Available online at www.SciMedJournal.org

SciMedicine Journal

(ISSN: 2704-9833)

Vol. 3, No. 2, June, 2021

Prediction of COVID-19 Confirmed Cases after Vaccination: Based on Statistical and Deep Learning Models

Meejoung Kim [1*]

[1] *Research Institute for Information and Communication Technology, Korea University, Seoul, Rep. of Korea.*

Received 08 February 2021; Revised 27 March 2021; Accepted 03 April 2021; Published 01 June 2021

Abstract

In this paper, we analyze and predict the number of daily confirmed cases of coronavirus (COVID-19) based on two statistical models and a deep learning (DL) model; the autoregressive integrated moving average (ARIMA), the generalized autoregressive conditional heteroscedasticity (GARCH), and the stacked long short-term memory deep neural network (LSTM DNN). We find the orders of the statistical models by the autocorrelation function and the partial autocorrelation function, and the hyperparameters of the DL model, such as the numbers of LSTM cells and blocks of a cell, by the exhaustive search. Ten datasets are used in the experiment; nine countries and the world datasets, from Dec. 31, 2019, to Feb. 22, 2021, provided by the WHO. We investigate the effects of data size and vaccination on performance. Numerical results show that performance depends on the used data's dates and vaccination. It also shows that the prediction by the LSTM DNN is better than those of the two statistical models. Based on the experimental results, the percentage improvements of LSTM DNN are up to 88.54% (86.63%) and 90.15% (87.74%) compared to ARIMA and GARCH, respectively, in mean absolute error (root mean squared error). While the performances of ARIMA and GARCH are varying according to the datasets. The obtained results may provide a criterion for the performance ranges and prediction accuracy of the COVID-19 daily confirmed cases.

Keywords: Covid-19; Predictive Model; Non-linear Fitting; Long Short-Term Memory Deep Neural Network; Autoregressive Integrated Moving Average; Generalized Autoregressive Conditional Heteroscedasticity.

1. Introduction

The coronavirus outbreak in Wuhan, China, in December 2019, and named COVID-19 by the World Health Organization (WHO), made 2020 the year of global disaster [1]. Since the first death from the disease was reported in January 2020, the numbers of confirmed cases and death cases have continuously increased until the vaccination began on Dec. 8th in the United Kingdom (UK). Currently, the numbers of confirmed cases are declining in countries where vaccinations have begun, such as the USA and the UK, while the numbers are still increasing or fluctuating in other countries where vaccinations have started late or have not yet begun.

Symptomatic treatment and supportive therapy are used to cure the COVID-19 patients. It includes basic disease treatment, symptom relief, effective protective and supportive treatment of internal organs, active prevention and treatment of complications, and respiratory support if necessary. Researchers are working on the development of treatments for the disease, and countries are supporting it. As a result, several types of vaccines have developed. However, the drugs that can cure the disease have not yet developed.

* Corresponding author: meejkim@korea.ac.kr

http://dx.doi.org/10.28991/SciMedJ-2021-0302-7

After the outbreak of the disease, each country does its best to protect its people. For example, various policies are in place, such as limiting people gathering events, restricting overseas travel, and quarantining people from abroad to prevent the influx of corona from abroad. Nevertheless, the increasing trend, rate of increment, and the number of confirmed cases vary from country to country depending on several factors, such as culture, policy, health care, and social habits. Imran et al. (2020) [3] analyzed the reaction of people from different cultures to the COVID-19 and sentiment about their subsequent actions. The authors applied the deep long short-term memory (LSTM) to the extracted tweets. The damage caused by the disease is expected to appear in all fields, including the economy, society, culture, and emotions of individuals, in the next few years worldwide.

There are case studies for COVID-19 [4–16], especially for Wuhan, China [4–8]. Various models for the disease have been studied in 2020, including mathematical models [7–19], statistical modeling [20–22], and artificial intelligence (AI) models [15, 22–24]. Details of the studies are presented in section 2.

In this paper, we analyze the COVID-19 based on the number of daily confirmed cases. As the data is time series, we consider time series models of statistics and deep learning (DL) technology to predict the number of daily confirmed cases; ARIMA, generalized autoregressive conditional heteroscedasticity (GARCH), and stacked LSTM deep neural network (LSTM DNN). The prediction procedure of models consists of three parts; preprocessing process, training process, and prediction process. The min-max transformation is used to preprocess the datasets. The autocorrelation function (ACF) and the partial autocorrelation function (PACF) are used to find the orders of the statistical models, while the sub-optimal hyperparameters of the DL model, such as the number of LSTM cells and the number of blocks in an LSTM cell, are found exhaustively. Data from Dec. 31, 2019, to Feb. 22, 2020, provided by the WHO [2], is used in the experiment. The models are applied to ten datasets of daily confirmed cases; nine countries across the continents and the world. Datasets of two sizes are used in the experiment to investigate the effects of data size and vaccination on performance. Numerical results show the effects and show the stacked LSTM DNN outperforms the statistical models.

The motivation is as follows: 1) Can the statistical models and the DL techniques provide acceptable predictive performances for the new disease before and after vaccination? 2) Which model can predict the disease best? Several articles have considered predicting the disease using statistical models and ML models [e.g. 15, 20, 21–23]. Since these studies used short-term data of the disease, it was insufficient for learning. The number of confirmed cases continues to increase as the disease spread. However, it tends to decline in countries where vaccination has begun. Therefore, it is meaningful to apply the models with the larger datasets and investigate the effect of vaccination on performance. Besides, this is the first study to apply the GARCH model to the dataset of COVID-19.

This study includes: Section 2 describes the predictive models and procedures for daily confirmed cases. Section 3 and Section 4 present performance measurements and the experimental results, respectively, and Section 5 provides a conclusion.

2. Related Works: Modeling the Disease

The case studies for COVID-19 can be found in [4–16], especially for Wuhan, China [4–8]. For example, Li et al. (2020) [4] described the characteristic of positive cases, the distributions of epidemiological time delay, the disease doubling time, and the breeding number, based on the data of Wuhan. Lin et al. (2020) [5] proposed conceptual models for the disease based on the individual behavioral reaction and governmental actions, while Prem et al. (2020) [6] considered synthetic location-specific contact patterns to estimate the effects of physical distance measures on the progression of the COVID-19 epidemic.

There are a lot of studies that investigate the models for the COVID-19 in 2020. The used methods for the model include the mathematical models [7–19], statistical modeling [20–22], and artificial intelligence (AI) models [15, 22–24]. A compartmental mathematical model was proposed as a spreading model of the disease, emphasizing the potential for transmission of super-spreaders individuals, in Ndairou et al. (2020) [7]. It studied the threshold of reproduction number, local stability of disease-free equilibrium using the number, and the model's sensitivity for parameters. Kucharski et al. (2020) [8] considered a stochastic transmission model to estimate transmission variation over time. The probability of newly confirmed cases that generate outbreaks in other areas was calculated based on the estimation. Zhao et al. (2020) [9] estimated the reproduction number in the early stage of the disease through the curve of confirmed cases in China. The reproduction number was also estimated in Shen et al. (2020) [10] study through a dynamic model, based on Chinese data, from which the epidemic peak time and size were predicted. A new epidemic model that can explain the impact of health care capacity was proposed in Cakan (2020) [11]. In the model, local stability and global stability were studied. Wu et al. (2020) [12] introduced the susceptible-exposed-infectious recovered (SEIR) model to simulate the Wuhan epidemic. The authors estimated the spread of the disease nationally as well as globally by the model. Shah et al. (2020) [13] proposed a generalized SEIR model for the disease, in which the behavior of transmission of the disease was investigated under different control strategies. In the model, the authors considered transmissions between humans and formulated the reproduction number to analyze transmission

dynamics of coronavirus outbreak. Intissar (2020) [14] reinvestigated the SEIR model in Shah et al. (2020) [13] work. They considered the local and global stability conditions by using a reproduction number and added some control parameters to force the trajectories to go to the equilibria in the five-dimensional Covid-19 system. Zheng et al. (2020) [15] proposed an improved susceptible-infected model to estimate the variety of infection rates for analyzing the transmission laws and development trend. The model contains the natural language processing (NLP) module and the LSTM. Fanelli & Piazza (2020) [16] analyzed the temporal dynamics of disease outbreaks in China, Italy, and France. It indicated the universality of epidemic spreading based on the analysis of simple day-lag maps and proposed simple mean-field models to collect a quantitative picture of the epidemic spreading. Choi & Ki (2020) [17] considered the transmission model, the reproduction number, and the effectiveness of preventive measures of the disease that fits S. Korea through the number of confirmed cases of S. Korea. Ivorra et al. (2020) [18] proposed the disease spread mathematical model and investigated the detected portion among all infected cases. Chen et al. (2020) [19] developed a simplified transmission network model for the disease by stimulating the potential transmission from the infection source to the human infection, and then computed the reproduction number based on the model. Roy et al. (2020) [20] predicted epidemiological patterns of prevalence and incidence of the disease with ARIMA, using cumulative confirmed cases of the disease in Indian states. A hybrid methodology, wavelet-autoregressive integrated moving average (W-ARIMA), was proposed in Singh et al. (2020) [21]. They used the number of daily deaths of five countries to validate their method, estimated one month-ahead prediction of death cases, and compared its performance with ARIMA. Singh et al. (2020) [22] considered ARIMA and least square support vector machine (LS-SVM) to predict confirmed cases. The data consisting of daily confirmed cases of SARS-CoV-2 in the most affected five countries was used for modeling and predicting one-month confirmed cases of this disease. Shahid et al. (2020) [23] predicted confirmed cases, deaths cases, and recoveries cases of the disease through ARIMA, support vector regression (SVR), LSTM, and bidirectional LSTM. The study used datasets of ten countries. For the early-stage treatment of the disease, the analysis of chest X-rays of infected patients was a crucial step. A model, based on an Auxiliary Classier Generative Adversarial Network (ACGAN), was developed to generate image data in Waheedi et al. (2020) [24].

There are several survey articles on the disease [25–27]. Latif et al. (2020) [25] surveyed various research activities on the disease, including statistical and artificial intelligence (AI) modelings and data visualization, which can be used in data management, such as storing, processing, training, predicting, and insight extracting. Emphasizing the importance of responding to the COVID-19 outbreak and preventing the severe effects of the disease pandemic, Pham et al. (2020) [26] overviewed AI and big data in various areas, identified the applications aimed at fighting against COVID-19, highlighted challenges and issues associated with state-of-the-art solutions, and recommended for effective control of the COVID-19 situation. Chamola et al. (2020) [27] investigated the key aspects of the disease, focusing on its impact on the global economy, and considered the use of technologies, including the internet of things (IoT) and AI, to mitigate the outbreak of disease.

3. Materials and Methods

Since the number of daily confirmed cases of COVID-19 is a time series, we denote it as a process $\{X_t\}_{t\in\mathbb{Z}}$. For the analysis of time series, the statistical models, such as ARIMA and GARCH, have been considered traditionally, and the NN models, such as multi-layer perceptron (MLP) and LSTM recurrent neural network (RNN), have been used recently (e.g. Kim (2020 and 2021) [28, 29]). In this study, we consider two statistical models and one DL model. We define terms and explain models and prediction procedures of the models in this section. Firstly, we define *n*-step ahead prediction as follows:

Definition 1. The *n-step ahead prediction* of $\{X_t\}_{t\in\mathbb{Z}}$ is given by $\mathrm{E}(X_{t+n} \mid \mathcal{F}_t), n \in \mathbb{N}$, the conditional expectation of X_{t+n} given that $\mathcal{F}_t$ is known. Here $\mathcal{F}_t$ is the entire history up to time t generated by $\{X_s : s \le t\}$.

3.1. Predictive Models and Prediction Procedures of COVID-19: ARIMA and GARCH

Definition 2. A time series ε_t is said to be *white noise* if the expectation, variance, and the auto-covariance function of ε_t are given by $E\varepsilon_t = 0$, $Var(\varepsilon_t) < \infty$, and $\gamma(h) = Cov(\varepsilon_t, \varepsilon_{t+h}) = 0$, respectively, for all $h \neq 0$. Here $\gamma(h)$ is given by $E(\varepsilon_t\varepsilon_{t+h}) - E(\varepsilon_t)E(\varepsilon_{t+h})$.

If a time series ε_t is white noise**,** we denote it by $\{\varepsilon_t\} \sim WN(0, \gamma(0))$.

Definition 3. The process $\{X_t\}_{t\in\mathbb{Z}}$ is called *ARMA(p,q)* if it satisfies

$$X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j}, \tag{1}$$

where $\{\varepsilon_t\} \sim WN(0, \sigma^2), \sigma > 0$, and α_i and β_i are some constants.

Definition 4. The process $\{X_t\}_{t\in\mathbb{Z}}$ is called *ARIMA(p,1,q)* if $\Delta X_t = X_t - X_{t-1}$ is stationary and invertible ARMA(*p*,*q*). The process $\{X_t\}_{t\in\mathbb{Z}}$ is called *ARIMA(p,d,q)* if $\Delta^d X_t$ is stationary and invertible ARMA(*p*,*q*).

In the definition 4, Δ^d is the operator defined by $(1-B)^d$, where *B* is a back-shift operator given by $BX_t = X_{t-1}$. For example, $\Delta^2 X_t = (1-B)^2 X_t = X_t - 2X_{t-1} + X_{t-2}$. Definitions of 'stationary' and 'invertible' can be found in Shumway & Stoffer (2017) [30] research.

Definition 5. The process $\{X_t\}_{t\in\mathbb{Z}}$ is called *GARCH(p,q)* if it satisfies the following conditions:

$$\begin{cases} (i)\ X_t = \sigma_t \varepsilon_t, \varepsilon_t \sim IID(0,1), \text{independent of } \sigma_t \\ (ii)\ \sigma_t^2 = \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \beta_0 + \sum_{j=1}^{q} \beta_j X_{t-j}^2 \end{cases}, \tag{2}$$

where $\beta_0 > 0,\ \alpha_i \geq 0, \beta_j \geq 0, 1 \leq i \leq p, 1 \leq j \leq q$. $IID(0,1)$ in Equation 2 means that ε_t follows independent identical distribution with mean zero and variance one, for all *t*. The process such that $\Delta^d X_t$ is GARCH(*p*,*q*), we call it GRACH(*p*,*d*,*q*).

The statistical methods dealing with time series require testing the stationary property of data in advance. If a given data is non-stationary, take the stationary test to the first differenced dataset of the data, $\{\Delta X_t\}$. The test repeats until getting the stationary process by increasing *d*. If the test passes, we have to find the model's orders of the process. ARIMA and GARCH generally use the ACF and the PACF to find the orders. The determined orders are used in the training process, and the time series is predicted based on the trained results.

Let $\{X_t\}_{t=1}^{N}$ be a given dataset, where *N* is the number of data. Algorithm 1 describes the prediction procedure of ARIMA (GARCH).

Algorithm 1. Prediction procedure of ARIMA (GARCH)

1. Test the stationarity of data $\{X_t\}_{t=1}^{N}$. If it is non-stationary, take the difference of data $\{\Delta X_t = X_{t+1} - X_t\}_{t=1}^{N-1}$ and test the stationarity of $\{\Delta X_t\}_{t=1}^{N-1}$.
2. Find the orders of ARIMA by ACF and PACF.
3. Training process:
   3.1 Divide $\{\Delta X_t\}_{t=1}^{N-1}$ into two disjoint subsets $\{\Delta X_t\}_{t=1}^{N_1-1}$ and $\{\Delta X_t\}_{t=N_1}^{N-1}$, where $N_1 - 1 = (N-1)\times r$ for initially given training ratio $r \in (0,1)$.
   3.2 Fit the model by using the orders.
4. Prediction process: Generate the prediction of difference $\{\triangle \hat{X}_t\}_{t=N_1}^{N-1}$ by the fitted model. Take the inverse transform of $\{\triangle \hat{X}_t\}_{t=N_1}^{N-1}$ and then obtain the predicted time series $\{\hat{X}_t\}_{t=N_1}^{N-1}$.
5. Compute the performance measures with $\{X_t\}_{t=N_1+1}^{N}$ and $\{\hat{X}_t\}_{t=N_1}^{N-1}$.

We will explain Algorithm 1 based on the dataset of daily confirmed cases. Figure 1 shows the number of daily confirmed cases of the world and nine countries over 420 days from Dec. 31, 2019, to Feb. 22, 2021. Then, X_t is the number of confirmed cases at time *t* and $N = 420$ (step 1). In China, X_t was suddenly increased at the beginning of the outbreak and then declined after sixty days of the first outbreak. X_t s for other countries tended to increase with fluctuations. After vaccination begins in the UK and the USA, X_t s for both countries tended to decline. All the datasets considered are checked non-stationary, and the datasets consisting of the first or the second difference between daily confirmed cases turned out to be stationary (step 1). Figure 2 illustrates the first difference ΔX_t (step 1) of the world dataset, while Figure 3 illustrates the ACF and PACF with 60 lags for the set (step 2). Based on Figure 3, (1,0) or (5,0) can be considered as (*p*, *q*) with $d = 1$ in ARIMA(*p*,*d*,*q*) model for the dataset. The ARIMA model is fitted based on the training set (step 3), and then future values for the difference of time series are generated, which is $\{\triangle \hat{X}_t\}_{t=N_1}^{N-1}$ (step 4).

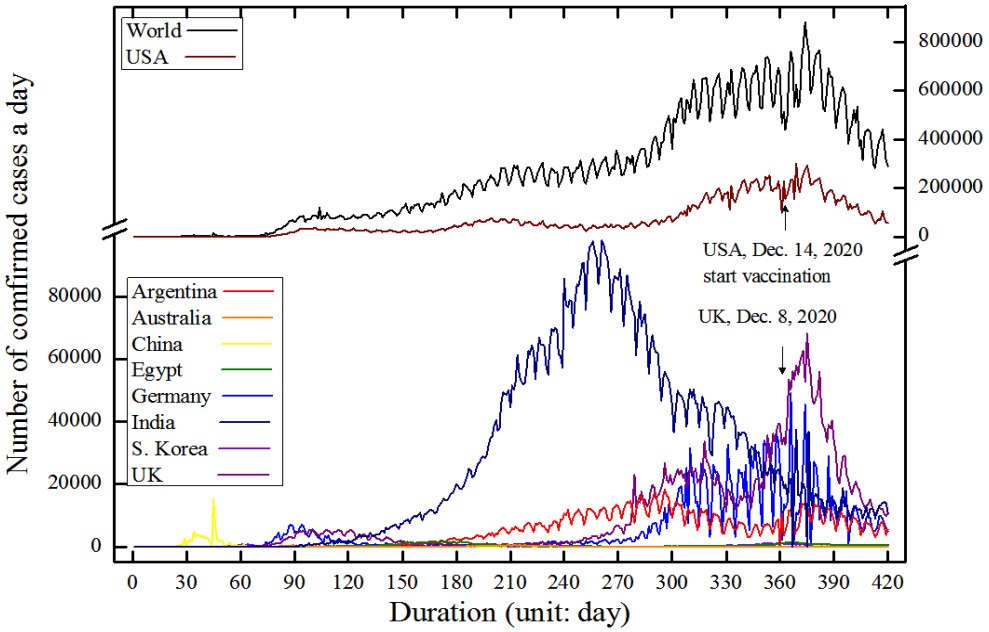


Figure 1. The numbers of daily confirmed cases of COVID-19: World and nine countries for 420 days

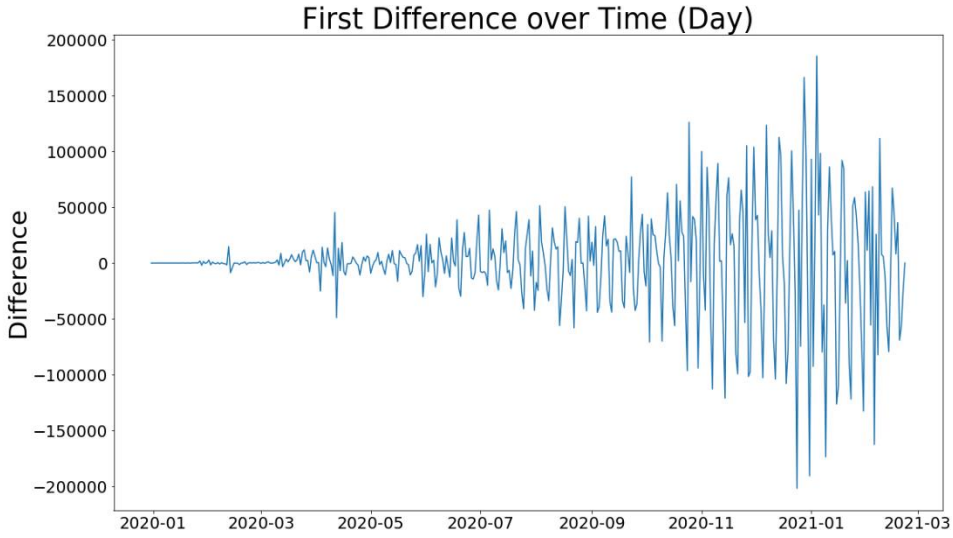


Figure 2. The first difference between daily confirmed cases of COVID-19: World

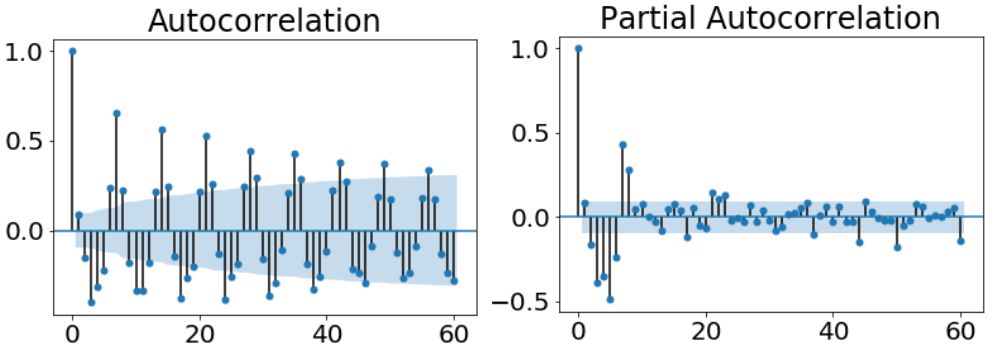


Figure 3. ACF and PACF for daily confirmed cases with 60 lags: World

3.2. Predictive Models and Prediction Procedures of COVID-19: Stacked LSTM DNN

LSTM is a model that considers the vanishing gradient problem in an RNN, which is dealing with time series. Figure 4 illustrates the structure of an LSTM cell.

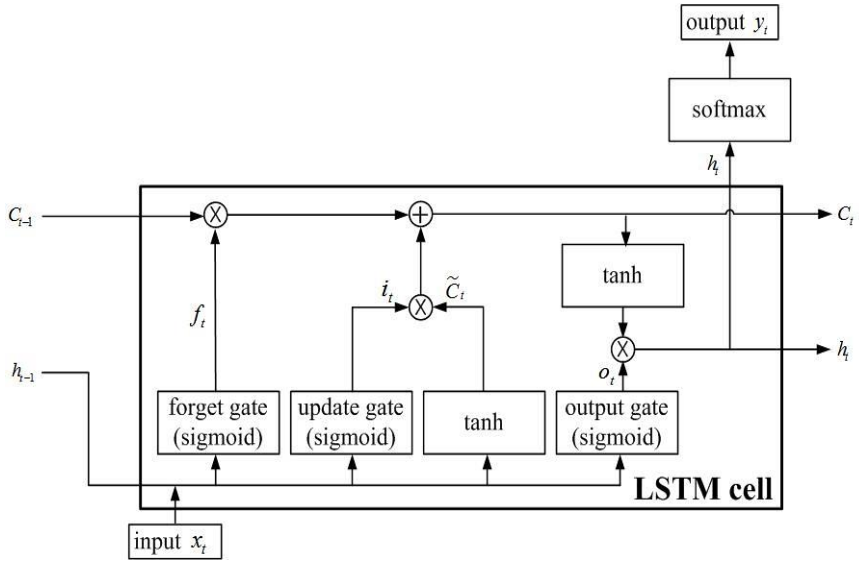


Figure 4. The structure of an LSTM cell

The functions of gates in Figure 4 are given by:

$$\begin{aligned} &f_t = S(\mathbf{W}_f \cdot [h_{t-1}, X_t] + \mathbf{b}_f), \quad i_t = S(\mathbf{W}_i \cdot [h_{t-1}, X_t] + \mathbf{b}_i), \quad C_t = \tanh(\mathbf{W}_C \cdot [h_{t-1}, X_t] + \mathbf{b}_C), \\ &o_t = S(\mathbf{W}_o \cdot [h_{t-1}, X_t] + \mathbf{b}_o), \; C_t = f_t \otimes C_{t-1} \oplus i_t \otimes C_t, \text{ and } h_t = o_t \otimes \tanh(C_t), \end{aligned} \tag{3}$$

where S and tanh are the sigmoid function and hyperbolic tangent function defined by:

$$S(x) = \frac{1}{1+e^{-x}} \quad \text{and} \quad \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}, \tag{4}$$

respectively. $\mathbf{W}_\bullet$ and $\mathbf{b}_\bullet$ are the weight matrix and bias vector for gate $\bullet$, respectively, and determined during the training process in such a way as to minimize the cost function. $\otimes$ and $\oplus$ represent entry-wise multiplication (Hadamard product) and entry-wise addition (direct sum), respectively. The subscripts of the matrices in Equation 3 only denote the gates. We deleted the notations describing layer and iteration number in the matrices for notational simplicity. A detailed explanation of LSTM and the explicit formulas of the functions in Equation 3 can be found in several articles, including Azzouni et al. (2017) [31].

LSTM cells can be stacked in a network to enhance prediction accuracy. Figure 5 illustrates the structure of a stacked LSTM DNN. All of the LSTM cells in Figure 5 have the same structure as shown in Figure 4. The dataset inputs to LSTM(1) and the output of LSTM(1), h_1, is passed to LSTM(2), and the same procedure is performed up to the kth LSTM stack. At the end of the LSTM cell, h_k converts to the output via softmax. Algorithm 2 explains the prediction procedure of LSTM DNN.

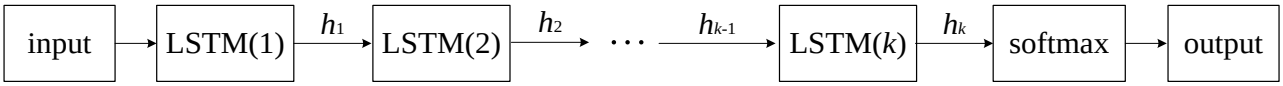


Figure 5. The structure of the stacked LSTM DNN

Algorithm 2. Prediction procedure of LSTM DNN

1. Preprocess dataset $\{X_t\}_{t=1}^{N}$: Let $\{X_{t,\text{pre}}\}_{t=1}^{N}$ be the preprocessed dataset.

2. Divide the preprocessed dataset into training and test sets: $\{X_{t,\text{pre}}\}_{t=1}^{N} = \{\{X_{t,\text{pre,tr}}\}_{t=1}^{N_1}, \{X_{t,\text{pre,te}}\}_{t=N_1+1}^{N}\}$, where $N_1 = N \times r$ for initially given training ratio $r \in (0,1)$.

3. Train LSTM DNN by $\{X_{t,\text{pre,tr}}\}_{t=1}^{N_1}$ to minimize the cost function.

4. Predict $\{X_{t,\text{pre,te}}\}_{t=N_1+1}^{N}$ by the trained network.

5. Compute the performance measures.

In step 1, the dataset is pre-processed by min-max transformation (MMT) defined by:

$$X_{t,\text{pre}} = MMT(X_t) = \frac{X_t - \min\limits_{1\le t\le N}(X_t)}{\max\limits_{1\le t\le N}(X_t) - \min\limits_{1\le t\le N}(X_t)}. \tag{5}$$

The mean squared error is used as the cost function in step 3. Table 1 summarizes the notation.

Table 1. Summary of notations

Notation	Description	Notation	Description
X_t	Raw data	p, q	Orders of ARIMA & GARCH
$X_{t,\text{pre}}$	Min-max transformed data	$\hat{X}_t$	1-step ahead prediction of X_t
$\mathcal{F}_t$	History up to t	$\alpha_i, \beta_j, \varepsilon_t$	Parameters in statistical models
$\mathbf{W}_\bullet$	Weight matrix of gate $\bullet$	S	Sigmoid function
$\mathbf{b}_\bullet$	Bias vector of gate $\bullet$	r	Training ratio

3.3. Performance Measures

To measure the accuracy of prediction, we consider the mean absolute error (MAE), root mean squared error (RMSE), normalized mean absolute error (NMAE), and the normalized mean squared error (NMSE), which are defined by:

$$MAE = \frac{1}{N-N_1}\sum_{j=N_1+1}^{N} |X_j - \hat{X}_j|, RMSE = MSE^{1/2}, NMAE = \frac{1}{N-N_1}\sum_{j=N_1+1}^{N}\frac{|X_j - \hat{X}_j|}{X_j}, \text{and } NMSE = \frac{MSE}{M_{\text{test}}}, \tag{6}$$

respectively, where MSE and M_{test} is are given by $MSE = \sum_{j=N_1+1}^{N}(X_j - \hat{X}_j)^2/(N-N_1)$ and $M_{\text{test}} = \sum_{j=N_1+1}^{N} X_j/(N-N_1)$, respectively.

We define another measure, percentage improvement (PI), to compare the performance of two models as follows:

$$PI(M_1, M_2) = \left\{1 - \frac{\mu(M_2)}{\mu(M_1)}\right\} \times 100, \tag{7}$$

where $\mu(M)$ is the measured performance of model M, such as MAE and RMSE. $PI(M_1, M_2) > 0$ $(PI(M_1, M_2) < 0)$ implies that model M_1 (M_2) improved M_2 (M_1) by $PI(M_1, M_2)$ % for the concerned measure.

4. Results and Discussion

4.1. Experimental Setting

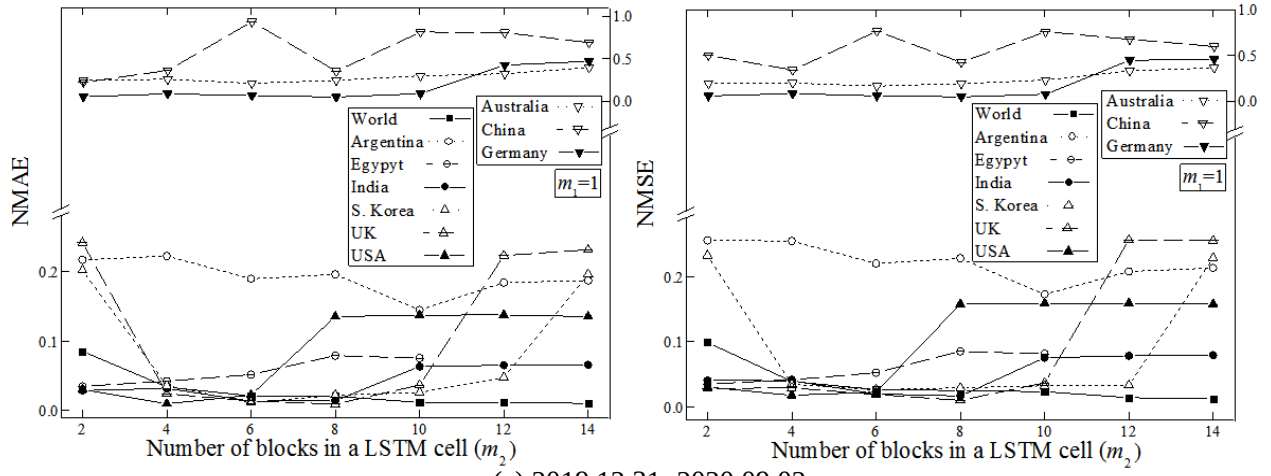
We conducted the experiments using Python 3.6 and Tensorflow v.1.7.0 on an Intel Core i7, 16 GB RAM. The data used is the daily confirmed cases of COVID-19 for 420 days from Dec. 31, 2019, to Feb. 22, 2021, extracted from the dataset obtained through GitHub [32], provided by the WHO [2]. The datasets of the world and nine countries shown in Figure 1 are selected. The selection intended to include countries from all continents, where culture and policies on the disease are different. The countries are Argentina, Australia, China, Egypt, Germany, India, S. Korea, the UK, and the USA, and the world dataset is used to investigate the global trend of the disease. All three models are applied to each dataset. Besides the ten datasets, we consider a portion of data from each dataset to investigate the effect of data size on performance. The selected portion is for 247 days from Dec. 31, 2019, to Sept. 2, 2020. From now on, we call this dataset a small-size dataset. It is observed that performances depend on training ratio and $r = 0.8 \sim 0.9$ turns out to provide better performances than other ratios. Therefore, we mainly use $r = 0.8$ in the experiments and use 0.9 for comparison of ratio effect only. We use the sigmoid for activation function, 100 epochs, and one batch size in the LSTM cells. Since a small-future-step predicts better, we consider a 1-step ahead prediction in the experiment.

4.2. Reports of Various Experimental Results

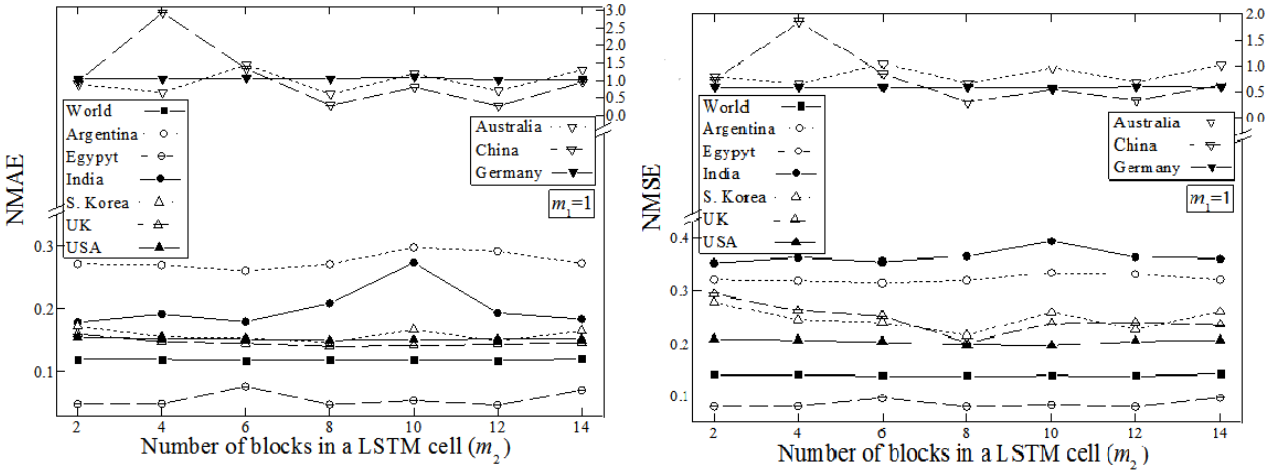
The orders of statistical models are determined by the ACF and the PACF, while the optimal hyperparameters of LSTM DNN, such as the numbers of LSTM cells and blocks of a cell, are obtained through exhaustive search. Let (p,d,q) and (m_1,m_2) be the obtained orders for statistical models and the hyperparameters for LSTM, which minimize

NMAE and NMSE or MAE and RMSE. Here m_1 and m_2 are the numbers of LSTM cells and block in an LSTM cell, respectively. From now on, we call the (p,d,q) and (m_1,m_2) *optimal parameters*. In the experiment, we considered up to five LSTM cells and fifteen blocks in an LSTM cell. We observed that one LSTM cell provides the best NMAE, except for three countries.

Figure 6 shows the values of NMAE and NMSE for varying numbers of blocks with one LSTM cell for the two different sizes of datasets. Figures 6(a) and 6(b) are obtained by the small-size datasets and a total of 470 days datasets, respectively. It shows that the fluctuations of NMAE and NMSE are large for small-size datasets, which seems due to more sensitivity of hyperparameters for small-size datasets. On the other hand, the values of NMAE and NMSE with small-size datasets are less than those with large-size datasets. It seems due to the relatively small changes in the predicted values of the small-size datasets since the increments of confirmed cases of the small-size datasets are small compared to the increments of confirmed cases of the total datasets.



(a) 2019.12.31~2020.09.02



(b) 2019.12.31~2021.02.22

Figure 6. The performance measures values for varying number of blocks with one LSTM cell

Figure 7 shows the daily confirmed cases of four datasets and their predictions of the three models using the optimal parameters. Black and red lines represent raw data and predictions, respectively. In China, many confirmed cases occurred in the early stage of the outbreak, and those seem to affect the model, which appears as the difference between predicted values and actual values. Table 3 will show it explicitly. The increment of confirmed cases in S. Korea was limited in the early stage of the disease. It dues to the policies that identify and disclose the paths of infected people and mandate the mask-wearing. However, it started to increase due to a not predicted public meeting held on August 15, 2020, and the relief of distance policy from the second stage to the first stage from Oct. 12. The distance policy change resulted in group infections through nursing hospitals and church meetings. The number of daily confirmed cases in the USA has soared in the early stage of the disease. It seems due to the culture of wearing masks. The number increased from 646.2 per 1 million people on Dec. 15, 2020, when vaccination began, to 753.3 on Jan. 11, 2021. Since then, it has decreased with slight fluctuation. According to the figure, LSTM DNN is better than the statistic models for the datasets.

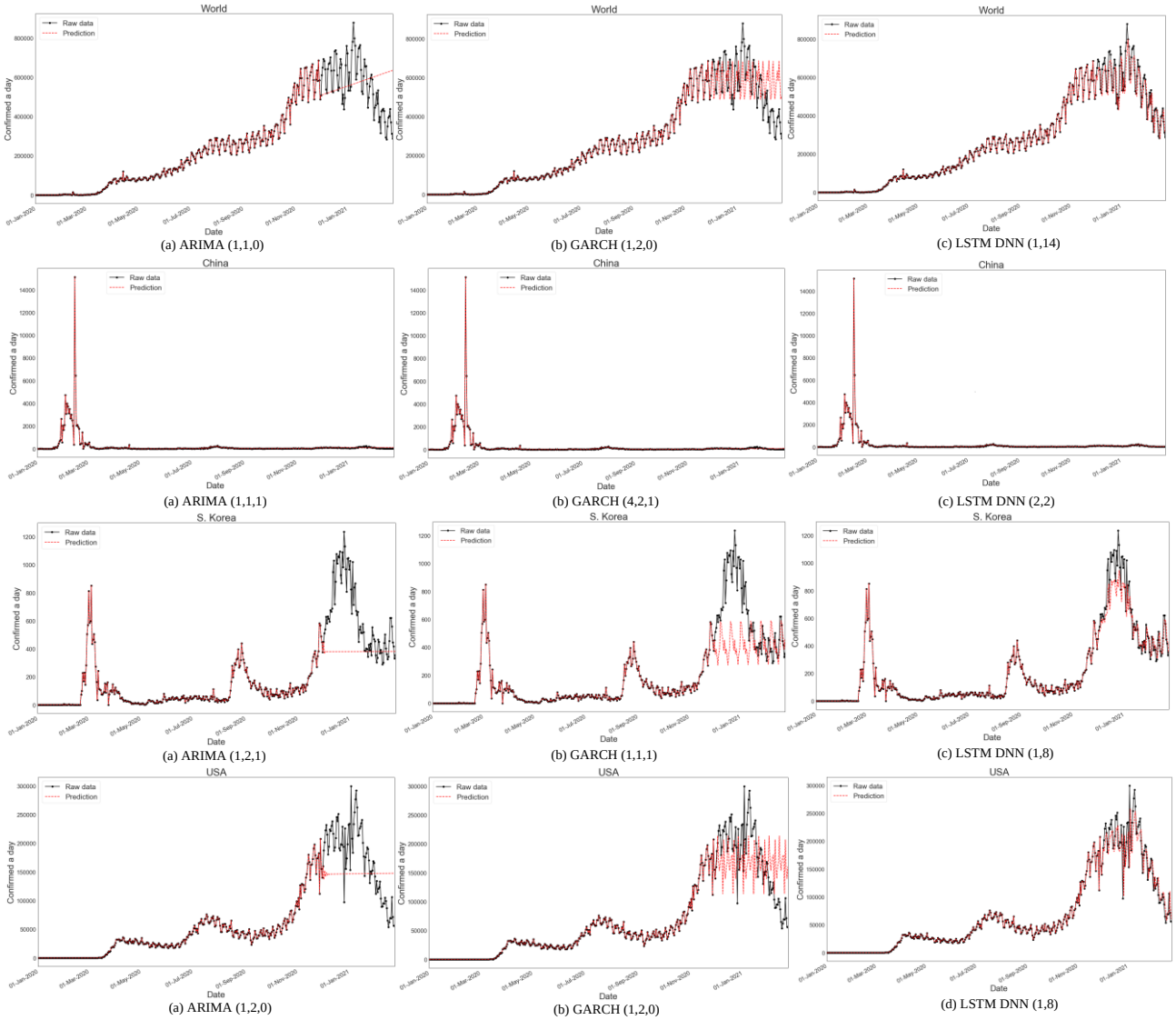


Figure 7. Raw data and 1-step ahead prediction of daily confirmed cases: World and three countries (r = 0.8)

Table 2 presents the corresponding NMAE and NMSE of Figure 7 and those of other countries. The value in the 'Optimal' column is an optimal parameter for the dataset. For example, for China corresponding to Figure 7, NMAE and NMSE for LSTM DNN are obtained by 0.2468 and 0.3190, respectively, and (1,1,1), (4,2,1), and (2,4) are the optimal parameters that provide the best NMAE for ARIMA, GARCH, and LSTM DNN, respectively. We notice one LSTM cell, m_1=1, is optimal for the datasets, except for three countries. The number of blocks in a cell varies depending on the dataset. Table 3 presents MAE and RMSE. We observed (p,d,q) that minimizes NMAE may not minimize MAE. The optimal hyperparameters for MAE of LSTM DNN are (2, 2), different from (2, 4) for NMAE.

Table 2. Comparison of best NMAE and NMSE with optimal parameters (r = 0.8)

Model / Country	ARIMA			GARCH			LSTM DNN		
	Optimal	**NMAE**	**NMSE**	**Optimal**	**NMAE**	**NMSE**	**Optimal**	**NMAE**	**NMSE**
World	(2,1,0)	0.2372	0.2868	(1,2,0)	0.2646	0.2705	(1,14)	**0.1384**	**0.1695**
Argentina	(2,1,0)	0.3227	0.4039	(1,2,0)	0.4330	0.4682	(2,6)	**0.2595**	**0.3131**
Australia	(7,2,1)	0.7063	0.7217	(7,1,1)	0.8375	0.7524	(2,4)	**0.5957**	**0.6390**
China	(1,1,1)	0.9624	0.5824	(4,2,1)	0.8056	0.6098	(2,4)	**0.2468**	**0.3190**
Egypt	(0,1,0)	0.3680	0.5906	(1,1,0)	0.4491	0.6446	(1,12)	**0.0470**	**0.0785**
India	(1,2,0)	0.9402	0.7848	(1,2,0)	1.5623	1.3186	(1,2)	**0.1784**	**0.3512**
Germany	(1,2,0)	**0.6894**	0.9290	(1,2,2)	1.4499	0.8222	(1,10)	1.1084	**0.5849**
S. Korea	(1,2,1)	0.3382	0.5679	(1,1,1)	0.3363	0.5456	(1.8)	**0.1465**	**0.2154**
UK	(1,1,1)	0.4482	0.7285	(8,2,1)	0.4276	0.7096	(1,8)	**0.1401**	**0.2018**
USA	(1,2,0)	0.3993	0.3866	(1,2,0)	0.4232	0.3721	(1,8)	**0.1494**	**0.1971**

Table 3. Comparison of best MAE and RMSE with optimal parameters (r = 0.8)

Model / Country	ARIMA			GARCH			LSTM DNN		
	Optimal	MAE	RMSE	Optimal	MAE	RMSE	Optimal	MAE	RMSE
World	(1,1,0)	135008.05	162471.80	(1,2,0)	132156.17	153008.71	(1,14)	**61783.26**	**76964.19**
Argentina	(7,1,1)	2206.05	2642.88	(1,2,0)	2985.45	3596.55	(2,6)	**1670.65**	**2403.63**
Australia	(7,2,1)	6.24	8.80	(1,1,1)	6.87	9.15	(2,4)	**6.03**	**7.80**
China	(1,1,1)	42.88	54.65	(4,2,1)	45.84	57.22	(2,2)	**19.70**	**29.15**
Egypt	(0,1,0)	331.42	442.05	(1,1,0)	385.81	482.41	(1,12)	**37.98**	**59.10**
India	(1,2,0)	12992.52	14516.73	(1,2,0)	22483.93	24388.98	(1,2)	**3285.32**	**6419.35**
Germany	(1,1,0)	8452.58	11189.64	(1,2,2)	10564.81	13014.75	(1,10)	**6599.46**	**9195.28**
S. Korea	(4,1,1)	240.45	310.12	(1,1,1)	251.64	344.48	(1.8)	**99.24**	**136.29**
UK	(1,1,1)	16155.33	21721.85	(1,1,1)	15456.04	20984.41	(1,8)	**4349.66**	**6057.66**
USA	(1,2,0)	58409.90	66901.96	(1,2,0)	54921.64	64402.18	(1,8)	**24381.41**	**34078.02**

Table 4 presents the performance improvements between models, obtained based on Table 3. It shows that LSTM improves ARIMA and GARCH by 3.36%~88.54% (9.05%~86.63%) and 12.22%~90.15% (14.15%~87.74%), respectively, for MAE (RMSE), while one of ARIMA and GARCH is better depending on data. For example, for the data of Egypt, LSTM improves ARIMA and GARCH by 88.54% (86.63%) and 90.15% (87.74%) for MAE (RMSE), respectively, and GARCH is better than ARIMA for the data of UK and USA.

Table 4. Comparison of percentage improvement (r = 0.8, unit: %)

PI / Country	PI(LSTM,ARIMA)		PI(LSTM, GARCH)		PI(ARIMA, GARCH)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
World	54.23	52.62	53.24	49.69	−2.15	−6.18
Argentina	24.26	9.05	44.04	33.16	26.10	26.51
Australia	3.36	11.36	12.22	14.15	9.17	3.82
China	54.05	44.66	57.02	49.05	6.45	4.49
Egypt	88.54	86.63	90.15	87.74	14.09	8.36
India	74.71	55.77	85.38	73.67	42.21	40.47
Germany	22.74	17.82	37.53	29.34	19.99	14.02
S. Korea	58.72	56.05	60.56	60.43	4.44	9.97
UK	73.07	72.11	71.85	71.13	−4.52	−3.51
USA	58.25	49.06	55.60	47.08	−6.35	−3.88

Table 5 compares our results with existing results. The datasets used in Singh et al. (2020) [21, 22] are the number of daily death cases from Jan. 21 to April 11, 2020, and the number of daily confirmed cases from May 10 to June 7, 2020, respectively. Singh et al. (2020) [21] considered ARIMA and wavelet-ARIMA (W-ARIMA), while Singh et al. (2020) [22] considered ARIMA and LS-SVM, and both used 0.8 for the training ratio. Both studies obtained the orders of ARIMA in the same way as ours. The datasets used in Shahid et al. (2020) [23] are the numbers of daily confirmed cases, deaths cases, and recovered cases from Jan. 22 to May 10, 2020. The study considered three models, ARIMA, SVR, and LSTM, with a training ratio of 0.7. In the ARIMA model, (p,d,q)=(1,1,1) was used. The values in the table are results for the daily confirmed cases.

Direct comparisons with these studies seem difficult because the data and training ratios used in the experiments are different. However, we can use the existing measurement ranges as the criterion for our results. Since existing results were obtained with short-term data before vaccination, we also used a similar small-size dataset. Besides, we considered different training ratios for the small-size data to investigate its effect on performance. For the small-size dataset, optimally obtained orders and hyperparameters are used, which are (1,1,1) ((1,2,0)) for ARIMA and GARCH and (1,8) ((1,4)) for LSTM DNN for the data of the UK (USA) for both training ratios. The table shows that MAE and RMSE for the small-size dataset are much less than those for the large dataset. It seems due to the numbers of daily confirmed cases in the small-size dataset, which are less than those in the large dataset. That is, training with a small-size dataset is suitable to predict a small number of confirmed cases. While, training a training set, in which the small numbers of confirmed cases contained more than the large numbers of confirmed cases, seems not suitable to predict the large numbers of confirmed cases. Based on this observation, we can assume that a large dataset may yield the worse result. However, this cannot be assumed in general, as optimal parameters may vary depending on datasets.

Besides, vaccination seems to affect performance. It is noteworthy that the results with the full dataset are better than those of Singh et al. (2020) [22]. Several reasons can be considered, such as vaccination and different orders and training ratios.

Table 5. Comparison with existing results

Comparison	Data	Model	*r*	UK		USA	
				MAE	RMSE	MAE	RMSE
This study	2019.12.31 ~2020.09.02	ARIMA	0.9	249.52	304.44	16465.28	17690.02
		GARCH		425.92	515.01	15609.36	17992.26
		LSTM DNN		262.85	292.89	6044.48	7067.24
		ARIMA	0.8	520.77	583.50	10054.21	11782.22
		GARCH		384.89	465.86	11505.35	14172.12
		LSTM DNN		195.98	**253.31**	5987.86	7823.42
	2019.12.31 ~2021.02.22	ARIMA	0.8	**16155.33**	**21721.85**	**58409.90**	**66901.96**
		GARCH		**15456.04**	**20984.41**	**54921.64**	**64402.18**
		LSTMDNN		**4349.66**	**6057.66**	**24381.41**	**34078.02**
Singh et al. (2020) [21]	2020.01.21 ~2020.04.11	ARIMA	0.8	1316	1724	2822	4103
		W-ARIMA		193	253	1341	1974
Singh et al. (2020) [22]	2020.05.10 ~2020.06.07	ARIMA	0.8	2750	3381	21339	31972
		LS-SVM		3520	3964	18405	22667
Shahid et al. (2020) [23]	2020.01.22 ~2020.05.10	ARIMA	0.7	83359.04	98881.48	34867.61	61859.84
		SVR_Poly		32152.97	35442.09	244528.11	273851.39
		SVR_RBF		33336.29	37554.32	257046.10	298513.60

5. Conclusion

The numbers of daily confirmed cases of COVID-19 are analyzed and predicted by three models: ARIMA, GARCH, and stacked LSTM DNN. Datasets of two sizes are used in the experiment to investigate the effects of data size and vaccination on performance. Experimental results show that LSTM DNN predicts best for all datasets in terms of MAE (NMAE) and RMSE (NMSE), while the performances of ARIMA and GARCH depend on datasets. The NN with one LSTM cell outperformed the NN with more LSTM cells in many cases, which should be investigated later. We will expand this study to the NN models, including GAN and meta-learning techniques, and the datasets including more components, such as the number of daily deaths of the disease. The proposed method also can be applied to image data of the disease, such as chest X-rays of patients.

6. Declarations

6.1. Funding

This work was supported by the Mid-career Research Program through the NRF Grant funded by the Korea government (MEST) (NRF-2019R1A2C1002706).

6.2. Ethical Approval

All procedures performed in studies involving human participants were in accordance with the Italian National Research Council (CNR) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

6.3. Data Availability Statement

The data presented in this study are available in article.

6.4. Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Yang, P., & Wang, X. (2020). COVID-19: a new challenge for human beings. Cellular & Molecular Immunology, 17(5), 555–557. doi:10.1038/s41423-020-0407-x.

[2] World Health Organization. Coronavirus. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed on 17 March 2021)

[3] Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets. IEEE Access, 8, 181074–181090. doi:10.1109/access.2020.3027350.

[4] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., … Feng, Z. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine, 382(13), 1199–1207. doi:10.1056/nejmoa2001316.

[5] Lin, Q., Zhao, S., Gao, D., Lou, Y., Yang, S., Musa, S. S., … He, D. (2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. International Journal of Infectious Diseases, 93, 211–216. doi:10.1016/j.ijid.2020.02.058.

[6] Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., … Hellewell, J. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health, 5(5), e261–e270. doi:10.1016/s2468-2667(20)30073-6.

[7] Ndaïrou, F., Area, I., Nieto, J. J., & Torres, D. F. M. (2020). Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. Chaos, Solitons & Fractals, 135, 109846. doi:10.1016/j.chaos.2020.109846.

[8] Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., … Flasche, S. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases, 20(5), 553–558. doi:10.1016/s1473-3099(20)30144-4.

[9] Zhao, S., Lin, Q., Ran, J., Musa, S. S., Yang, G., Wang, W., … Wang, M. H. (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. International Journal of Infectious Diseases, 92, 214–217. doi:10.1016/j.ijid.2020.01.050.

[10] Shen, M., Peng, Z., Xiao, Y., & Zhang, L. (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China. bioRxiv. doi:10.1101/2020.01.23.916726.

[11] Çakan, S. (2020). Dynamic analysis of a mathematical model with health care capacity for COVID-19 pandemic. Chaos, Solitons & Fractals, 139, 110033. doi:10.1016/j.chaos.2020.110033.

[12] Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet, 395(10225), 689–697. doi:10.1016/s0140-6736(20)30260-9

[13] Shah, N. H., Suthar, A. H., & Jayswal, E. N. (2020). Control Strategies to Curtail Transmission of COVID-19. International Journal of Mathematics and Mathematical Sciences, 1, 1–12. doi:10.1101/2020.04.04.20053173.

[14] Intissar, A. (2020). A Mathematical Study of a Generalized SEIR Model of COVID-19. SciMedicine Journal, 2, 30–67. doi:10.28991/scimedj-2020-02-si-4

[15] Zheng, N., Du, S., Wang, J., Zhang, H., Cui, W., Kang, Z., … Xin, J. (2020). Predicting COVID-19 in China Using Hybrid AI Model. IEEE Transactions on Cybernetics, 50(7), 2891–2904. doi:10.1109/tcyb.2020.2990162.

[16] Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons & Fractals, 134, 109761. doi:10.1016/j.chaos.2020.109761.

[17] Choi, S., & Ki, M. (2020). Estimating the reproductive number and the outbreak size of COVID-19 in Korea. Epidemiology and Health, 42, e2020011. doi:10.4178/epih.e2020011.

[18] Ivorra, B., Ferrández, M. R., Vela-Pérez, M., & Ramos, A. M. (2020). Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. Communications in Nonlinear Science and Numerical Simulation, 88, 105303. doi:10.1016/j.cnsns.2020.105303.

[19] Chen, T.-M., Rui, J., Wang, Q.-P., Zhao, Z.-Y., Cui, J.-A., & Yin, L. (2020). A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. Infectious Diseases of Poverty, 9(24). doi:10.1186/s40249-020-00640-3.

[20] Roy, S., Bhunia, G. S., & Shit, P. K. (2020). Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. Modeling Earth Systems and Environment. doi:10.1007/s40808-020-00890-y.

[21] Singh, S., Parmar, K. S., Kumar, J., & Makkhan, S. J. S. (2020). Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19. Chaos, Solitons & Fractals, 135, 109866. doi:10.1016/j.chaos.2020.109866.

[22] Singh, S., Parmar, K. S., Makkhan, S. J. S., Kaur, J., Peshoria, S., & Kumar, J. (2020). Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries. Chaos, Solitons & Fractals, 139, 110086. doi:10.1016/j.chaos.2020.110086.

[23] Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos, Solitons & Fractals, 140, 110212. doi:10.1016/j.chaos.2020.110212.

[24] Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., & Pinheiro, P. R. (2020). CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. IEEE Access, 8, 91916–91923. doi:10.1109/access.2020.2994762.

[25] Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., … Crowcroft, J. (2020). Leveraging Data Science To Combat COVID-19: A Comprehensive Review. doi:10.36227/techrxiv.12212516.

[26] Pham, Q.-V., Nguyen, D. C., Huynh-The, T., Hwang, W.-J., & Pathirana, P. N. (2020). Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts. IEEE Access, 8, 130820–130839. doi:10.1109/access.2020.3009328.

[27] Chamola, V., Hassija, V., Gupta, V., & Guizani, M. (2020). A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact. IEEE Access, 8, 90225–90265. doi:10.1109/access.2020.2992341.

[28] Kim, M. (2020). Network traffic prediction based on INGARCH model. Wireless Networks 26(8), 6189–6202. doi:10.1007/s11276-020-02431-y.

[29] Kim, M. (2021). ML/CGAN: Network Attack Analysis using CGAN as Meta-Learning. IEEE Communications Letters 25(2), 499-502. doi:10.1109/lcomm.2020.3029580.

[30] Shumway, R. H., & Stoffer, D. S. (2017). Time Series Analysis and Its Applications. Springer Texts in Statistics. doi:10.1007/978-3-319-52452-8

[31] Azzouni, A., & Pujolle, G. (2017). A long short‐term memory recurrent neural network framework for network traffic matrix prediction. arxiv preprint arxiv: 1705.05690 v3. https://arxiv.org/abs/1705.05690.

[32] Available online: https://github.com/owid/covid-19-data/tree/master/public/data (accessed on 17 February 2021).