Prevention (CDC) that used state-level data. Specifically, we aggregate our county-level results to match their scale for comparison. Additionally, after the prediction of the COVID-19 pandemic for all counties, we showcase our predictive model for the most severely affected county in each of the top 10 states with the highest number of confirmed cases, considering their paramount public health significance. Thus, COVINet's interpretability sheds light on the "black box" of deep learning, providing a clear understanding of how actionable features impact the trajectory of the COVID-19 pandemic. 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1.: The list of five health factors related to their categories, meanings, and sources. The ranks of factors are the variable importance rankings of the random forest models for cumulative confirmed cases and deaths, respectively.

Category	**Factors (Ranks)**	**Meanings**	**Sources**
Physical Environment	Traffic volume (3, 3)	Average traffic volume per meter of major roadways in the county	Environmental Justice Screening and Mapping Tool
	Severe housing problems (4, 4)	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
	Air pollution (6, 5)	Average daily density of fine particulate matter in micrograms per cubic meter ($PM_{2.5}$)	Environmental Public Health Tracking Network
Social & Economic Factors	Some college (2, 1)	Percentage of adults ages 25-44 with some post-secondary education	American Community Survey, 5-year estimates
Demographics	Population (1, 2)	Resident population	Census Population Estimates

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

2. Methods

2.1. Data Sources

We collect the daily numbers of cumulative confirmed cases and deaths from January 21, 2020, to March 23, 2023, for infected counties in the US from the New York Times [33]. The daily cumulative confirmed cases and deaths are collected from health departments and the US CDC, where patients are identified as "confirmed" based on positive laboratory tests and clinical symptoms and exposure [33]. All risk factors are compiled from 2020 annual data on the County Health Rankings and Roadmaps program's official website [32]. In addition, the longitude and latitude of each infected county are collected from Census TIGER 2000 [25]. Data analysis is conducted in Python 3.7 with TensorFlow-GPU 1.14.0 and Keras 2.3.0.

2.2. The selection of features

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

$$\mathbf{X}_{..k}^{(cases)} = \begin{pmatrix} x_{1,k}^{(cases)} & \cdots & x_{14,k}^{(cases)} \\ \vdots & \cdots & \vdots \\ x_{T-20,k}^{(cases)} & \cdots & x_{T-7,k}^{(cases)} \end{pmatrix}_{(T-20)\times 14}, \mathbf{X}_{..k}^{(deaths)} = \begin{pmatrix} x_{1,k}^{(deaths)} & \cdots & x_{14,k}^{(deaths)} \\ \vdots & \cdots & \vdots \\ x_{T-20,k}^{(deaths)} & \cdots & x_{T-7,k}^{(deaths)} \end{pmatrix}_{(T-20)\times 14},$$

$$\mathbf{Y}_{k}^{(cases)} = \begin{pmatrix} x_{21,k}^{(cases)} \\ \vdots \\ x_{T,k}^{(cases)} \end{pmatrix}_{(T-20)\times 1}, \mathbf{Y}_{k}^{(deaths)} = \begin{pmatrix} x_{21,k}^{(deaths)} \\ \vdots \\ x_{T,k}^{(deaths)} \end{pmatrix}_{(T-20)\times 1}, k = 1,\ 2, \ldots, K,$$

where T is the length of the training period, and K is the total number of counties. $x_{i,k}^{(cases)}$ are the cumulative confirmed cases and $x_{i,k}^{(deaths)}$ are the total deaths at the corresponding date. For example, $i = 1$ corresponds to the first day when the confirmed cases and deaths were officially reported. These cumulative confirmed cases and total deaths give rise to fourteen historical epidemic features as the first part of the input data. The other part of the inputs includes J county features, $\mathbf{X}_{k}^{(cov)} = \left[x_{1k}^{(cov)}, \ldots, x_{Jk}^{(cov)}\right]^T$. These features are three actionable factors in addition to the longitude and latitude of infected counties. Thus, $J = 5$ applies to the second part of our input data. Although the longitude and latitude of infected counties are not actionable features, we include them in our model because of their established importance in prediction [29, 31].

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

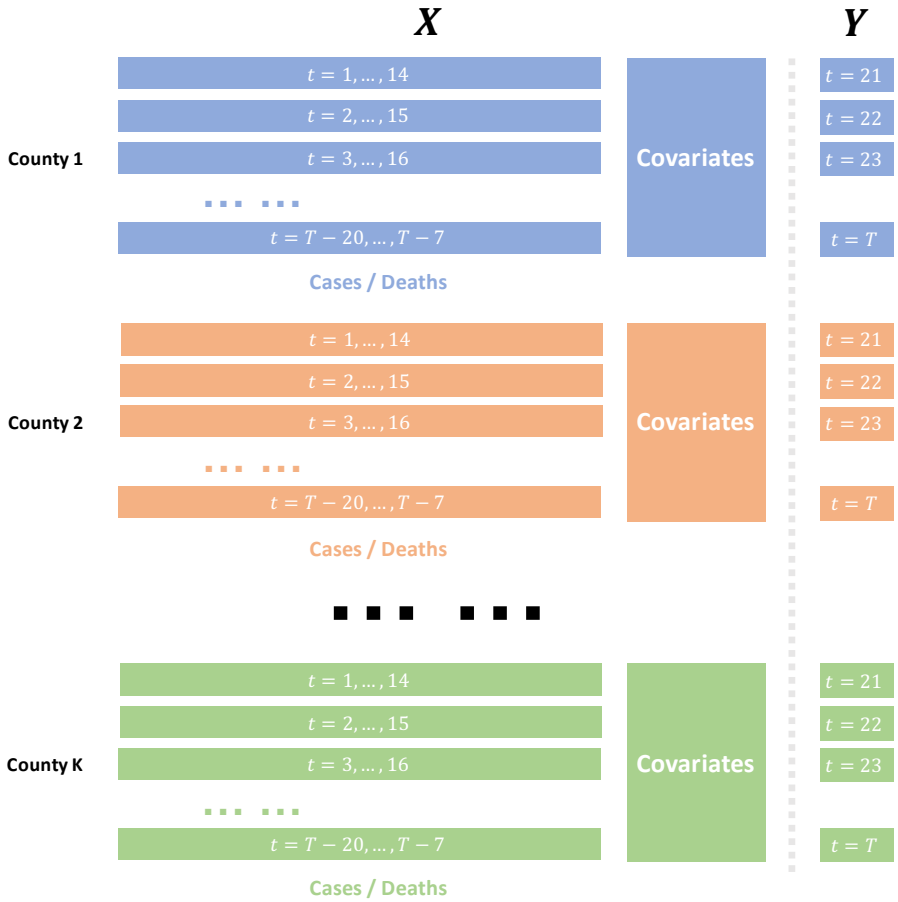

Figure 1.: The structure of data usage in the models. Cumulative data for each county (confirmed or death cases) from the preceding 1st to 14th days serve as independent variables (X) for predicting the cumulative data (confirmed or death cases) as response variable Y on the 21st day. This process is repeated for subsequent days for each county. The method also integrates covariate data from different counties, collectively inputting them into the model, and conducts separate modeling for confirmed and death cases.

Our goal is to incorporate important features that can enhance the accuracy and interpretability of COVINet. To achieve this, we employ the random forest to screen the three actionable features. In a random forest, a common practice is to select features with the largest variances [8]. This approach selects the following three features: traffic volume, severe housing problems, and air pollution ($PM_{2.5}$) (Table 2). Therefore, as presented in Figure 1, our proposed model uses nineteen features as the input data, comprising fourteen historical epidemic features and five county features (three selected actionable features, longitude, and latitude). Note that the input data are not predicted from the model.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

2.3. COVINet

2.3.1. Model architecture

Our proposed model integrates an LSTM layer, a GRU layer [5, 11, 19], and a fully connected layer, formulated as:

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

where $g^{(dense)}$ is a fully connected layer, $g^{(LSTM)}$ is an LSTM layer, and $g^{(GRU)}$ is a GRU layer. The time series of historical epidemic data $\mathbf{X}_{..k}^{(main)}$ 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 $\mathbf{X}_{k}^{(cov)}$ in a fully connected layer.

An LSTM layer ($g^{(LSTM)}$) 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. $\mathbf{X}_{t\cdot k}^{(main)}$ is a tth row of $\mathbf{X}_{..k}^{(main)}$ used as the input vector of the LSTM layer, then the iterative formula for each item is shown as follows:

$$in_t = \sigma\left(W_i \mathbf{X}_{t\cdot k}^{(main)} + U_i h_{t-1}^{(LSTM)} + b_i\right),$$

$$f_t = \sigma\left(W_f \mathbf{X}_{t\cdot k}^{(main)} + U_f h_{t-1}^{(LSTM)} + b_f\right),$$

$$o_t = \sigma\left(W_o \mathbf{X}_{t\cdot k}^{(main)} + U_o h_{t-1}^{(LSTM)} + b_o\right),$$

$$\tilde{C}_t = tanh\left(W_c \mathbf{X}_{t\cdot k}^{(main)} + U_c h_{t-1}^{(LSTM)} + b_c\right),$$

$$C_t = f_t \bigotimes C_{t-1} \bigoplus in_t \bigotimes \tilde{C}_t,$$

$$h_t^{(LSTM)} = o_t \bigotimes tanh\left(C_t\right).$$

Comparatively, a GRU layer ($g^{(GRU)}$) 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:

$$r_t = \sigma\left(W_r \mathbf{X}_{t\cdot k}^{(main)} + U_r h_{t-1}^{(GRU)} + b_r\right),$$

$$z_t = \sigma\left(W_z \mathbf{X}_{t\cdot k}^{(main)} + U_z h_{t-1}^{(GRU)} + b_z\right),$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

$$\tilde{h}_t = tanh\left(W_h \mathbf{X}_{t \cdot k}^{(main)} + U_h\left(r_t \bigotimes h_{t-1}^{(GRU)}\right) + b_h\right),$$

$$h_t^{(GRU)} = (1 - z_t) \bigotimes h_{t-1}^{(GRU)} \bigoplus z_t \bigotimes \tilde{h}_t,$$

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, $\bigotimes$ and $\bigoplus$ are pointwise multiplication, pointwise addition, respectively.

For a fully connected layer ($g^{(dense)}$), 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

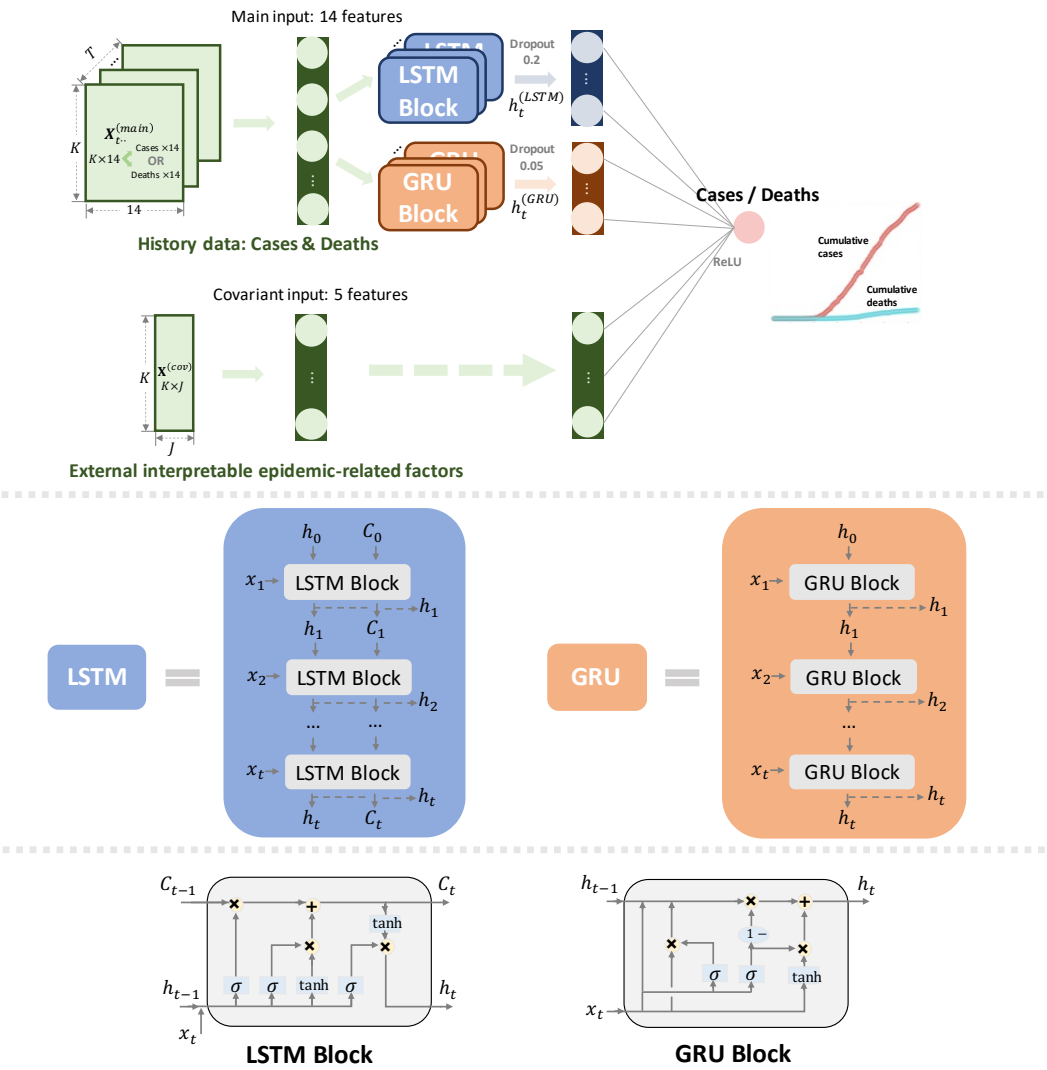

Figure 2.: The COVINet combines Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) using J (5) county features.

All data involved in the model are min-max normalized within each state before being used. The data from New York City (New York), Macomb (Michigan), Oakland (Michigan), Wayne (Michigan), Cook (Illinois), and Wayne (Illinois), Tarrant (Texas) are normalized separately from the rest of the data in their respective states, because their scales are much larger. 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 in the past fourteen days in each county of the US are used to predict the cumulative confirmed cases and deaths in the 7th day in the future. COVINet is trained to learn

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

the observed patterns of COVID-19 and then to validate the learned patterns, where the accuracy of the models is evaluated by Mean Absolute Errors(MAE_t) and Mean Relative Errors (MRE_t) as validation loss:

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

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

where Actual_i are the actual cumulative confirmed cases or total deaths at the ith 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 (MSE).

We assess the performance of all models through temporal domains. In the comparison between COVINet and the ten CDC models, we utilize data from January 21, 2020, to January 26, 2021, as the training set and data from January 27 to March 23, 2021, as the test set. For additional evaluation, we assess the prediction accuracy for the last eight weeks leading up to March 23, 2023, focusing on the county with the most severe infections in each of the top 10 states.

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 [5]. 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 α=0.001 (step size or learning rate), β_1=0.9, β_2=0.999 (exponential decay rates for the moment estimates), and ε=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 200 epochs. For the learning rate, if the MRE does not decrease for ten consecutive epochs, we reduce the learning rate to its 30% until the 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

3. Results

3.1. Comparison between COVINet and COVID-19 forecast hub models

Table 2.: Comparison of the performance of COVINet and ten CDC models in predicting the disease dynamics using the MAE and MRE as the evaluation metrics. The results are reported for the top 10 states and all states in the US for a 7-day prediction. The results of COVINet have been averaged over 50 repetitions.

Method	Top 10 States		All States	
	$\mathbf{MAE_7}$	$\mathbf{MRE_7}$	$\mathbf{MAE_7}$	$\mathbf{MRE_7}$
COVINet(Proposed model)	**159.00**	**0.0049**	**49.48**	0.0107
UMass-MechBayes[17]	163.05	0.0058	58.00	0.0079
COVIDhub CDC-ensemble[30]	167.88	0.0060	58.10	**0.0077**
LANL-GrowthRate[26]	173.55	0.0063	62.72	0.0080
MOBS-GLEAM COVID[1]	179.64	0.0065	66.37	0.0083
COVIDhub-baseline [15]	186.20	0.0067	71.64	0.0100
IowaStateLW-STEM [37]	187.25	0.0065	73.46	0.0082
UT-Mobility[39]	196.75	0.0072	72.33	0.0083
CU-select[42]	221.26	0.0074	82.07	0.0096
CU-nochange[27]	221.78	0.0075	82.18	0.0096
JHU-IDD-CovidSP[22]	409.73	0.0157	132.71	0.0216

Table 2 shows the results for the top 10 states and all states in the US for a 7-day prediction. COVINet exhibits outstanding performance with the lowest MAE values among the top 10 states (159.00) and for all states (49.48). Also, COVINet achieves favorable MRE outcomes, with 0.0049 for the top 10 states and 0.0107 for all states. The latter MRE value is close to the minimum MRE of 0.0077 achieved by COVIDhub CDC-ensemble for all states.

3.2. Prediction of future trajectories of COVID-19 in the most severe county in each of the top 10 states

The MRE_7 and MRE_{30} between the observed and projected counts from the day after training periods to March 23, 2023, are computed to assess the accuracy of the temporal prediction for the most severe county in each of the top 10 states, because those hot-hit areas were of the most severe public health interest. Table 3 presents individual MRE_7 and $\mathrm{MRE}_{30}$for those ten counties using COVINet. Overall, the MRE_7 and MRE_{30} 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 most severe county in each of the top 10 states.

The 30-day projected trajectories of the cumulative confirmed cases and deaths using the COVINet from August 10, 2022, to March 23, 2023, are presented in Figure 3. From Figure 3, the predicted cumulative confirmed cases from August 10, 2022, to March 23, 2023, are remarkably close to the actual ones for the six counties. The situation is similar in predicting the death counts. The projected values of the confirmed cases for the six counties would increase at a slow rate in the near future.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

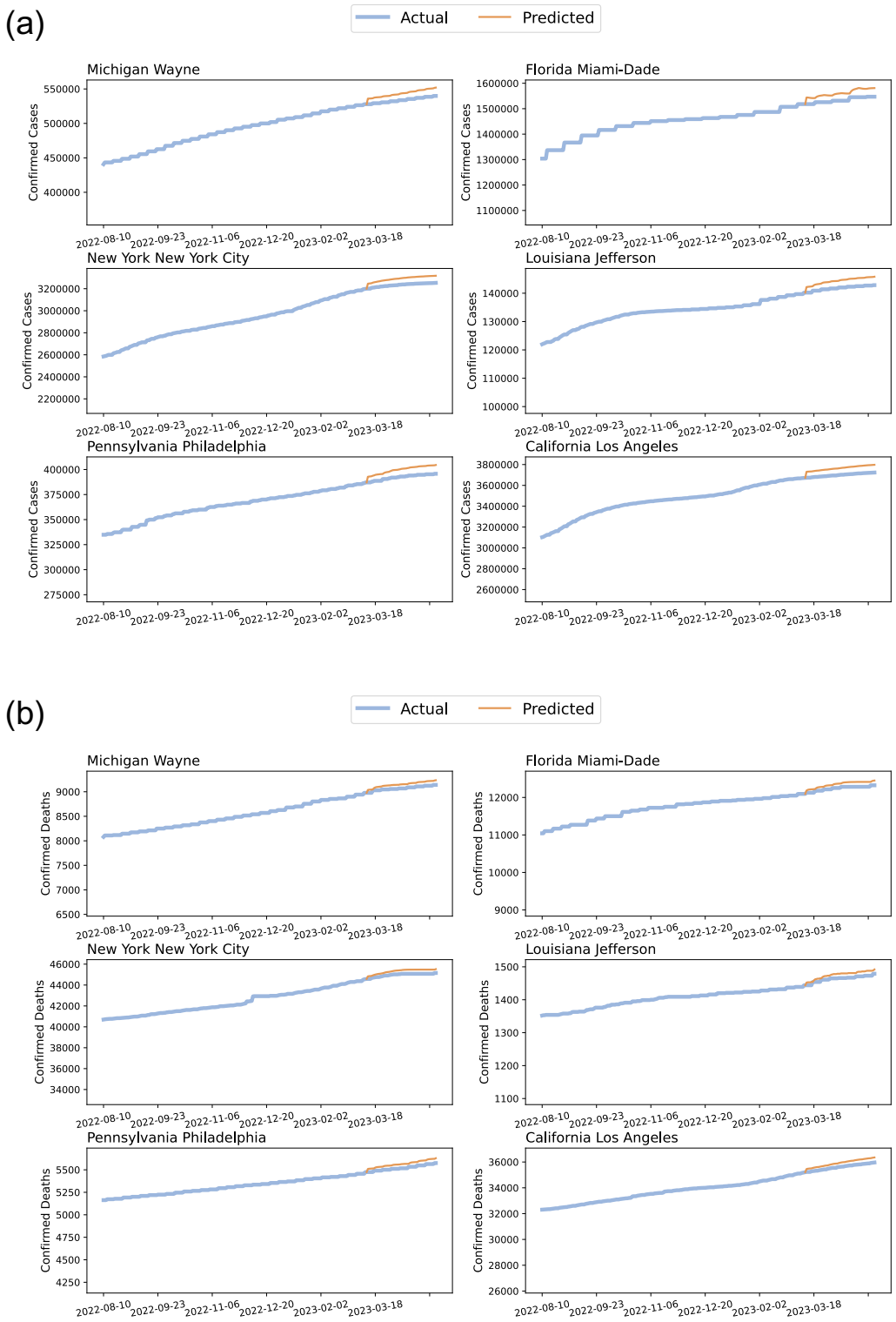

Figure 3.: The trajectories of COVID-19 cumulative confirmed cases (a) and total deaths (b) for six counties from August 10, 2022, to March 23, 2023, are displayed. The blue curves indicate the actual cumulative confirmed cases and total deaths, while the orange curves indicate the predicted ones from February 17, 2023, to March 23, 2023.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3.: MRE_7 and MRE_{30} of cumulative confirmed cases and deaths using COVINet model with three selected features for each of the ten most severe counties of COVID-19.

State, County	**MRE_7**		**MRE_{30}**	
	Confirmed cases	Deaths	Confirmed cases	Deaths
Florida, Miami-Dade	0.0168	0.0087	0.0732	0.0226
Louisiana, Jefferson	0.0167	0.0080	0.0752	0.0161
Connecticut, Fairfield	0.0162	0.0103	0.0831	0.0649
California, Los Angeles	0.0162	0.0089	0.0806	0.0193
Michigan, Wayne	0.0169	0.0086	0.0723	0.0241
Pennsylvania, Philadelphia	0.0176	0.0081	0.0741	0.0199
Illinois, Cook	0.0166	0.0081	0.0701	0.0263
Massachusetts, Middlesex	0.0158	0.0079	0.0745	0.0111
New Jersey, Bergen	0.0177	0.0056	0.0772	0.0177
New York, New York City	0.0163	0.0075	0.0769	0.0226

3.3. Feature effects on COVID-19

Our COVINet model incorporates three 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 1.897×10^{-3} and 4.107×10^{-4} for the confirmed cases and 2.012×10^{-3} and 1.021×10^{-3} 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 March 23, 2023, are presented in Figure 4 and are consistent with our prediction. There are more infected counties in the Northern and Eastern regions.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

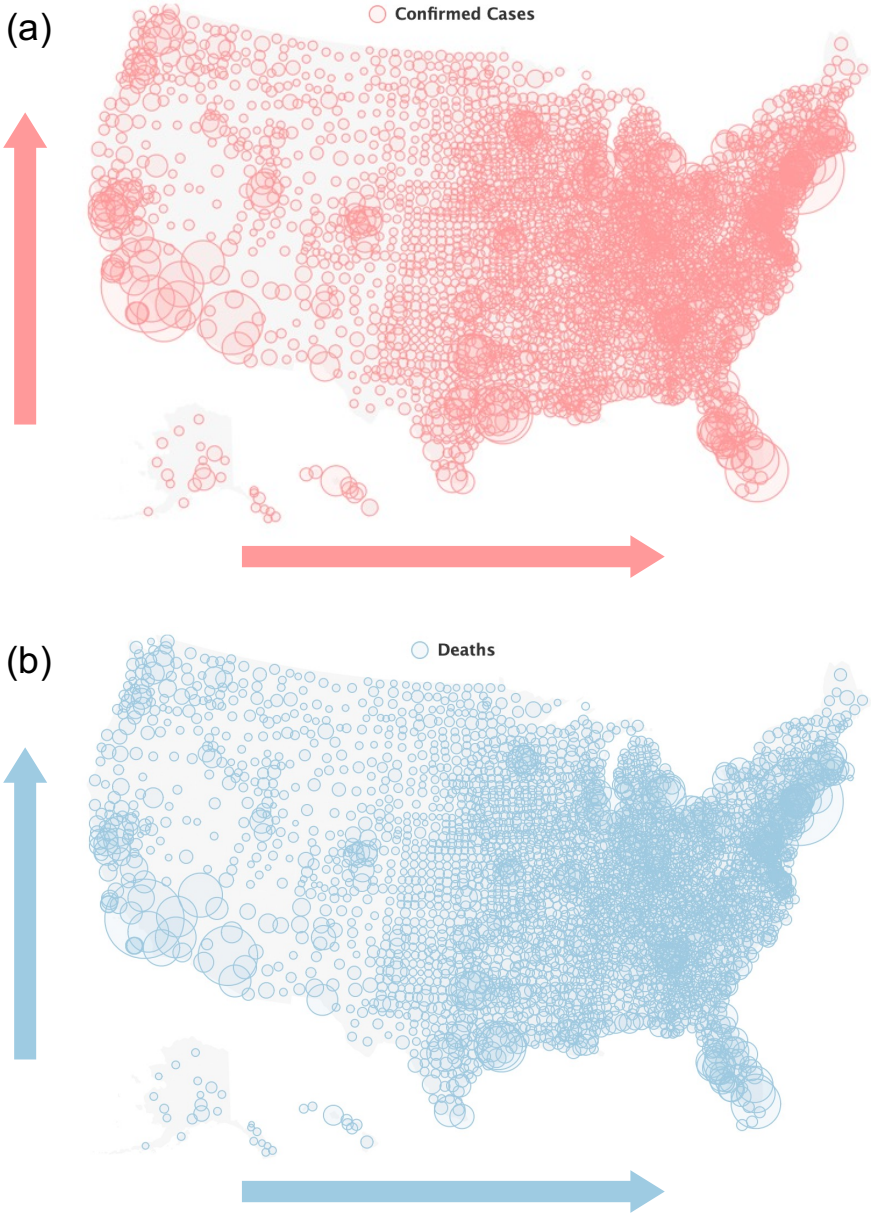

Figure 4.: The map of all infected counties. The circle sizes indicate the number of cumulative confirmed cases (a) and deaths (b) on March 23, 2023. The arrows indicate the trend of change in confirmed cases and deaths over longitudes and latitudes.

The weights of the three selected adverse health risk factors are positive for both confirmed cases and deaths. For example, the largest values of weights for confirmed cases and deaths are the traffic volume at 1.783×10^{-3} and 1.626×10^{-3}, respectively. Specifically, an increase in the traffic volume, severe housing problems, and air pollution would increase both the cumulative confirmed cases and deaths.

To offer insight into the prediction dynamics of COVINet, we vary the levels of the three actional features and present the resulting trajectories of COVID-19 for Los Angeles County, California, as shown in Figure 5. Moreover, for better visibility, we draw the projected trajectories of COVID-19 from March 3, 2023, to March 18,

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4.: The weights of five adverse health risk factors.

Factors	**Weights (Confirmed Cases)**	**Weights (Deaths)**
Traffic volume	1.783×10^{-3}	1.626×10^{-3}
Severe housing problems	7.418×10^{-4}	1.236×10^{-3}
Air pollution	1.603×10^{-4}	3.184×10^{-4}
Longitude	1.897×10^{-3}	2.012×10^{-3}
Latitude	4.107×10^{-4}	1.021×10^{-3}

2023. The impact of the three actional features on COVID-19 in both the cumulative confirmed cases and deaths is visible, depending on the weights of the features. Overall, the number of cumulative confirmed cases and deaths are projected to rise slowly in the following days in Los Angeles County, California. The changes in traffic volume and severe housing problems have a greater impact on the number of confirmed cases and deaths than the changes in air pollution ($PM_{2.5}$), as varying their levels leads to diverging trajectories of confirmed cases and total deaths. The impact of air pollution on the COVID-19 pandemic is relatively slight, as shown by the minimal changes in the cumulative confirmed cases and total deaths across different levels of exposure.

4. Discussion

Our COVINet is built by deep learning and is shown to be an effective model, which elegantly predicts the cumulative confirmed cases and deaths in US counties. 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.

LSTM combined with GRU was shown to capture more temporal information, consistent with the work proposed by Dutta et al. [5]. 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 [5].

To train COVINet, we use the cumulative data (confirmed or death cases) of each county from the previous fourteen days to predict the cumulative data on the 21st day. This time window is chosen because the data from the previous fourteen days contains enough information to capture the trend and the periodicity of the COVID-19 spread. Moreover, the rolling of the data may remove the weekly effect, leading to the model's better fit of the pattern of COVID-19 trajectories. By rolling the data every day, we can eliminate the weekly effect that may introduce noise or bias to the prediction. For example, the number of confirmed cases might be lower on weekends due to less testing or reporting [7].

In our study, we find that the higher the traffic volume, the higher the risk of COVID-19 spread. A study [44] found that traffic volume was positively associated with COVID-19 incidence and mortality after controlling for population density, income, and others. Traffic volume may reflect the level of human mobility, social contact,

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

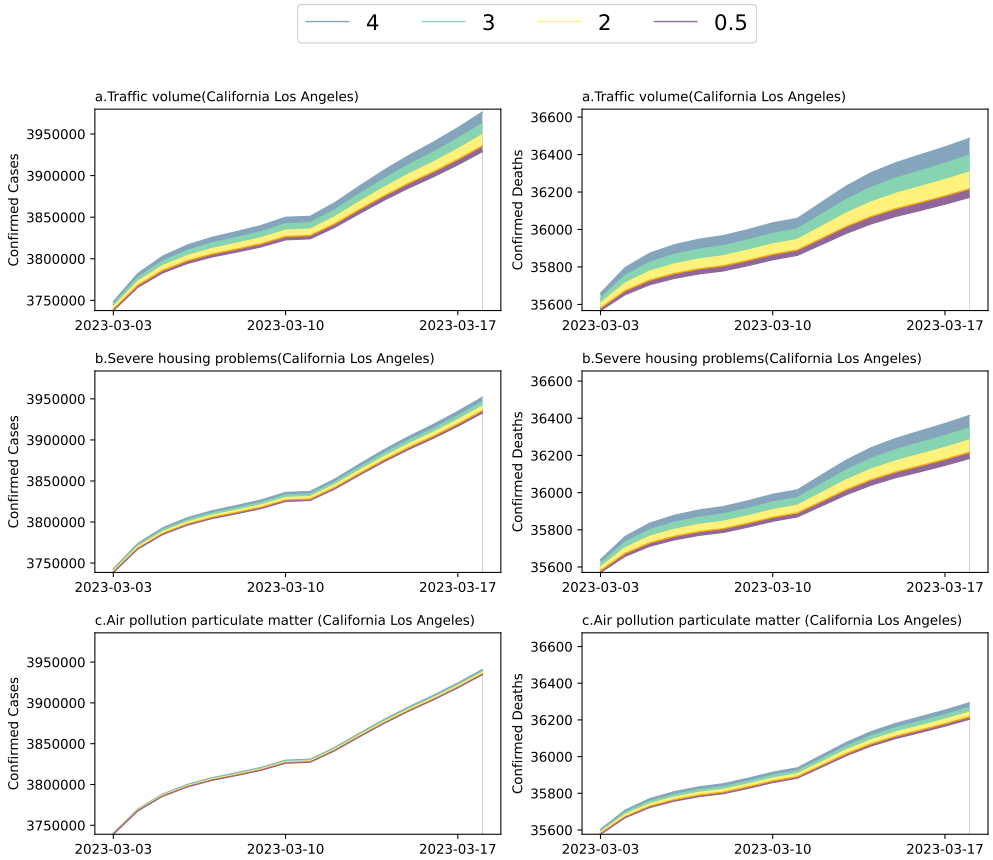

Figure 5.: The projected relative trajectories of COVID-19 for Los Angeles County, California, of cumulative confirmed cases and deaths from March 3, 2023 to March 18, 2023. The levels of the three risk factors are changed from 0.5 times to 4 times since January 27, 2023.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

and exposure to the virus, which are all crucial for the transmission and outcome of the disease. Moreover, the quality of housing, which may affect the immune system, the respiratory system, and the mental health of the residents, has also been linked to higher COVID-19 infection and death rates. Studies in the US [3] and UK [35] have shown that poor housing conditions, such as overcrowding, dampness, and lack of ventilation, make people more susceptible and vulnerable to COVID-19.

As for air pollution, studies indicate that pre-existing cardiovascular disease could increase the severity of COVID-19 [16, 45], so does the air pollution [12, 13]. The residential proximity to high vehicle traffic at a distance would increase exposure to air pollution and risk of cardiovascular disease (CVD) [6, 9, 25]. However, studies [2, 10, 28] have shown that air pollution has a slight impact on COVID-19 infections, which is in line with the small weights assigned by COVINet to this covariate compared to others.

Overall, if the values of those adverse health factors increase, the trajectories of COVID-19 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 significantly in the COVID-19 trajectories. As a result of the COVID-19 pandemic, it is a public health matter and an issue of social responsibility.

The estimated weights of covariates in Table 4 align with the variable importance rank obtained from the random forest estimation in Table 1, as well as with the simulated results in Figure 5. The rank for traffic volume, severe housing problems, and air pollution ($PM_{2.5}$) is consistently from large to small. The high degree of consistency in variable importance across different models provides evidence to a certain extent that our model is credible and reliable. There might be other factors that we could consider in building the COVINet. However, we chose to use the three actionable adverse health factors based on a criterion in the random forest, and they may be controllable by local authorities relatively quickly.

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 [29, 31]. As the most severe county in the US, the Los Angeles County of California is located in the southwest of the US with the highest number of cases of COVID-19 since 2020. However, for the overall hot-spot areas of COVID-19, approaching north (higher values in the latitude) and east (higher values in the longitude) areas of the US, 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 US.

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. COVINet, to some extent, demystifies the black box of deep learning, providing decision-makers with intuitive insights into the impact of health factors on the epidemic. Consequently, it is easy to understand and act by the decision-makers.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

5. Conclusions

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.

Disclosure statement

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Data are publicly available from the New York Times. The code implementation is available at `https://github.com/tingT0929/COVINet-COVID-19`.

Competing interests

The authors declare that they have no competing interests.

References

[1] *Modeling of covid-19 epidemic in the united states.* `http://covid19.gleamproject.org/#about`.

[2] A. Adhikari and J. Yin, *Short-term effects of ambient ozone, pm2. 5, and meteorological factors on covid-19 confirmed cases and deaths in queens, new york*, International journal of environmental research and public health 17 (2020), p. 4047.

[3] K. Ahmad, S. Erqou, N. Shah, U. Nazir, A.R. Morrison, G. Choudhary, and W.C. Wu, *Association of poor housing conditions with covid-19 incidence and mortality across us counties*, PloS one 15 (2020), p. e0241327.

[4] S. Balter, L.S. Gupta, S. Lim, J. Fu, S.E. Perlman, and N.Y.C.H.F.I. Team, *Pandemic (h1n1) 2009 surveillance for severe illness and response, new york, new york, usa, apriljuly 2009*, Emerging infectious diseases 16 (2010), p. 1259.

[5] S.K. Bandyopadhyay and S. Dutta, *Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release*, medRxiv (2020), p. 2020.03.25.20043505.

[6] L.M. Baumann, C.L. Robinson, J.M. Combe, A. Gomez, K. Romero, R.H. Gilman, L. Cabrera, N.N. Hansel, R.A. Wise, and P.N. Breysse, *Effects of distance from a heavily transited avenue on asthma and atopy in a periurban shantytown in lima, peru*, Journal of Allergy Clinical Immunology 127 (2011), pp. 875–882.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

[7] A. Bergman, Y. Sella, P. Agre, and A. Casadevall, *Oscillations in us covid-19 incidence and mortality data reflect diagnostic and reporting factors*, Msystems 5 (2020), pp. e00544–20.

[8] L. Breiman, *Random forests*, Machine learning 45 (2001), pp. 5–32.

[9] B. Brunekreef, R. Beelen, G. Hoek, L. Schouten, S. Bausch-Goldbohm, P. Fischer, B. Armstrong, E. Hughes, and M. Jerrett, *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) (2009), pp. 5–71; discussion 73–89.

[10] F. Cai, K. Yin, and M. Hao, *Covid-19 pandemic, air quality, and pm2. 5 reduction-induced health benefits: a comparative study for three significant periods in beijing*, Frontiers in Ecology and Evolution 10 (2022), p. 885955.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, *Empirical evaluation of gated recurrent neural networks on sequence modeling*, arXiv: Neural and Evolutionary Computing (2014).

[12] 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 (2020), p. 114465.

[13] D. Contini and F. Costabile, *Does air pollution influence covid-19 outbreaks?* (2020).

[14] E.Y. Cramer, Y. Huang, Y. Wang, E.L. Ray, M. Cornell, J. Bracher, A. Brennen, A.J. Castro Rivadeneira, A. Gerding, K. House, D. Jayawardena, A.H. Kanji, A. Khandelwal, K. Le, J. Niemi, A. Stark, A. Shah, N. Wattanachit, M.W. Zorn, N.G. Reich, and U.C..F.H. Consortium, *The united states covid-19 forecast hub dataset*, Scientific Data (2022). Available at `https://doi.org/10.1038/s41597-022-01517-w`.

[15] E.Y. Cramer, E.L. Ray, V.K. Lopez, J. Bracher, A. Brennen, A.J. Castro Rivadeneira, A. Gerding, T. Gneiting, K.H. House, Y. Huang, *et al.*, *Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states*, Proceedings of the National Academy of Sciences 119 (2022), p. e2113561119.

[16] E. Driggin, M.V. Madhavan, B. Bikdeli, T. Chuich, J. Laracy, G. Biondi-Zoccai, T.S. Brown, C. Der Nigoghossian, D.A. Zidar, and J. Haythe, *Cardiovascular considerations for patients, health care workers, and health systems during the covid-19 pandemic*, Journal of the American College of Cardiology 75 (2020), pp. 2352–2371.

[17] G.C. Gibson, N.G. Reich, and D. Sheldon, *Real-time mechanistic bayesian forecasts of covid-19 mortality*, medRxiv (2020).

[18] 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).

[19] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural Computation 9 (1997), pp. 1735–1780.

[20] Z. Hu, Q. Ge, S. Li, L. Jin, and M. Xiong, *Artificial intelligence forecasting of covid-19 in china*, arXiv preprint arXiv:.07112 (2020).

[21] D.P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv: 1412.6980 (2014).

[22] J.C. Lemaitre, K.H. Grantz, J. Kaminsky, H.R. Meredith, S.A. Truelove, S.A. Lauer, L.T. Keegan, S. Shah, J. Wills, K. Kaminsky, *et al.*, *A scenario modeling pipeline for covid-19 emergency planning*, Scientific reports 11 (2021), p. 7534.

[23] 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 140 (2020), p. 110156.

[24] 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).

[25] National Weather Service, *Counties of the u.s used by nws to issue county based forecasts and warnings*, `https://www.weather.gov/gis/Counties` (2020).

[26] D. Osthus, *Lanl covid-19 cases and deaths forecasts*, Website (2020). `https://covid-19.bsvgateway.org`.

[27] S. Pei and J. Shaman, *Initial simulation of sars-cov2 spread and intervention effects in*

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

the continental us, MedRxiv (2020), pp. 2020–03.

[28] O. Ranzani, A. Alari, S. Olmos, C. Milà, A. Rico, J. Ballester, X. Basagaña, C. Chaccour, P. Dadvand, T. Duarte-Salles, *et al.*, *Long-term exposure to air pollution and severe covid-19 in catalonia: a population-based cohort study*, Nature Communications 14 (2023), p. 2916.

[29] M.M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, and A. Amoroso, *Temperature and latitude analysis to predict potential spread and seasonality for covid-19*, Available at SSRN 3550308 (2020).

[30] K. Santosh, *Covid-19 prediction models and unexploited data*, Journal of medical systems 44 (2020), p. 170.

[31] M. Sarmadi, *Association of covid-19 global distribution and environmental and demographic factors: An updated three-month study*, Environmental Research (2020), p. 109748.

[32] The County Health Rankings & Roadmaps program, *State rankings data & reports*, `https://www.countyhealthrankings.org/reports/county-health-rankings-reports` (2020).

[33] The New York Times, *Coronavirus in the u.s.: Latest map and case count*, `https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html` (2020).

[34] 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 10 (2021), pp. 1–8.

[35] A. Tinson and A. Clair, *Better housing is crucial for our health and the covid-19 recovery*, The Health Foundation 20 (2020), pp. 1–25.

[36] 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 203 (2011), pp. 828–837.

[37] L. Wang, G. Wang, L. Gao, X. Li, S. Yu, M. Kim, Y. Wang, and Z. Gu, *Spatiotemporal dynamics, nowcasting and forecasting of covid-19 in the united states*, arXiv preprint arXiv:2004.14103 (2020).

[38] 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 18 (2020), pp. 409–432.

[39] S. Woody, M. Tec, M. Dahan, K. Gaither, M. Lachmann, S.J. Fox, L.A. Meyers, J. Scott, and U. of Texas at Austin COVID-19 Modeling Consortium, *Projections for first-wave covid-19 deaths across the us using social-distancing measures derived from mobile phones*, Medrxiv (2020), pp. 2020–04.

[40] World Health Organization, *Who director-general's opening remarks at the media briefing on covid-19*, `https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020` (2020).

[41] 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 395 (2020), pp. 689–697.

[42] T. Yamana, S. Pei, S. Kandula, and J. Shaman, *Projection of covid-19 cases and deaths in the us as individual states re-open may 4, 2020*, MedRxiv (2020), pp. 2020–05.

[43] Z. Yang, Z. Zeng, K. Wang, S.S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, and J. He, *Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions*, Journal of Thoracic Disease 12 (2020), pp. 165–174.

[44] Y.J. Yasin, M. Grivna, and F.M. Abu-Zidan, *Global impact of covid-19 pandemic on road traffic collisions*, World journal of emergency surgery 16 (2021), pp. 1–14.

[45] Y.Y. Zheng, Y.T. Ma, J.Y. Zhang, and X. Xie, *Covid-19 and the cardiovascular system*,

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20113787; this version posted December 18, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Nature Reviews Cardiology 17 (2020), pp. 259–260.