2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application

Enhance COVID-19 Mortality Prediction with Human Mobility Trend and Medical Information

Yogesh Chaudhari, Indrajeet Javeri, Ismailcem Arpinar and John A. Miller
Department of Computer Science
The University of Georgia
Athens, USA
{yogesh.chaudhari; indrajeet.javeri; budak; jamill}@uga.edu

Xiaochuan Li, Bingnan Li and Yuan Ke
Department of Statistics
The University of Georgia
Athens, USA
{xiaochuan.li; Bingnan.Li1; yuan.ke}@uga.edu

Mohammadhossein Toutiaee
Khoury College of Computer Sciences
Northeastern University
San Jose, USA
m.toutiaee@northeastern.edu

Nicole Lazar
Department of Statistics
The Pennsylvania State University
University Park, USA
nfl5182@psu.edu

***Abstract*—In this work, we study national and state-level COVID-19 pandemic data in the United States with the help of human mobility trend data and auxiliary medical information. We analyze and compare various state-of-the-art time-series prediction techniques. We assess a spatio-temporal graph neural network model which forecasts the pandemic course by utilizing a hybrid deep learning architecture and human mobility data. Nodes in the graph represent the state-level deaths due to COVID-19 at any particular time point, edges represent the human mobility trend and temporal edges correspond to node attributes across time. We also study statistical modeling and machine learning techniques for mortality prediction in the United States. We evaluate these techniques on both state and national level COVID-19 data in the United States and claim that the SARIMAX and GCN-LSTM model generated forecast values using exogenous hospital information variables can enrich the underlying model to improve the prediction accuracy at both levels. Our best machine learning models perform 50% and 60% better than the baseline on an average on the national level and state-level data, respectively, while the statistical models perform 63% and 42% better.**

***Index Terms*—COVID-19, Time-Series Analysis, Graph Neural Networks, SARIMAX, Medical Information**

I. Introduction

The COVID-19 pandemic has infected over 223 million individuals and has resulted in the deaths of over 4.6 million worldwide since its beginning in late 2019 [1] (some evidence is that patient zero dates from November 17^{th}) until the 10^{th} September 2021. One of the worst impacted countries in the world is the United States with more than 40 million infections and 649 thousand deaths [2]. Data collected within the United States show different infection and mortality patterns of the disease in different states. For example, California has the highest number of reported cases and deaths, Tennessee has the highest cases per million population, and New Jersey has the highest deaths per million population. Although one can drill down to the county level, the error canceling capabilities of state-level data over different counties can lead to more accurate reporting.

COVID-19 provides a unique opportunity in studying pandemics due to increased data collection. One challenge in the data collection is to ensure the information is accurate. Since the beginning of the COVID-19, many organizations have been collecting and aggregating pandemic-related medical data, such as the number of infected individuals, number of deaths, number of hospitalized patients, number of patients in the Intensive Care Unit (ICU), number of individuals vaccinated, etc. However, these data have a lot of discrepancies due to reporting errors, miscalculations, systematic biases, and rollbacks in various data sources. Therefore, first and foremost, it is critical for us to ensure that the dataset we study is of good quality.

The ability to forecast the number of infections and deaths is vital not only for those who work in the healthcare system to look ahead but also for policymakers since they have the power to manage healthcare resources, control disease upsurges, and take preventive actions when necessary to protect public health. The COVID-19 Forecast Hub currently lists models developed by 35[1] research groups and their relative performances. These 35 models can be loosely categorized into three modeling groups: SEIR-like, Curve-fitting, and Predictive models, with 16, 4, and 12 models, respectively. There are also 3 that are categorized as ensembles. Our work mainly focuses on predictive modeling techniques.

SEIR-like models: SEIR stands for Susceptible, Exposed, Infectious, and Recovered which are four major states that an individual can experience in a pandemic. A set of mathematical rules is used to govern the spread of the disease and perform forecasting [3]. An extension of the SEIR model that can be used to forecast hospitalizations (H) and deaths (D) as

[1]https://covid19forecasthub.org/eval-reports/

978-1-6654-9457-1/21/$31.00 ©2021 IEEE
DOI 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00190

well, is shown in the state transition diagram in Figure 1 and is called the SEIHRD model.

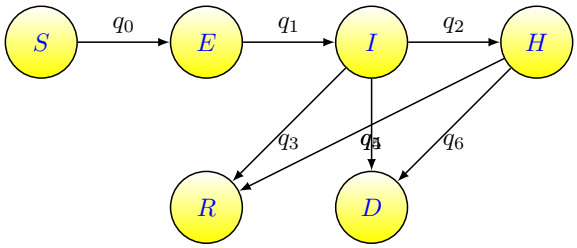

Fig. 1. State Transition Diagram for SEIHRD Models

If there is a significant reinfection rate, an edge can be added from state R to state S. The model equations are parameterized by the state transition rates which control the dynamics of the system.

Curve fitting models: The curve fitting method refers to extrapolating an analytical function that can be used for performing forecasts. Noise is added to the function or its derivative to capture local fluctuations [4]. When the sample size is small, complicated models are hard to train, and simple models with few parameters can be useful to capture fundamental characteristics of pandemics. For instance, modified exponential curves tend to work well for pandemics [5].

Predictive models: Predictive models for forecasting include statistical models such as Auto-Regressive, Integrated, Moving Average (ARIMA), Vector Auto-Regressive (VAR), Minimax Concave Penalty (MCP), etc., and machine learning models such as Neural Networks, Gradient Boosting Regression, etc. As such models can be highly parameterized and are inherently data-dependent, they have the potential to perform very well for short and near-term forecasting. Other techniques that incorporate the mechanics of the pandemic (e.g., SEIR models) may do better for longer-term forecasting. [6]

Over the last year, many academic research groups, government agencies, industrial research teams, and individuals have generated COVID-19 forecasts using one or more of the above-mentioned techniques. An ensemble model that uses the forecasts from 23 [7] to 35 research groups that regularly submit has been created by the United States Centers for Disease Control and Prevention (CDC) to analyze the progress of the pandemic as well as communicate with the general public. Analyzing the performance of different models can be complicated since not all the studies use the same metrics for evaluating their models. The COVID-19 Forecast Hub compares all the models that are used as a part of the CDC ensemble model with a baseline model. To be specific, their metric is the relative Mean Absolute Error (MAE) between the models and the baseline.

In this study, we extend the current research of pandemic modeling by studying statistical and machine learning methods that can be used to forecast *daily deaths* caused by COVID-19 in the United States. We focus on daily deaths not only because of their extreme importance but also due to their impact on model interpretability. Although we have examined cumulative deaths, all models tend to be relatively accurate as they are predicting slowly changing large numbers, whereas, models become highly differentiated when looking at daily deaths. Figure 2 shows daily deaths and the weekly moving average in the US.

We also propose a spatio-temporal graph convolutional network that can capture complex dynamics by including mobility patterns across different states. In addition, we utilize exogenous medical information-related variables like "COVID-19 hospitalizations" and "number vaccinated" to improve the prediction performance of our models. We demonstrate that such exogenous medical variables combined with lagged variables within the predictive models have strong potential to enhance prediction and monitor virus spread.

The rest of this paper is organized as follows: Section II discusses different datasets available on COVID-19. Related work is covered in Section III. Section IV and Section V focus on the methodologies and experiments. Results are discussed in Section VI, and conclusions and future work are presented in Section VII.

II. Datasets

A. Data Sources

Since the beginning of the pandemic, there have been multiple publicly available COVID-19 data sources for the number of cases, deaths, hospitalizations, etc. However, these data sources all have strengths and weaknesses. For example, the COVID Tracking Project dataset which was used by multiple research groups began collecting data on January 13^{th}, 2020, but unfortunately stopped collecting data on March 7^{th}, 2021. Some of the most reliable and complete sources of pandemic and related data are listed below.

1) CDC COVID Data Tracker: The official CDC COVID Data Tracker dataset[2] provides the latest time-series data on COVID-19 cases, deaths, and tests performed. Since this dataset does not contain any additional medical information such as hospitalizations or patients in the ICU, it is insufficient for our study.

2) CSSE-JHU COVID-19 Dataset: The COVID-19 dataset by the Center for Systems Science and Engineering at Johns Hopkins University (CSSE-JHU) [8] has many numerical discrepancies when compared with the CDC COVID Data tracker. This dataset, however, is used as the "ground truth" data for COVID-19 cases and deaths on the COVID-19 forecast hub [9] and thus has been used by several research groups. Though it provides complete time series on COVID cases and deaths in the US, other variables such as hospitalizations are incomplete in this dataset.

3) OWID COVID-19 Dataset: The COVID-19 dataset provided by Our World In Data (OWID) [10] is currently the most complete dataset on COVID-19 cases, deaths, hospitalizations,

[2]Avaliable at https://covid.cdc.gov/covid-data-tracker

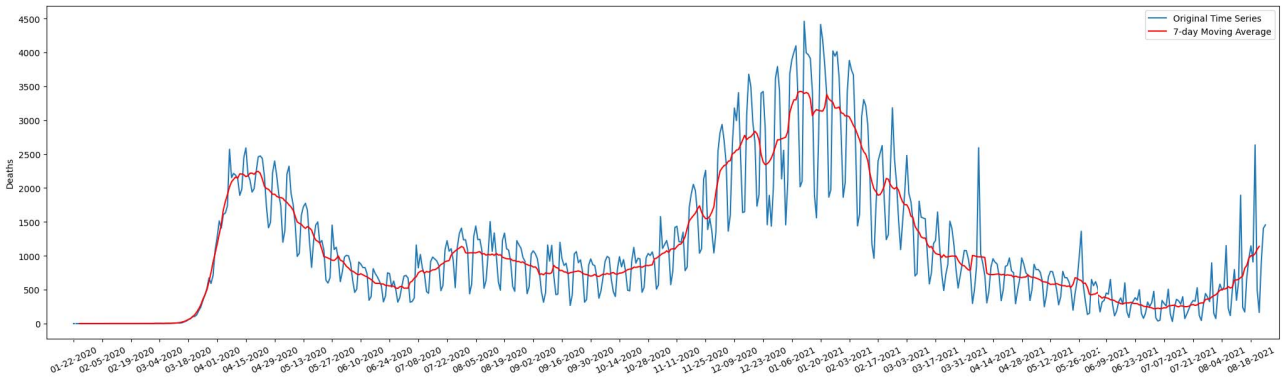

Fig. 2. Daily Deaths in the US

ICU patients, tests, vaccination, etc. The data are aggregated from multiple sources. For example, the death and cases data are collected from the CSSE-JHU dataset. We noticed minor discrepancies between the CSSE-JHU dataset and the OWID dataset. Fortunately, that only has a limited impact on our study. A brief comparison of the three datasets can be found at our GitHub repository[3].

B. Datasets for Evaluation

1) National Level Dataset: We have used the dataset provided by OWID discussed in Section II-A3 to generate forecasts for daily deaths on the national level. It also contains several medical variables such as "hospitalized patients", "ICU patients", "people vaccinated", "people fully vaccinated", etc. The data we collected spans from January 22^{nd}, 2020, to August 25^{th}, 2021 (582 days). We ignore the first 38 days due to missing values and start from February 29^{th}, 2020, when the first COVID-19 death in the US was recorded. As a result, the national level time-series studied in this paper contains 544 days. We set the first 222 days as the initial training set and use a two-week ahead rolling window forecast on the remaining 322 days as the test set. (Note: The training set size at the national level was set to be comparable to the training set size for the state level.)

2) Aggregate Mobility Data: The mobility data used in the study are obtained from COVID-19 US Flows [11]. This dataset consists of dynamic human mobility patterns across the United States in the form of the daily and weekly population flows at three geographic scales: census tract, county, and state. The spatio-temporal data are obtained by analyzing, computing, and aggregating the millions of anonymous mobile phone users' visit trajectories to various places provided by "SafeGraph". To be specific, we use the daily state-to-state population flow starting from April 13^{th}, 2020, to April 15^{th}, 2021. The raw data contains the unique identifiers, latitudes and longitudes for the origin and destination states, date, visitor flows (estimated number of visitors detected by SafeGraph between two geographic units), and the population flow (estimated population flows between two geographic units, inferred from visitor flows).

3) State-Level Dataset: The state-level data we collected spans from April 13^{th}, 2020, to August 25^{th}, 2021. However, since the mobility data after April 15^{th}, 2021, is absent, our models use the state-level deaths data only until April 15^{th}, 2021. This yields a time series consisting of 368 days. Credible sources for medical variables like "hospitalized patients" and "ICU patients" are not available for the state-level data. For this study, we have used "mobility between the states" and "daily new cases" as the extra parameters. We use the first 228 days as the initial training set and implement a two-week ahead rolling window forecast to assess the performance of various methods.

The national level, state-level, and mobitility datasets used in this study are available at our GitHub repository[4].

III. Related Work

A great amount of research has been conducted primarily on forecasting the cumulative number of people affected by the COVID-19 pandemic. Early studies like [12] used the Susceptible, Infected and Recovered (SIR) model for forecasting the infection rate in different countries. Another study [13] proposed an SEIR model which also took into account mobility restrictions due to regional lockdowns. [14] proposed a simple curve fitting model that can be implemented in Microsoft Excel using a log-normal function. Agent-based models, like the one discussed in [15], simulate details of the pandemic such as mobility restrictions, hospitalizations, vaccinations, etc. on a micro-level and can be effective in modeling the transmission of disease. However, agent-based models can be computationally expensive when the sample size is large. Other work, e.g. [16], used ARIMA to predict the daily death rate in different states in the United States. [17] introduced an AUG-NN model that enriches neural network models by augmentation which resulted in significant improvements in accuracy. They reported the forecast values on the national level data. [18] introduced the DeepAR model,

[3]github.com/scalation/data/blob/master/COVID/dataset-comparison.xlsx

[4]http://github.com/scalation/data

an auto-regressive recurrent network approach for forecasting. Although DeepAR is shown to be a powerful tool, there is not much work on its use in forecasting COVID-19.

Time-series forecasting using Graph Neural Networks (GNN) has been introduced in various domains in the past, though it is less studied in epidemic diseases. For example, [19] used a Temporal Graph Convolutional Network (T-GCN) for traffic prediction and [20] used GNN for stock market prediction. [21] discussed a GNN approach for COVID-19 using spatio-temporal mobility data. However, they only reported COVID-19 forecasting on a very small scale, i.e., the top 20 most populated counties in the United States.

IV. Methodology

A. SARIMAX

The Partial Auto-Correlation Function (PACF) on the daily deaths data, as presented in Figure 3a, shows the first 15 lags are significant and there is a clear weekly pattern. Weekly seasonality of daily deaths can also be observed using the Auto-Correlation Function (ACF), as shown in Figure 3b. Then, we propose to fit the daily deaths by a Seasonal Auto-Regressive, Integrated, Moving Average (SARIMA) [22] model defined as

$$\varphi_p(B)\Phi_P\left(B^s\right)\nabla^u\nabla_s^v z_t = \theta_q(B)\Theta_Q\left(B^s\right)\varepsilon_t, \qquad (1)$$

where z_t is a variable to forecast, i.e., the logarithm of *daily deaths*, $t = 1, 2, \ldots$, $\varphi_p(B)$ is a regular AR polynomial of order p, $\theta_q(B)$ is a regular MA polynomial of order q, $\Phi_P\left(B^s\right)$ is a seasonal AR polynomial of order P, and $\Theta_Q\left(B^s\right)$ is a seasonal MA polynomial of order Q. The differencing operator ∇^u and the seasonal differencing operator ∇_s^v eliminate the non-seasonal and seasonal non-stationarity, respectively.

The SARIMA with eXogenous factor (SARIMAX) model is an extension of the SARIMA model in (1), which can include exogenous variables, such as hospitalization, ICU occupancy rate, and vaccination rate. The SARIMAX model can be defined as:

$$\varphi_p(B)\Phi_P\left(B^s\right)\nabla^u\nabla_s^v z_t = \theta_q(B)\Theta_Q\left(B^s\right)\varepsilon_t + \sum_{i=1}^{m}\beta_i x_t^i,$$

where $\{x_t^1, \ldots, x_t^m\}$ are the m exogenous variables defined at time t with coefficients $\{\beta_1, \ldots, \beta_m\}$.

B. Minimax Concave Penalty

We also consider a penalized linear regression approach to predict daily deaths by historical data and medical variables. We denote $z_t^* = \log z_t - \log z_{(t-7)}$ as the weekly log-return of *daily deaths* at time t and $\mathbf{x_t}$ are m explanatory medical variables at time t. We linearly regress z_t^* on $\{z_{t-h}^*, \ldots, z_{t-(h+k-1)}^*\}$ and $\{\mathbf{x}_{t-h}, \ldots, \mathbf{x}_{t-(h+k-1)}\}$, where $k = 14$ and $1 \le h \le 14$ is a horizon parameter. The *horizon* refers to the number of days into the future for which forecast values are to be generated. Since the state-level medical variables are highly correlated, we first implement a sure independent screening [23] procedure to reduce the dimensionality of $\mathbf{x_t}$. As a result, only the top seven explanatory variables are included in the following partial penalized regression model:

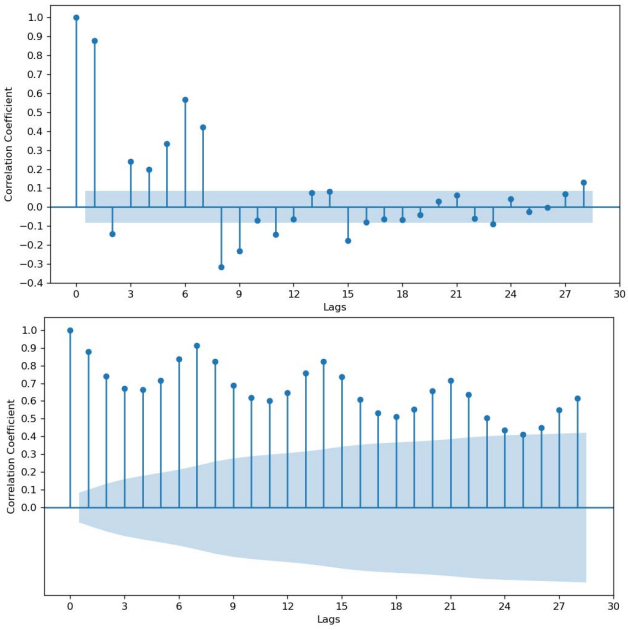

Fig. 3. a) PACF, b) ACF plot of the daily deaths

$$Q(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{z}) = \frac{1}{2N}\|\mathbf{z}^* - \mathbf{X}^\dagger\boldsymbol{\beta}\|^2 + \sum_{j=1}^{k(m+1)} P_\gamma\left(\beta_j; \lambda\right),$$

where $\mathbf{z}^* = (z_1^*, \ldots, z_N^*)^{\mathrm{T}} \in \mathbb{R}^N$ is the vector of weekly log-return of *daily deaths* over a sample size N, $\mathbf{X}^\dagger \in \mathbb{R}^{N\times k(m+1)}$ is the augmented design matrix of all predictors (e.g. lags of z_t and explanatory variables), and $\boldsymbol{\beta} = (\beta_1, ..., \beta_{k(m+1)})^{\mathrm{T}}$ is a vector of unknown regression coefficients. $P_\gamma(\cdot; \lambda)$ is the Minimax Concave Penalty (MCP) [24] which satisfies

$$P_\gamma(\beta; \lambda) = \begin{cases} \lambda|x| - \frac{\beta^2}{2\gamma}, & \text{if } |\beta| \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |\beta| > \gamma\lambda. \end{cases}$$

Since $k(m+1)$ predictors in the regression model tend to overfit, the number of predictors has been reduced from $k(m+1)$ to $m^* = 7$. The $m^* = 7$ explanatory variables are the top m^*, which have larger marginal Pearson's correlation coefficients with the response variable. We also tried the L_1 penalty (LASSO) and smoothly clipped absolute deviation (SCAD) as two alternative penalty functions. We find MCP performs the best among the three due to its smaller penalty away from zero.

C. Vector Auto-Regressive Model

Vector Auto-Regression [25] is another popular approach to model and predict multivariate time-series. For a d-dimensional response vector of interest, say:

$$\mathbf{z}_t = (z_t^{(1)}, z_t^{(2)}, \ldots, z_t^{(d)})^{\mathrm{T}},$$

a vector auto-regressive model of order q, i.e. VAR(q), is defined as

$$\mathbf{z}_t = \boldsymbol{\alpha} + \Phi^{(0)}\mathbf{z}_{t-q} + \Phi^{(1)}\mathbf{z}_{t-q+1} + \ldots + \Phi^{(q-1)}\mathbf{z}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\alpha} \in \mathbb{R}^d$ is an intercept vector, $\Phi^{(s)} \in \mathbb{R}^{d\times d}$ for $s = 0, \ldots, p-1$ are regression coefficient matrices, and $\boldsymbol{\epsilon}_t \in \mathbb{R}^d$ is an error vector.

Similar to the data pre-processing procedure described in Section IV-B, a weekly log-return has been applied to both the response vector and medical variables to remove the seasonality.

D. GCN-LSTM

GCN: The main idea behind Graph Convolutional Network (GCN) models is that a given input signal (node) can be enriched via information propagation from its neighbors to improve a future prediction task. The neighbors are often defined by constructing a network of inputs where nodes and connections represent features and relations between them.

Mobility Network Graph: Models can be improved by adding human mobility data across regions. Mobility data form a spatial graph, where a region i corresponds to a node, every node can be connected to other nodes $j, k, z, \ldots$, and weighted edges represent the strength of relations between the nodes. We constructed a binary adjacency matrix to feed into the trainer. This adjacency matrix was created by the following steps: (1) the average of mobility data along the time point was obtained, and (2) if the number of movements from an origin to a destination was among the top 40%, the corresponding cell was assigned 1, and 0 otherwise. (3) Finally, the matrix was corrected to be a full-rank matrix.

GCN-LSTM: Disease epidemic forecasting is a quintessential example of spatio-temporal problems for which we present a deep neural network framework that captures the number of deaths using spatio-temporal data. The task is challenging due to two main inter-twined factors: (1) the complex spatial dependency between time-series of each state, and (2) non-linear temporal dynamics with changing non-pharmaceutical interventions (NPI) such as mobility trends.

We attempt to populate a temporal knowledge graph using mobility patterns, since people moving around the regions with similar epidemic patterns may contribute to the forecasting process. The people traveling serve as the ground truth for training a GCN for identifying the underlying graph between sub-regions. Next, the constructed graph embedding from the GCN model is used to feed into a Long Short-Term Memory (LSTM) model to forecast the pandemic in the future. Notice that the graph embedding provides knowledge about the forecasting system, and the LSTM provides a direction for how to leverage the GCN output in the COVID-19 forecasting.

E. Wavelet-ANN

The Wavelet-ANN model, as described in [26], is a hybrid model that uses Wavelet Analysis and Artificial Neural Networks (ANN) to perform time-series forecasting. Wavelet Analysis transforms the original signal into a different domain using continuous, discrete, or multiresolution-based wavelet transforms. The Wavelet-ANN model decomposes the original time-series into multiple sub-series, performs wavelet denoising, and then reconstructs the original series to perform forecasting. [27]

F. Augmented Neural Networks

The Augmented Neural Network (AUG-NN) model introduced in [17] deals with the problem of forecasting on intermediate size time-series where the average length is in the hundreds of data points. AUG-NN uses a hybrid of statistical and machine learning approaches to improve forecasting results. First horizon forecasts are generated on the time-series using a statistical approach like the SARIMA model. These forecasts are then used to calculate half-a-day time point values and the time-series is thus doubled in size.

G. Deep-AR

The Deep-AR model [18] is based on an auto-regressive recurrent network architecture. Denote $\mathbf{z}_{i,t_1:t_2} = (z_{i,t_1}, \ldots, z_{i,t_2})^{\mathrm{T}}$ the data collected from the i time-series between time t_1 and t_2. Deep-AR models the current and future of each time-series given its past by the conditional distribution $P(\mathbf{z}_{i,t_0:T}|\mathbf{z}_{i,1:t_0-1}, \mathbf{f}_{i,1:T})$, where $\mathbf{f}_{i,1:T}$ are common factors for the ith time-series. The conditional distribution function is modelled by the product of log-likelihood functions at each time steps from t_0 to T as follows

$$\begin{aligned} P(\mathbf{z}_{i,t_0:T}|\mathbf{z}_{i,1:t_0-1}, \mathbf{f}_{i,t}) &\approx \prod_{t=t_0}^{T} Q_t(z_{i,t}|\mathbf{z}_{i,1:t-1}, \mathbf{f}_{i,t}; \theta_t) \\ &= \prod_{t=t_0}^{T} \ell\left(z_{i,t}|\theta_t(\boldsymbol{h}_{i,t}, \Theta)\right). \end{aligned}$$

The log-likelihood $\ell\left(z_{i,t}|\theta_t(\boldsymbol{h}_{i,t}, \Theta)\right)$ is parametric distribution whose parameters are given by the output $\boldsymbol{h}_{i,t}$ of an auto-regressive recurrent neural network

$$\boldsymbol{h}_{i,t} = h(\boldsymbol{h}_{i_t-1}, z_{i,t-1}, \boldsymbol{f}_{i,t}, \Theta),$$

where h is a function implemented by a multi-layer recurrent neural network with LSTM cells. The neural network $h(\cdot)$ is auto-regressive as it takes the observation at the last time step, i.e. $z_{i,t-1}$ as an input. Besides, $h(\cdot)$ is recurrent since the network output at the last time step, i.e. $\boldsymbol{h}_{i_t-1}$ is fed back to the network as well.

H. GRU and LSTM

Recurrent Neural Networks (RNN) are well known for extracting sequential characteristics from the data, learning patterns, and predicting the next set of sequences. Gated Recurrent Units (GRU) and LSTMs provide the facility for the RNN to learn the information over long sequences of data because of the availability of short-term memory. [28]

I. Random Walk

By definition, a candidate series follows a random walk if the first differences are random (non-stationary). Random Walk (RW) is a common technique in graphical models [29], and it is widely used in webpage ranking, image segmentation, and time-series analysis. Many studies have shown that the RW method is applicable to most time-series data, especially when the samples have the same distribution and are independent of each other. A Gaussian Random Walk for variable z_t can be written by:

$$z_t = z_{t-1} + \epsilon_t, \tag{2}$$

where ϵ_t follows a Gaussian distribution.

J. Rolling Validation for Multiple Horizons

Classical multi-folds cross-validation approaches are not directly applicable to time-series data due to the existence of serial dependence. We follow a rolling-validation scheme in this study. A 14-day ahead forecast is obtained using the model trained on the training set. Then, we move the rolling window forward by a one-step which involves appending the first value from the test set to the training set and removing the first value from the training set. The rolling window moves all the way to the end of the test set and 14-day forecasts are obtained on each step. Models can be trained only once or they can be retrained at a regular frequency.

V. Experiments

A. Hyperparameters and Architectures

SARIMAX: The medical variables are selected among several hospitalization and vaccination related variables, such as "icu_patients", "hosp_patients", "people_vaccinated", etc. The hyperparameter, p, is tuned from 1 to 6. The model that gave the best result is a SARIMAX$(1,1,4)\times(3,1,1,7)$ model with medical variables "hosp_patients" and "people_vaccinated" at the national level, and a SARIMAX$(3,1,4)\times(3,1,1,7)$ model with the state-wise daily cases variable at the state level.

MCP: We fit an MCP model and consider the serial dependence among daily new cases in hospitals. Each day's new hospitalized count is dependent on the previous 14 days of observations. The regularized parameters in MCP are selected by a multi-fold cross-validation. As a result, 7 predictors are included in the MCP model.

VAR: Similar to SARIMAX model, VAR model contains multiple predictors including "icu_patients" and "hosp_patients" in addition to "new_deaths" variable. The hyperparameters are selected by the Bayesian Information Criterion (BIC).

AUG-NN: We use the SARIMAX$(1,1,4)\times(3,1,1,7)$ model mentioned above to generate the first horizon forecasts for augmenting the time-series. Actual 14-day (28-half days) ahead time-series forecasting is generated on a 4 layer neural network, the layers of which are: a) Input layer, b) Dense layer of size 32 with Rectified Linear Unit (ReLU) activation, c) Dense layer of size 1024 with ReLU activation, d) output layer. The neural network architecture is obtained by the network architecture search tool 'AutoKeras'.

Deep-AR: We fit the Deep-AR model to predict "new_death" by inputting 9 related vaccination and hospitalization variables, such as "new_cases", "icu_patients", "people_vaccinated", "people_fully_vaccinated", etc. The hidden layer in RNN is set as 2 layers. The number of epochs, which is the maximum number of passes over the training data, is set as 400, and the early stopping patience is set as 40, with the learning rate set as 5×10^{-4}.

GCN-LSTM: The GCN-LSTM architecture used in this experiment consists of two GCN layers followed by one LSTM layer. The model is built using the `StellarGraph` library [30]. Only one architecture of the model is used for forecasting on all 14 horizons. The size of the two GCN layers used is 32 while the size of the LSTM layer is 300. Information from 10 previous lags is used to forecast future instances. 182 samples from the training set are used for actual training and the rest of the 46 samples are used as the validation set.

Code for all our experiments is available on GitHub.[5]

VI. Results

We compare the performance of all the models with a baseline Random Walk (RW) model. For each model, we make a 14-day ahead rolling window forecast as introduced in Section IV-J. The forecast performance is measured by the symmetric Mean Absolute Percentage Error (sMAPE) score at each horizon $1 \le h \le 14$.

The national-level forecast results are reported in Table I. The SARIMAX model performs the best on this dataset with the lowest sMAPE score. MCP and VAR are the second and third best methods in terms of the average sMAPE over all horizons. All the statistical models, including SARIMA which does not use any exogenous variables, outperformed machine learning models.

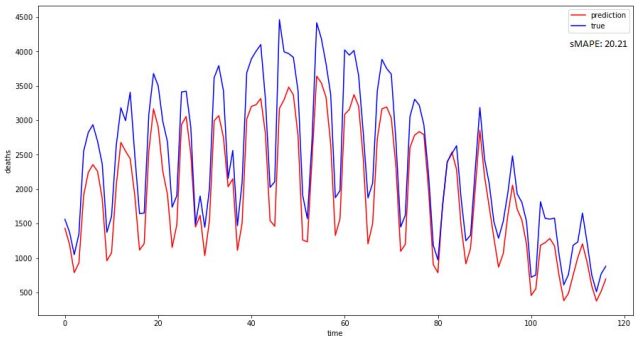

Fig. 4. GCN-LSTM forecast for the 1^{st} horizon on state-level data. (December 7^{th}, 2020 to April 2^{nd}, 2021)

Table II shows the forecast results on the state-level data. For each state in the US, we make a 14-day ahead rolling window forecast with horizons $1 \le h \le 14$. Then, we aggregate the state-level forecast into a forecast for the national

[5]https://github.com/yogeshchaudhari/COVID-19-Forecasting

TABLE I
MULTI-HORIZON (h) ROLLING FORECASTS FOR THE UNITED STATES (NATIONAL LEVEL DATA): COMPETITIVE MODELS (SMAPE).

Horizon	RW	SARIMAX	MCP	VAR	SARIMA	AUG-NN	Wavelet-ANN	DeepAR	NN	LSTM	GRU
$h=1$	38.06	15.75	18.75	19.90	17.12	20.34	21.20	22.77	24.94	32.70	35.61
$h=2$	54.10	16.64	18.86	19.09	18.44	23.06	22.79	22.67	26.79	29.87	35.19
$h=3$	60.66	16.56	19.11	19.19	18.69	21.82	23.16	22.83	26.68	30.50	36.53
$h=4$	59.86	16.49	19.20	19.32	19.23	21.91	23.82	23.76	27.87	29.50	36.21
$h=5$	53.89	16.67	19.11	19.26	20.05	22.17	24.56	25.16	29.40	33.04	35.43
$h=6$	38.70	16.89	19.13	19.21	19.90	22.06	24.32	25.18	29.09	34.25	35.89
$h=7$	21.71	16.64	19.05	19.35	20.96	22.03	25.19	26.54	28.39	35.76	38.58
$h=8$	40.51	17.96	22.09	22.99	24.00	24.49	27.81	29.50	31.18	38.23	39.19
$h=9$	54.93	18.40	22.12	22.51	24.67	24.63	28.48	30.51	34.06	36.33	39.60
$h=10$	59.96	18.02	22.06	22.34	25.17	24.84	28.84	32.27	32.32	35.78	40.31
$h=11$	59.42	18.15	22.49	22.57	26.83	24.94	28.82	34.52	33.41	35.64	40.77
$h=12$	53.29	18.42	22.68	22.34	27.80	25.78	29.70	33.97	34.03	37.38	41.60
$h=13$	40.90	18.54	22.50	22.47	28.48	25.27	28.90	35.32	35.18	38.31	43.34
$h=14$	29.08	18.45	22.41	22.88	28.94	26.89	31.14	35.98	34.85	41.55	45.74
Average	47.51	17.40	20.68	20.96	22.88	23.59	26.34	28.64	30.59	34.92	38.86

TABLE II
MULTI-HORIZON (h) ROLLING FORECASTS FOR THE UNITED STATES (STATE LEVEL DATA): COMPETITIVE MODELS (SMAPE).

Horizon	RW	GCN	SARIMAX	MCP	SARIMA
$h=1$	27.53	20.21	16.80	19.12	17.85
$h=2$	43.88	7.91	17.62	19.30	20.29
$h=3$	50.64	9.72	18.41	19.35	19.45
$h=4$	50.14	17.82	18.47	19.68	21.35
$h=5$	43.18	8.77	19.13	19.33	22.01
$h=6$	28.53	21.64	19.25	19.18	20.82
$h=7$	18.31	16.77	19.88	19.45	21.13
$h=8$	31.01	9.41	23.05	26.34	27.90
$h=9$	43.37	12.88	24.44	26.97	29.79
$h=10$	49.34	9.14	25.49	27.21	27.90
$h=11$	49.61	17.57	25.78	27.59	30.25
$h=12$	42.98	18.68	26.68	27.21	32.10
$h=13$	31.26	23.58	27.44	27.47	32.62
$h=14$	24.99	21.60	28.08	27.37	32.57
Average	38.20	15.41	22.18	23.26	25.43

TABLE III
MULTI-HORIZON (h) ROLLING FORECASTS FOR THE UNITED STATES (NATIONAL LEVEL DATA): COMPETITIVE MODELS (MAE).

Horizon	SARIMAX	Wavelet-ANN	AUG-NN
$h=1$	171.00	209.18	236.25
$h=2$	183.51	238.08	269.02
$h=3$	185.83	261.72	272.65
$h=4$	186.06	273.84	280.05
$h=5$	187.75	283.04	292.21
$h=6$	188.53	280.04	295.96
$h=7$	190.87	288.95	287.83
$h=8$	202.68	313.74	308.60
$h=9$	211.07	319.35	319.13
$h=10$	215.91	324.82	317.65
$h=11$	216.35	322.31	311.53
$h=12$	221.20	328.26	321.83
$h=13$	225.07	327.10	328.89
$h=14$	222.01	351.49	350.34
Average	200.56	294.42	299.42

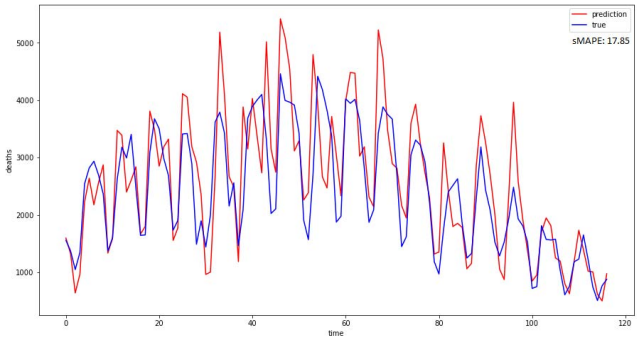

Fig. 5. SARIMA forecast for the 1^{st} horizon on state-level data. (December 7^{th}, 2020 to April 2^{nd}, 2021)

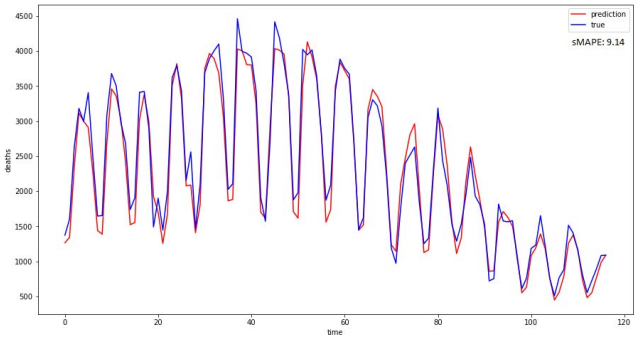

Fig. 6. GCN-LSTM forecast for the 10^{th} horizon on state-level data. (December 16^{th}, 2020 to April 11^{th}, 2021)

level daily deaths. The sMAPE scores on this aggregated forecast are then used to compare all models. GCN-LSTM model performs the best in this experiment. Figures 4 and 5 compare the forecasts of GCN-LSTM model and SARIMA model for the first horizon, while figures 6 and 7 compare the forecasts for the tenth horizon. Although the GCN-LSTM model performs worse than SARIMA on the first horizon, the model was able to better capture the fluctuations in the data.

Table III compares the performances of SARIMAX, Wavelet-ANN, and AUG-NN using the Mean Absolute Error

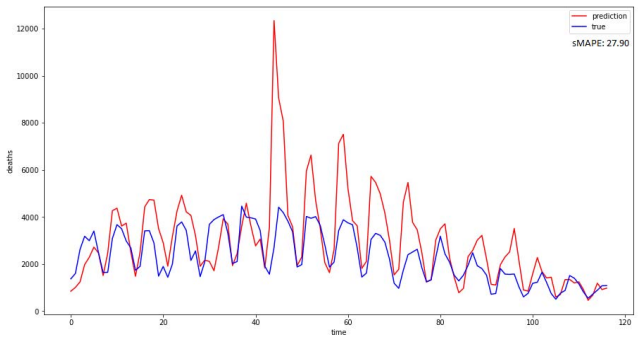

Fig. 7. SARIMA forecast for the 10^{th} horizon on state-level data. (December 16^{th}, 2020 to April 11^{th}, 2021)

(MAE) metric. When compared with the sMAPE scores in Table I, AUG-NN outperforms Wavelet-ANN when the metric is sMAPE but performs worse than Wavelet-ANN when the metric is MAE.

VII. Conclusion and Future Work

In this paper, we evaluated and compared several statistical and machine learning models to forecast daily deaths due to COVID-19 in the United States, using national and state-level data, which lead to several interesting findings. First, we added a network graph learned from the human morbidity data across the United States to enhance the forecast made by a deep spatio-temporal neural network. Second, we discovered that several medical exogenous variables, like "hospitalized patients" and "ICU patients", contain useful information to boost the performance of statistical time-series modeling. Third, in terms of out-of-sample rolling window prediction, the winner of our model comparison is a classical seasonal auto-regressive integrated moving average model with medical exogenous variables. Besides pursuing deep learning models, reinvigorating classical statistical models with data-specific features may lead to surprising results.

Our work adds to the growing body of epidemic disease modeling with novel approaches to combine the multivariate time-series analysis with human mobility data and exogenous medical variables. We expect the findings in this paper can shed some light on future studies in the epidemic forecast.

References

[1] A. Tichopád, L. Pecen, and V. Sedlák, "Could the new coronavirus have infected humans prior november 2019?," *PLOS ONE*, vol. 16, pp. 1–9, 08 2021.

[2] "Who coronavirus (covid-19) dashboard, https://covid19.who.int/," 2021.

[3] S. He, Y. Peng, and K. Sun, "Seir modeling of the covid-19 and its dynamics," *Nonlinear Dynamics*, vol. 101, pp. 1667–1680, Aug 2020.

[4] Z. Zhao, K. Nehil-Puleo, and Y. Zhao, "How well can we forecast the covid-19 pandemic with curve fitting and recurrent neural networks?," *medRxiv*, 2020.

[5] W. Langel, "Extrapolation of infection data for the covid-19 virus and estimate of the pandemic time scale," *medRxiv*, 2020.

[6] P. Furtado, "Epidemiology sir with regression, arima, and prophet in forecasting covid-19," in *Engineering Proceedings*, vol. 5, p. 52, Multidisciplinary Digital Publishing Institute, 2021.

[7] E. Y. Cramer, V. K. Lopez, J. Niemi, G. E. George, J. C. Cegan, I. D. Dettwiller, W. P. England, M. W. Farthing, R. H. Hunter, B. Lafferty, *et al.*, "Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the us," *medRxiv*, 2021.

[8] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *Infectious Diseases*, vol. 20, 2020.

[9] T. R. L. at UMass-Amherst and, *Ground truth data*. Available at https://github.com/reichlab/covid19-forecast-hub/blob/master/data-processed/README.md#ground-truth-data.

[10] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, B. M. Joe Hasell, D. Beltekian, and M. Roser, "Coronavirus pandemic (covid-19)," *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[11] Y. Kang, S. Gao, Y. Liang, M. Li, J. Rao, and J. Kruse, "Multiscale dynamic human mobility flow dataset in the us during the covid-19 epidemic," *Scientific data*, vol. 7, no. 1, pp. 1–13, 2020.

[12] G. Barmparis and G. Tsironis, "Estimating the infection horizon of covid-19 in eight countries with a data-driven approach," *Chaos, Solitons & Fractals*, vol. 135, p. 109842, Jun 2020.

[13] N. Picchiotti, M. Salvioli, E. Zanardini, and F. Missale, "Covid-19 pandemic: a mobility-dependent seir model with undetected cases in italy, europe and us," *arXiv preprint arXiv:2005.08882*, 2020.

[14] Y. Nishimoto and K. Inoue, "Curve-fitting approach for covid-19 data and its physical background," *medRxiv*, 2020.

[15] M. S. Shamil, F. Farheen, N. Ibtehaz, I. M. Khan, and M. S. Rahman, "An agent-based modeling of covid-19: Validation, analysis, and recommendations," *Cognitive Computation*, Feb 2021.

[16] S. Fazeli, B. Moatamed, and M. Sarrafzadeh, "Statistical analytics and regional representation learning for covid-19 pandemic understanding," 2020.

[17] I. Y. Javeri, M. Toutiaee, I. B. Arpinar, T. W. Miller, and J. A. Miller, "Improving neural networks for time series forecasting using data augmentation and automl," *arXiv preprint arXiv:2103.01992*, 2021.

[18] D. Salinas, V. Flunkert, and J. Gasthaus, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," 2019.

[19] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[20] D. Matsunaga, T. Suzumura, and T. Takahashi, "Exploring graph neural networks for stock market predictions with rolling window analysis," 2019.

[21] A. Kapoor, X. Ben, L. Liu, B. Perozzi, M. Barnes, M. Blais, and S. O'Banion, "Examining covid-19 forecasting using spatio-temporal graph neural networks," *arXiv preprint arXiv:2007.03113*, 2020.

[22] N. S. Arunraj, D. Ahrens, and M. Fernandes, "Application of sarimax model to forecast daily sales in food retail industry," *International Journal of Operations Research and Information Systems (IJORIS)*, vol. 7, no. 2, pp. 1–21, 2016.

[23] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[24] C.-H. Zhang *et al.*, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[25] E. Zivot and J. Wang, "Vector autoregressive models for multivariate time series," *Modeling financial time series with S-PLUS®*, pp. 385–429, 2006.

[26] N. Thi Ngoc Anh, N. Quang Dat, N. Thi Van, N. Ngoc Doanh, and N. Le An, "Wavelet-artificial neural network model for water level forecasting," in *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*, pp. 1–6, 2018.

[27] A. H. Nury, K. Hasan, and M. J. B. Alam, "Comparative study of wavelet-arima and wavelet-ann models for temperature time series data in northeastern bangladesh," *Journal of King Saud University - Science*, vol. 29, no. 1, pp. 47–61, 2017.

[28] R. Fu, Z. Zhang, and L. Li, "Using lstm and gru neural network methods for traffic flow prediction," in *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 324–328, 2016.

[29] D. Aldous and J. Fill, "Reversible markov chains and random walks on graphs, 2002," *Unfinished monograph, recompiled*, vol. 2002, 2014.

[30] C. Data61, "Stellargraph machine learning library." https://github.com/stellargraph/stellargraph, 2018.