Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(9): 8935–8962.
DOI: 10.3934/mbe.2022415
Received: 06 January 2022
Revised: 01 March 2022
Accepted: 21 March 2022
Published: 20 June 2022

Research article

Data driven time-varying SEIR-LSTM/GRU algorithms to track the spread of COVID-19

Lin Feng [1], Ziren Chen [1], Harold A. Lay, Jr. [2,*], Khaled Furati [3] and Abdul Khaliq [4]

[1] Department of Mathematics, Iowa State University, Ames, IA 50011, USA

[2] Thompson Machinery Commerce Corporation, 1245 Bridgestone Blvd LaVergne, TN 37086, USA

[3] Department of Mathematics, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

[4] Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132, USA

* **Correspondence:** Email: jj.lay@tmcat.com.

Abstract: COVID-19 is an infectious disease caused by a newly discovered coronavirus, which has become a worldwide pandemic greatly impacting our daily life and work. A large number of mathematical models, including the susceptible-exposed-infected-removed (SEIR) model and deep learning methods, such as long-short-term-memory (LSTM) and gated recurrent units (GRU)-based methods, have been employed for the analysis and prediction of the COVID-19 outbreak. This paper describes a SEIR-LSTM/GRU algorithm with time-varying parameters that can predict the number of active cases and removed cases in the US. Time-varying reproductive numbers that can illustrate the progress of the epidemic are also produced via this process. The investigation is based on the active cases and total cases data for the USA, as collected from the website "Worldometer". The root mean square error, mean absolute percentage error and r_2 score were utilized to assess the model's accuracy.

Keywords: SEIR; LSTM; GRU; time-varying parameters; data-driven; COVID-19; time-varying reproduction number

1. Introduction

In early December 2019, the first case of Coronavirus 2019 (COVID-19 [1]) was reported in Wuhan, Hubei Province of China. The disease broke out on a large scale and spread rapidly around the world, becoming one of the most fatal pandemics [2] in human history. The COVID-19 virus is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus Type 2 (SARS-CoV-2). COVID-

19 poses a continuous threat to human health with its high transmission rate, serious health effects and changing genetic makeup.

It is critical to understand and analyze the rate of spread and trend of a disease during a pandemic. Only when we have a sufficient understanding of how the disease spreads can we propose targeted measures to slow it, such as using masks, closing non-essential facilities and isolating. The mathematical modeling of epidemics can help us to better understand the underlying mechanisms that affect the spread of diseases, and to help quantify possible control strategies by calculating descriptive quantities to track the disease's spread [3]. Salvadore et al. developed a model to quantify the impact of various control strategies used in the regions of Italy [4]. One very important threshold quantity is the basic reproduction number, which is usually denoted by R_0, and is also known as the basic reproduction ratio [5,6] or basic reproductive rate [7,8].

With the continuous spread and mutation of COVID-19, a significant challenge for researchers in several science areas has become how to help slow or halt its spread. Various models, estimation methods and forecasting approaches have been introduced to help understand and manage this pandemic [9]. The susceptible-exposed-infected-recovered (SEIR) and susceptible-infected-recovered (SIR) models are two of the most commonly used and convincing mathematical models. However, due to the continuous mutation of the virus, the inconsistent reporting of cases and the differences in the response measures taken by people and governments during different periods, the data-driven parameter estimation of mathematical models has become a major challenge faced by researchers [10]. Many parameter estimation methods for the SIR/SEIR model have been proposed and applied to COVID-19 data. Bentout et al. [11] applied the least squares method to estimate the epidemic parameters and the basic reproduction number R_0. Oliveira et al. [12] used a Bayesian method (MCMC) to estimate the parameters of the SIR model. The biggest limitation of these methods is that they can only fit fixed parameters for the entire time period, or fixed parameters for segmented time periods, and thus cannot produce dynamic parameter predictions. In fact, the parameters in the SEIR model are all time-dependent, and they will be affected by various factors over time. The traditional SEIR models with fixed parameters greatly limit our prediction of the epidemic because of the timeliness of their parameters. Therefore, effectively estimating time-dependent parameters has become a difficult task and challenge.

Most recently, machine learning has been applied to a variety of problems in many fields [13]. Machine learning methods are being used analyze and predict the epidemic trend of COVID-19 [14]. Various RNN methods such as long-short-term-memory (LSTM) and gated recurrent units (GRU) are commonly used as well-performing machine learning methods for time series data sets such as those used to analyze COVID-19. Zeroual et al. [15] compared five common machine learning methods, including LSTM and GRU, to study and predict the number of new and recovered cases. In the study by Shahid et al. [16], five machine learning methods including LSTM and GRU, were compared and evaluated via time-series forecasting of the population, death and recoveries in 10 major countries affected by COVID-19. Fokas et al. applied data from several countries to a birational model and a bi-directional LSTM network [17]. Many studies have demonstrated the effectiveness of machine learning methods, such as LSTM and GRU, for epidemic analysis and prediction. However, they have the common problem that they can only analyze and predict final result data, such as infected cases and removed cases. However, the analysis of the parameters proposed in some models, such as the transmission rate, removed rate and their derivative reproduction rate in the SEIR model, are helpful to study the rate of spread. Simply having the number of cases limits the discussion and analysis of the outbreak of COVID-19 in the research.

To address the problems described above, we propose a SEIR-LSTM/GRU model that optimally combines the mathematical models and machine learning algorithms. This model maintains the ability of the LSTM or GRU to predict the trend of the epidemic and produce more accurate results compared to those obtained by using the constant parameters in the basic SEIR model. the parameters in the SEIR-LSTM/GRU model, including the infection rate and removed rate, are time-dependent and estimated by LSTM and GRU models. Also, the time-varying reproduction rate is generated by using a time-varying infection rate and removed rate. We will discuss the relationship between the reproduction number and the epidemic trend of COVID-19 based on the results. To verify the effectiveness of the model, we compared its results to those of the LSTM and GRU models directly. The results show that our model performs similar to or better than the LSTM and GRU models alone. The four data-driven forecasting models used in this study are as follows: LSTM, GRU, SEIR-LSTM, and SEIR-GRU models. They were applied to a COVID-19 time series for the number of active cases and removed cases for the United States of America, and the models' accuracies were compared using several indicators. Recently, Long et al. [18] used physics informed neural networks combined with LSTM for the prediction and identification of time-varying parameters of COVID-19 in a way that was similar to this pstudy; however, they used different methods; we refer the reader to this paper as a reference. In this study, we used the first 240 data points (April 15, 2020 to December 10, 2020) as training data and the last 21 data points (December 11, 2020 to December 31, 2020) as test data to compare the models.

2. Methods

2.1. SEIR model

2.1.1. Model introduction

The SEIR model divides the population into four compartments: susceptible individuals, exposed individuals, infected individuals, and removed individuals. The model requires the following assumptions:

1) Population dynamics such as birth, natural death and mobility are not considered.
2) Removed individuals will not be infected again.
3) Exposed individuals cannot be infectious. In another words, the infectious group is the only group that can infect other individuals.

Figure 1 presents the flowchart of the SEIR model.

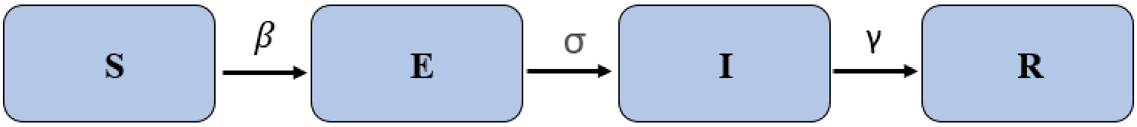


Figure 1. Flow chart of SEIR model.

The variables and parameters of the model are as follows:

S represents the number of susceptible individuals, which is the population that could be infected. At the beginning of the outbreak, we can assume almost all of the population is susceptible, because the number of the infectious individuals is very small compared to the whole population at initial breakout time.

E represents the number of exposed individuals, which is the population that has been infected but does not show symptoms. Generally, we state that this group is in the incubation period.

I is the number of infected individuals following the incubation period.

R represents the number of removed individuals, or the total population of recovered individuals and individuals who died from the disease. The reason why recovered individuals are included in the removed group is because the traditional SEIR model assumes that people who have been infected are immune to the disease and will not be infected again.

β is the transmission rate. In the SEIR model, β is the parameter that transports people from the susceptible group S to the exposed group E.

σ is the incubation rate, which is the inverse of the average incubation time. It controls the time from asymptomatic to symptomatic for a person who has been in contact with an infected person. σ is the parameter that transports people from the exposed group E to the infectious group I.

γ is the removal rate, which is the summation of the recovery rate and the death rate for the disease. It is the parameter that transports people from the infectious group I to the removed group R.

At the very beginning of a pandemic, the number of infected individuals is lowest, which implies the number of susceptible individuals is at its highest. The number of susceptible individuals continues to decrease as time passes, while the number of infectious individuals continues increasing. The following differential equation system reflects these changes:

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\frac{\beta S(t)I(t)}{N} \\
\frac{dE(t)}{dt} &= \frac{\beta S(t)I(t)}{N} - \sigma E(t) \\
\frac{dI(t)}{dt} &= \sigma E(t) - \gamma I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t)
\end{aligned} \tag{1}$$

where

$$N = S(t) + E(t) + I(t) + R(t) \tag{2}$$

is the total local population of the investigated area, and $S(t)$, $E(t)$, $I(t)$ and $R(t)$ are the varying susceptible, exposed, infected, and removed individuals, respectively.

Furthermore, it can be seen that

$$\frac{dN}{dt} = \frac{d(S + E + I + R)}{dt} = 0 \tag{3}$$

2.1.2. Basic reproduction number (R_0)

In epidemiology, the basic reproduction number, R_0, of an epidemic refers to the expected number of cases directly produced by one case in a population where all individuals are susceptible to infection and without the influence of external forces. In the SEIR model, R_0 can be calculated as detailed in [19].

$$R_0 = \frac{\beta}{\gamma} \tag{4}$$

Regarding R_0, there are two aspects that need special explanation. One is the interpretation of R_0 in the spread of an infectious disease, and the other is the limitation of the application of R_0.

(1) Significance of the exploration of R_0

There are three scenarios that indicate the possible spread or decline of a disease based on the value of R_0:

1) $R_0 < 1$: each infected individual infects less than one new individual, which implies the disease will die out at some future time.

2) $R_0 = 1$: each infected individual infects exactly one new individual, which implies the disease has reached an equilibrium.

3) $R_0 > 1$: each infected individual infects more than one new individual, which implies the disease will spread among individuals, resulting in an outbreak or epidemic.

(2) Limitation of R_0

From the definition, we can see that R_0 shows the average number of new infections from people who have the disease. It is a suitable metric when most of the population has not previously been infected and have not been vaccinated. Once an immunity is established or the contact rate among individuals is reduced due to the influence of external forces, then R_0 will begin to change. In this scenario, other models such as the piece-wise SEIR can be leveraged [20].

Thus, a constant value of R_0 for a disease is only applicable when most of the population is susceptible to the disease. The conditions are as follows:

1) No vaccine is available;

2) No one has developed immunity to the disease by contracting it; and

3) There is no way to control the spread of the disease.

Therefore, a constant R_0 only applies to the initial stage of the outbreak when the three conditions above are present.

2.1.3. Challenges of parameter identification

Parameter estimation is very important for the numerical solution of the SEIR model, because it can directly affect the accuracy of the results. Many researchers use traditional statistical methods, such as least squares [11] and Bayesian (MCMC) methods [12] to estimate the parameters of the SEIR model. The parameters estimated by these mature methods are generally the optimal parameters that can satisfy the conditions of the current epidemic trend and are usually expressed as constants.

However, the parameters of the SEIR model, including the transmission rate β, incubation rate σ, and removed rate γ, are all time-dependent due to the effects of factors such as the reporting rate, government policies, and medical advances. In this case, the parameters that are estimated using general estimation methods can only be applied to solve for specific windows of time during the epidemic, and the results of long-term simulations will deviate due to changes in parameters. Section 2.3 introduces a novel method to solve this problem.

2.2. Deep learning

In order to solve the problems of SEIR parameter estimation discussed in the last section, we propose a method of learning and estimating SEIR model parameters using the deep learning methods LSTM and GRU.

2.2.1. Artificial neural networks

Artificial neural network (ANNs) make up a type of a deep learning algorithm that is based on the idea of the human brain's biological neural network. ANNs try to simulate the operation of the human brain, but while its working principle is very similar to that of biological neural networks there are important differences.

2.2.2. Recurrent neural networks

Recurrent neural networks (RNNs) are a kind of ANN with memory. They can learn by saving past information and using it to perform future predictions.

Figure 2 presents the repeating module in a RNN. The values in the module are defined as follows:

x_t represents the input at time t;

h_t represents the hidden memory of the cell at time t;

$W_{x(t)}$ represents the weight matrix of x at time t; and

$W_{h(t)}$ represents the weight matrix of h_{t-1} at time t.

At time t, the new input and the memory of the previous cell are input at the same time and are combined into a new vector using the two different weight matrices. This vector contains the current input information and the previous memory, and the new hidden memory at time t is calculated using the activation function tanh. This information is passed to the next cell with the information at time t as the input. This process can be represented by Eq (5), where b is the bias.

$$h_t = tanh(W_{h(t)} * h_{t-1} + W_{x(t)*x_t} + b) \tag{5}$$

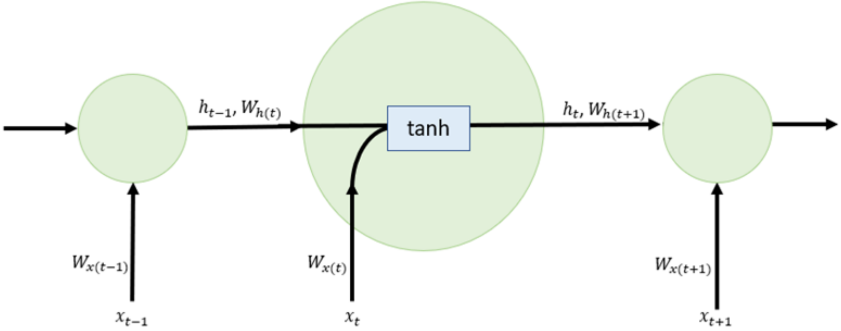


Figure 2. Repeating module in an RNN.

RNNs are primarily applied to sequential prediction problems [21–23]. Thus, we use RNNs for time series processing.

However, due to the difficulty in training, storing, and obtaining long-term memory information, basic RNN methods are not typically used when dealing with long sequence data. The most popular RNN methods that can solve these challenges effectively are LSTM- and GRU-based methods [24,25].

2.2.3. LSTM

LSTM networks are a long-term short-term storage networks that are used in the field of deep learning. They constitute a special type of RNN that can learn long-term dependencies, which is commonly used in sequence prediction problems.

The cell state and four gates are the core concept of LSTM networks. The cell state serves as a memory bank that runs through the entire sequence of processing. It can record relevant information during the process and pass it on. It is responsible for storing and transferring the long-term information through the sequence chain, which can be regarded as the "memory" of the neural network. As the sequence processing progresses, new or old information is added or removed from the cell state via various gates. These gates can learn and decide what information should be added and stored or be forgotten and removed during the training. Figure 3 depicts the repeating module for an LSTM network. There are a total of four gates in the repeating module of an LSTM network: forget, input, cell, and output.

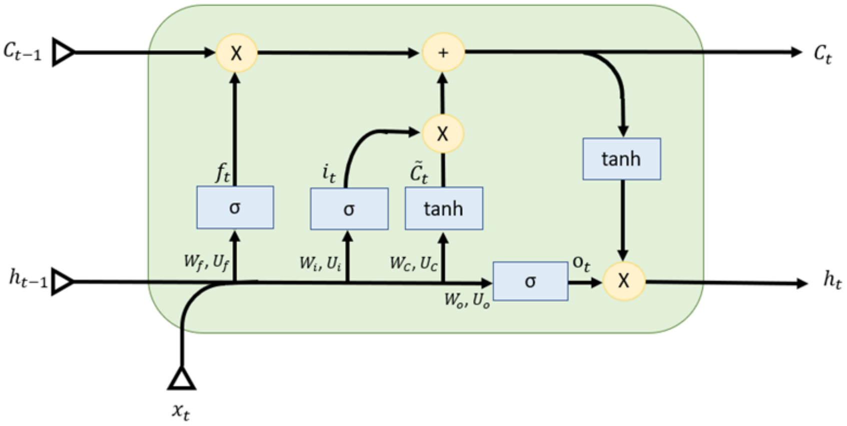


Figure 3. Repeating module(a cell) in an LSTM network.

The input of the repeating module of an LSTM network at time t includes the input of time $t(x_t)$, the output of the last cell (h_{t-1}), which brings the short-term memory and the cell state (C_{t-1}) from the previous cell, which keeps the long-term memory. In the diagram of the repeating module of the LSTM network, the blue box represents the activation function, and the yellow circle represents the arithmetic. Let W and U represent the weighted matrix of x_t and h_{t-1}, respectively, b represent the bias, and the subscripts "f", "i", "C" and "o" represent the forget gate, input gate, cell gate and output gate, respectively. When x_t and h_{t-1} enter each gate, they will combine the information through the corresponding weighted matrix. For example, when x_t and h_{t-1} enter the forget gate, the combined information can be expressed as $x_t + U_f h_{t-1} + b_f$.

(1) Forget gate

The first step in LSTM is to decide what information will be abandoned or kept from the cell state by implementing a sigmoid layer called the "forget gate". The inputs of the gate are h_{t-1} and x_t, and the output is a weight (0-1) matrix of the cell state C_{t-1}, where '1' represents "completely keep" and '0' represents "completely discard".

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{6}$$

(2) Input gate

The second step of LSTM is to decide what old information should be updated and what new information will be added to the cell state. This step includes two parts. First, it decides what information should be changed/updated in the cell state in a sigmoid layer, and then it creates a vector of new candidates that would be added to the cell state, $\widetilde{C}_t$, in a tanh layer. This step is called an "input gate".

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{7}$$

$$\widetilde{C}_t = tanh(W_C x_t + U_C h_{t-1} + b_C) \tag{8}$$

(3) Cell state

In this step, we update the old cell state C_{t-1} using the information we received from the previous gates. First, we multiply the updated and forgotten weight matrix of the old state obtained in the "forget gate" with the old state and filter the old information to determine what is preserved and what is discarded. Multiplying the old information by '1' means that the information is completely retained, and multiplying the old information by '0' means that the information is completely discarded. Then we multiply the results obtained in the input gate to obtain the new information that needs to be added and combine the updated old information to form the new information. The new information is finally recorded in the cell state.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{9}$$

(4) Output gate

Finally, we need to decide what will be the output of this repeating module from the cell state. First, we generate a weighted matrix to decide the output parts of the cell state by applying a sigmoid layer, where '1' denotes outputting all information and '0' means that nothing will be output. Then, we push the values of the cell state to be between −1 and 1 by using a tanh function. Finally, the result is multiplied by the weighted matrix to output the parts of the cell state (h_t).

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{10}$$

$$h_t = o_t * tanh(C_t) \tag{11}$$

2.2.4. GRU

GRUs are a type of LSTM network; they combine the forget and input gates into a single "update gate" and merge the cell state and hidden state to keep the long-term and short-term information together. Therefore, GRUs are more efficient than the traditional LSTM network. Their learning and prediction performance will vary based on the data set.

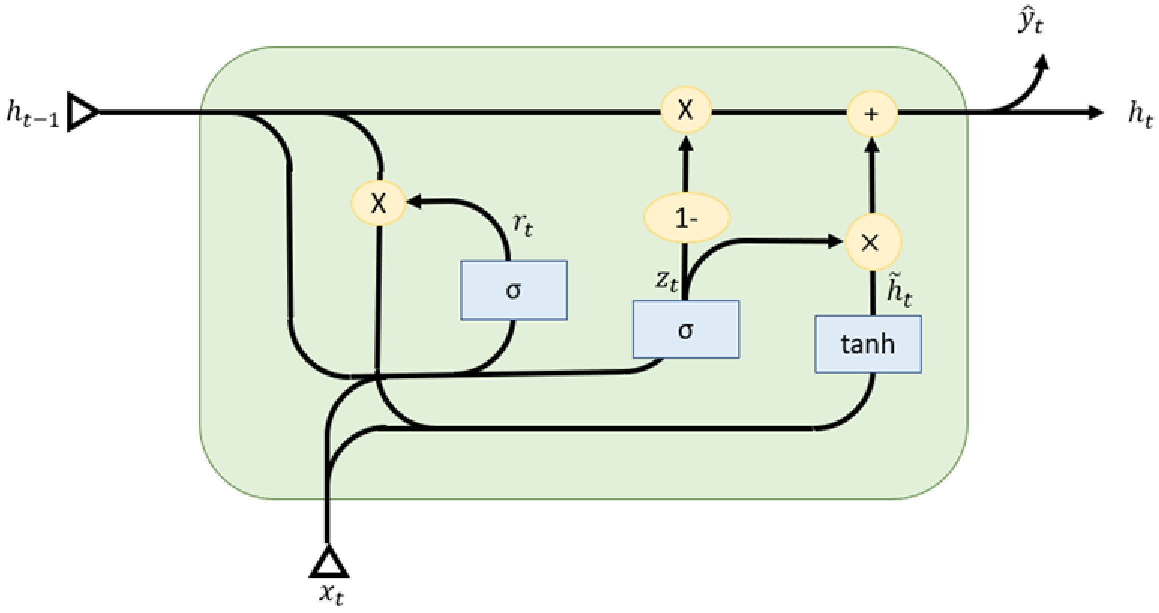


Figure 4. Repeating module in an GRU.

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_r + U_r h_{t-1} + b_r) \\ \widetilde{h_t} &= tanh(W x_t + U(r_t * h_{t-1}) + b_{\widetilde{h}}) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \end{aligned} \tag{12}$$

2.3. *SEIR-LSTM/GRU algorithm*

This section presents the time-varying SEIR-LSTM/GRU algorithms; the algorithm's framework is outlined in Figure 5. All subsections are developed around the sequence of the flowchart.

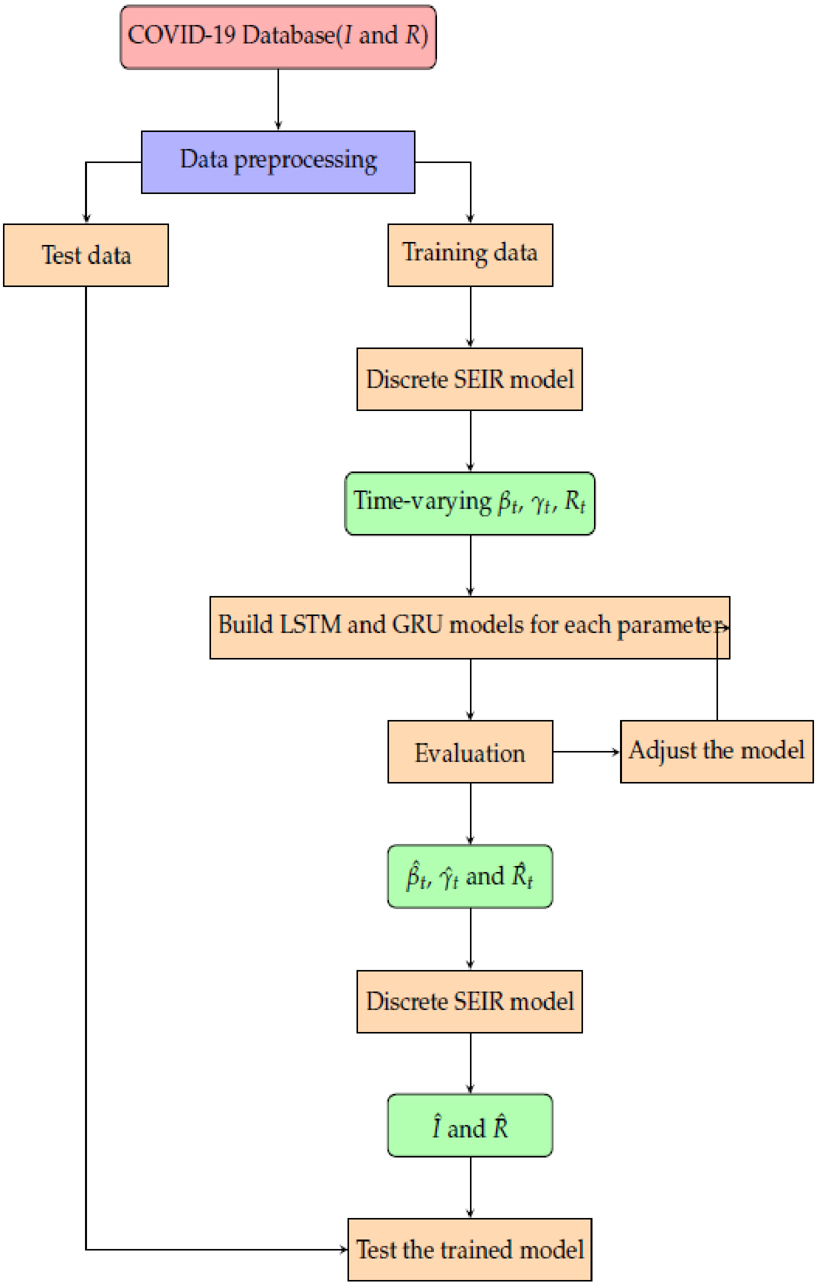
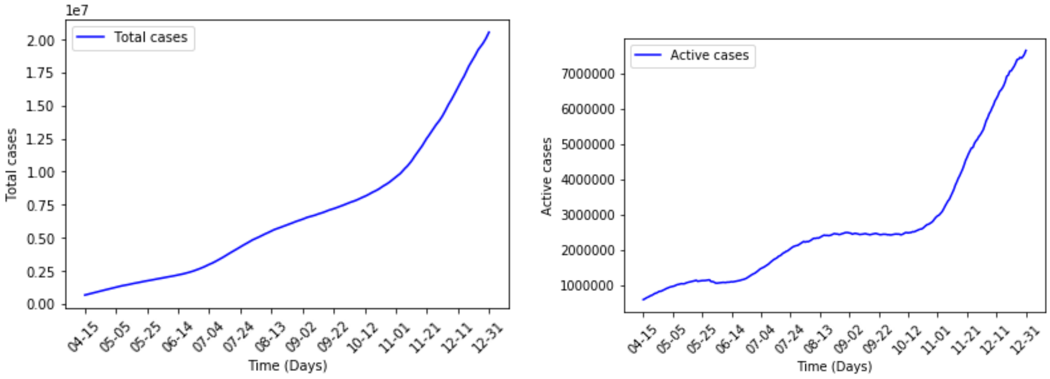

COVID-19 Database(I and R)
Data preprocessing
Test data
Training data
Discrete SEIR model
Time-varying β_t, γ_t, R_t
Build LSTM and GRU models for each parameter
Evaluation
Adjust the model
$\hat{\beta}_t$, $\hat{\gamma}_t$ and $\hat{R}_t$
Discrete SEIR model
$\hat{I}$ and $\hat{R}$
Test the trained model

Figure 5. Flowchart for time-varying SEIR-LSTM/GRU.

Figure 5 presents the flowchart of our method. First, we collect the data "total cases" and "active cases" from the website "Worldometer" [26] and perform preprocessing to obtain the training data and test data presented in Section 2.3.1. Then, as detailed in Section 2.3.2, we use the training data in the discrete SEIR model established by the forward euler method to estimate the time-varying numerical values for the parameters, transmission rate β and removed rate γ. In order to make the discrete model solvable, we fixed the parameter σ (incubation period) to be $\frac{1}{6}$ in accordance with the literature (see Section 2.3.2). At the same time, another important time dependent parameter, the recovery rate R_t is obtained by using the formula $R_t = \frac{\beta_t}{\gamma_t}$ where t is in days. Taking these time dependent parameters as our input, we train the LSTM and GRU models to obtain a reasonable model to identify the time-varing parameters of the SEIR model described in Section 2.3.3. Finally, in Section 2.3.4, we show how we utilize the estimated parameters in the discrete SEIR mdoel mentioned in Section 2.3.2 to solve the fit and prediction of the target data, active cases and removed cases. The details about the model are shown in the following subsections.

2.3.1. Data

Data collection. Our data comes from the website *"Worldometer"*. There are five types of data in total: total cases, daily new cases, active cases, total deaths and daily deaths. We downloaded the total cases and active cases from April 15, 2020 to December 31, 2021 for the United States (USA), which are shown in Figure 6 (a),(b), respectively. The total cases are the total cases that have been reported at that time, including active cases and removed cases. The active cases is the number of infected cases in the SEIR model.



(a) Total Cases

(b) Active Cases

Figure 6. (a) Total cases and (b) active cases for the USA between Feb 15, 2020 and Feb 14, 2021 (130 days).

Data preprocessing. Data preprocessing is a common data mining technique [27–30] used to transform the original data into a more efficient and useful data format for the subsequent data analysis process. The original data may have missing values or contain a lot of noise, which is unfavorable for the training of the model. Moreover, different algorithms may need different preprocessing approaches. Data preprocessing has a significant impact on the performance of the LSTM and GRU models used in this study [31].

Next, we perform simple preprocessing on the raw data to prepare it for the algorithm.

a. Left censoring. It has been observed that regardless of which country is studied, the data from the early stages of the pandemic are inaccurate. This could be due to the monitoring and reporting systems not being complete or effective, thus resulting resulted in incomplete information collection or censored information. In order to reduce the impact of this incomplete information on the results, we decided to delete the data points with insufficient information at the beginning and reset the start time of study for each group of data. We assumed that the monitoring and reporting system would be more complete two months after the outbreak started, so we chose April 15, 2020 as the new start time.

b. Right censoring. With the introduction of the COVID-19 vaccine, the epidemic situation in many places was brought under control. However, the vaccines at this stage were immature, and the virus continued to aggressively spread. In order to reduce the impact of the vaccine on the data, we chose to remove the data after the vaccine was produced. Most countries began to distribute vaccines from mid-to-late December. Therefore, we discarded data after December 31 for this study.

c. Derivation of removed data. Our data consisted of the total cases and the active cases for the USA from April 15, 2020 to December 31, 2020. In order to estimate the total removed cases (both recovered and deaths), we took into account the fact that the total cases at time t are all infected cases from the outbreak of COVID-19 to time t, and that the active cases are the currently infected individuals. Therefore, the difference between the two is the number of individuals who have been infected but removed. That is,

$$\text{Removed Cases} = \text{Total Cases} - \text{Active Cases}$$

The derived removed data are presented in Figure 7.

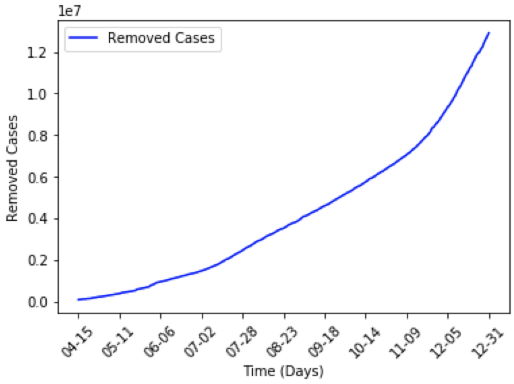
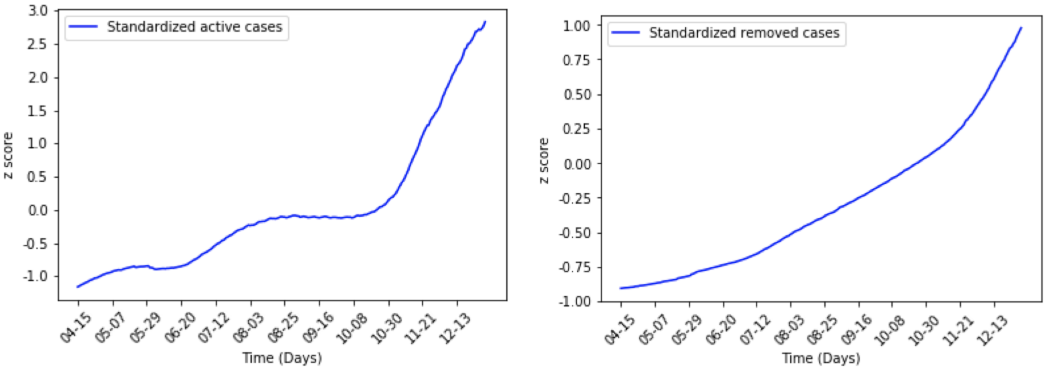


Figure 7. Derived removed cases for USA from April 15, 2020 to December 31, 2020

d. Data standardization. In machine learning, data standardization can be performed to indirectly avoid the impact of outliers and extreme values in the data on the training process. Here we chose the z-score standardization method to preprocess the data. The mean and standard deviation of the processed data are 0 and 1, respectively. The data standardization formula is

$$x' = \frac{x - \bar{x}}{\sigma_x}$$

where $\bar{x}$ and σ_x are the mean and standard deviation of the raw data, respectively.



(a) Standardized Active Cases (b) Standardized Removed Cases

Figure 8. Standardized (a) active cases; and (b) removed cases for the USA between April 15, 2020 and December 31, 2020.

Figure 8 presents the data for the active cases and removed cases after standardization.

e. Training and test data selection. The epidemic situations in different countries are affected by many different external factors at different stages, such as the reporting rate, different measures to respond to the epidemic and population movement measures; our model does not take these factors into account. Thus, our model is not suitable for the prediction of long-term data. Given these factors, we do not use the commonly used 80–20% division method to establish training data and test data but instead use the first 240 points of data as training data. We then use the obtained model to predict the remaining three weeks (21 days) of observations. The division is 92% training data - 8% test data.

2.3.2. Theoretical/numerical solutions of time-varying parameters of SEIR model

As mentioned in Section 2.1.1, the standard SEIR model can be expressed by the equations system (1). In order to solve the parameter problem mentioned in Section 2.1.3, we consider the time-varying parameters $\beta(t)$ and $\gamma(t)$ instead of the fixed values of β and γ; this provides the system of equations for the SEIR model with time-varying parameters as follows:

$$\begin{aligned}\frac{dS(t)}{dt} &= -\frac{\beta(t)I(t)S(t)}{N}\\ \frac{dE(t)}{dt} &= \frac{\beta(t)I(t)S(t)}{N} - \sigma E(t)\\ \frac{dI(t)}{dt} &= \sigma E(t) - \gamma(t)I(t)\\ \frac{dR(t)}{dt} &= \gamma(t)I(t)\end{aligned} \tag{13}$$

where $N = S(t) + E(t) + I(t) + R(t)$ is the total population of the model.

From our previous explorations, we know that S (susceptible), E (exposed), I (infected), R (removed), β (infected rate), γ (recovered rate) and σ (incubation rate) are all time-dependent variables. Since our model is based on the SEIR model without vital dynamics, we also know that the total population of the model is constant through time. To simplify the model, we set the incubation rate to be a constant. The incubation period for the coronavirus disease 2019 is 2–14 days [32]. Stephen et. al. [33] concluded that 5.1 days (95% CI, 4.5 to 5.8 days) is the median incubation period and that 97.5% of people will show symptoms within 11.5 days (CI, 8.2 to 15.6 days) of infection. Jantien et. al. [34] used the Weibull distribution to fit the data; they calculated the range of the incubation period to be from 2.1 to 11.1 days with a mean of 6.4 days (95% CI: 5.6 to 7.7 days). Here, we chose six days as the incubation period, which resulted in an incubation rate σ of $\frac{1}{incubationperiod} = \frac{1}{6}$.

In order to obtain reasonable values for the other parameters, we used the forward Euler method to discretize the ordinary differential equation (ODE) system of the SEIR model. Taking the step size to be one day, that is, $h = 1$, then the SEIR model can be expressed by the following system:

$$\begin{aligned} S_{t+1} &= S_t - \frac{\beta_t S_t I_t}{N} \\ E_{t+1} &= E_t + \frac{\beta_t S_t I_t}{N} - \sigma E_t \\ I_{t+1} &= I_t + \sigma E_t - \gamma_t I_t \\ R_{t+1} &= R_t + \gamma_t I_t \end{aligned} \tag{14}$$

with the symbol definitions in Table 1.

Table 1. Symbol definitions for the discrete SEIR model.

Symbol	Interpretation
S_t	Individuals not yet infected at time t
E_t	Individuals have been infected but are not yet infectious at time t
I_t	Individuals have been infected at time t and are now infectious
R_t	Individuals previously infected and then removed at time t
β_t	Transmission rate at time t
γ_t	Removed rate at time t
σ	Incubation rate
N	Total local population

Since the population remains constant, the sum of all terms on the left side of the equation is equal to the sum of all terms on the right side of the equation, that is,

$$\begin{aligned} N &= S_t + E_t + I_t + R_t \\ &= S_{t+1} + E_{t+1} + I_{t+1} + R_{t+1} \end{aligned} \tag{15}$$

Finally, the time-varying reproduction number at time t is represented by

$$R_t = \frac{\beta_t}{\gamma_t} \tag{16}$$

We now have actual data for the time period of interest for both the active cases and removed cases. We have established the incubation rate for the virus. Finally, we know the total population of the country we are studying. Thus, we know the values for R_t, R_{t+1}, I_t, I_{t+1}, σ and N; the unknown variables are β_t, γ_t, S_t, S_{t+1}, E_t and E_{t+1}. For the six equations of this system, we have six unknowns, so we can obtain numerical solutions for the parameters β and γ by solving the equation system. Algorithm 1 shows the process.

```
**Algorithm 1** Numerical solutions for the time-varying parameters of SEIR model
**Input:**
Local population of the surveyed area: $N$;
Number of iterations or days: $n$;
All sequential values of the real data $I$ and $R$ from $t = 0$ to $t = n$:
$I = \{I_0, I_1, ..., I_n\}$, $R = \{R_0, R_1, ..., R_n\}$;
Optimal incubation rate: $\sigma$;
**Output**
$\gamma = \{\gamma_0, \gamma_1, ..., \gamma_{n-1}\}$; $E = \{E_0, E_1, ..., E_{n-1}\}$; $S = \{S_0, S_1, ..., S_{n-1}\}$;
$\beta = \{\beta_0, \beta_1, ..., \beta_{n-2}\}$
**Procedure**
**For** $t$ in 0 to $n-1$, do
    $\gamma_t = \frac{R_{t+1}-R_t}{I_t}$
    $E_t = \frac{I_{t+1}-I_t+\gamma_t I_t}{\sigma}$
    $S_t = N - E_t - I_t - R_t$
**For** $t$ in 0 to $n-2$, do    $\beta_t = \frac{(S_{t+1}-S_t)N}{I_t S_t}$ or $\beta_t = \frac{(E_{t+1}-E_t+\sigma E_t)N}{I_t S_t}$
**For** $t$ in 0 to $n-2$, do    $R_t = \frac{\beta_t}{\gamma_t}$
```

2.3.3. Time-varying parameter identification of SEIR model with LSTM and GRUs

After obtaining the theoretical time-varying transmission rate and removal rate from April 15, 2020 to December 15, 2020 using the algorithm given above, the data were separated into two groups: training data (the first 240 data points) and test data (the remaining data of three weeks). Next, the training data were used as input for the LSTM and GRU models respectively described in Section 2.2.3 and 2.2.4. Finally, we perform the prediction for the next three weeks to obtain the values for β and γ using Algorithm 2.

For both the LSTM and GRU methods, we found that two hidden layers can provide accurate results through experimentation. We used the Adam optimization algorithm for the adaptive learning rate and a value of 2 for the step-size and batch size. Other parameters, like the epochs and hidden units in each layer are shown in the Table 2. The suitable numbers of epochs and hidden units were obtained by performing repeated experiments, starting from 100 epochs and 10 units for each hidden layer. Gradually we increased the number if under-fitting (the model fits the training data poorly) occurred or decreased the number if over-fitting (the prediction of the training data is very consistent, but the prediction of the test data is very bad) occurred.

Algorithm 2 Time-varying parameter learning and prediction using LSTM/GRUs

```
Input:
data (Numerical data of $\gamma$ or $\beta$ from Algorithm 1:
$\gamma = \{\gamma_0, \gamma_1, ..., \gamma_n\}$; $\beta = \{\beta_0, \beta_1, ..., \beta_n\}$);
epochs;
timesteps;
batch size;
Output:
Fitting and prediction of $\gamma$ & $\beta$
# Standardize the data
sc = StandardScaler()
data = sc.fit_transform(np.float64(data))
# Split data into $a\%$ training data and $(100 - a)\%$ testing data
$l$ = length(data)
train = data[0 : $l * a\%$]; test = data[$l * a\%$ : $l$]
# Creating a data structure with time steps
x_train = [ ]; y_train = [ ]
For $i$ in range(timesteps, $l * a\%$ + timesteps):
    x_train.append(train[$i$ - timesteps : $i$, 0])
    y_train.append(train[$i$ : $i$ + timesteps, 0])
# Learn the LSTM/GRU model for training data
Procedure fit_lstm(batch size, timesteps neurons)
model = Sequential()
model.add(LSTM(neurons, stateful=TRUE))
model.compile(optimizer='adam', loss = 'mse')
For $i$ in range(epochs), do
    model.fit(x_train, y_train, shuffle=False, epochs=1, batch_size=bath size))
    model.reset_states()
end for
return model
# Forecast
Procedure forecast_lstm(model, x_train)
y_predict=model.predict(x_train)
return y_predict
# Inverse transform
y_predict = sc.inverse_transform(y_predict)
# Plot the testing data and prediction
plt.plot(test)
plt.plot(y_predict)
plt.show()
# Fit the LSTM/GRU model
lstm_model = fit_lstm(train, epoch, neurons)
# Forecast the traning dataset
lstm_model.predict(tran)
```

Table 2. Parameter selection results obtained by using the LSTM/GRU network.

Step-size	Batch size	Epochs	Units in hidden layer 1	Units in hidden layer 2
2	2	200	10	10

We can now use Eq (16) to estimate the time-varying reproduction number (R_t) for each day from April 15, 2020 to December 31, 2020.

We can obtain the theoretical and predicted values of all model parameters using the above process. Applying regression analysis to the predicted results and theoretical results, we can compare and analyze the changes and trends over time. Furthermore, we can analyze the trend of COVID-19 in the United States of American by the analyzing the changes in the parameters.

2.3.4. Predictions for active cases and removed cases

After estimating the forecasting time-varying parameters, the transmission rate β, and the removed rate γ, we apply them to Algorithm 3 for the discrete SEIR model to predict the active cases and removed cases.

Algorithm 3 Predictions for active cases and removed cases

```
Input
Local population of the surveyed area: $N$;
Number of iterations or days: $n$;
Initial value of variables: $S_0$, $E_0$, $I_0$ and $R_0$;
All sequential predicted values of parameters:
$\beta = \{\beta_0, \beta_1, ..., \beta_n\})$; $\gamma = \{\gamma_0, \gamma_1, ..., \gamma_n\}$;
Optimal incubation rate: $\sigma$;
Output
$S = \{S_0, S_1, ..., S_n\}$, $E = \{E_0, E_1, ..., E_n\}$, $I = \{I_0, I_1, ..., I_n\}$, $R = \{R_0, ..., R_n\}$
Procedure
For $t$ in 1 to $n$, do
    $S_t = S_{t-1} - \frac{\beta S_{t-1} I_{t-1}}{N}$
    $E_t = E_{t-1} + \frac{\beta S_{t-1} I_{t-1}}{N} - \sigma E_{t-1}$
    $I_t = I_{t-1} + \sigma E_{t-1} - \gamma I_{t-1}$
    $R_t = R_{t-1} + \gamma I_{t-1}$
```

3. Results

3.1. Evaluation metrics

We utilized three commonly used error metrics, i.e., the root mean square error (RMSE), mean absolute percentage error (MAPE), and r_2 score, to measure the accuracy of the results. Let N be the number of data points, y_i be the actual value of the i^{th} data, and $\hat{y}_i$ be the prediction of the y_i, then the error metrics are as follows.

3.1.1. RMSE

The RMSE is a commonly used error metric; it is the square root of the quadratic mean of the differences between predicted and actual values. The RMSE is expressed mathematically as

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2} \tag{17}$$

The range of the MSE is $[0, +\infty)$. A large RMSE implies that the model would have a worse fit than a model with a smaller error. When $RMSE = 0$, it means that the predicted values match the actual values. While the instinct would be to select models with lower RMSE values, care must be taken to avoid overfitting.

3.1.2. MAPE

The MAPE measures the error in percentage; it is the absolute mean of the ratio of the predicted error to the actual value [35,36]. It reflects the relative error based on the actual data in the form of a ratio. This provides a means of comparing errors by eliminating potential differences in the scale of the errors. The MAPE is calculated as follows:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \tag{18}$$

As with the RMSE, the smaller the MAPE, the better the model. This metric works well for data with no extreme values or zero values.

3.1.3. r_2 score

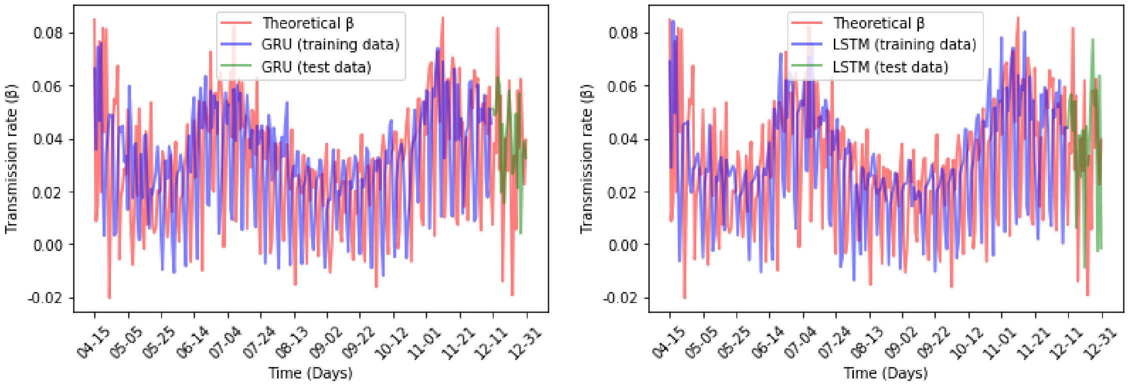
The r_2 score [37–39] is the proportion of the variance of the true value that the predicted value can explain, and it can reflect how well the predicted value fits the true value. The r_2 score is calculated as follows:

$$r_2(y, \hat{y}) = 1 - \frac{\sum_{t=1}^{N}(\hat{y}_t - y_t)^2}{\sum_{t=1}^{N}(y_t - \bar{y}_t)^2} \tag{19}$$

The range of the r_2 score is $(-\infty, 1]$ for the non-linear regression, and the closer the value of r_2 score is to 1, the better is the model is.

3.2. *Results for the parameters*

3.2.1. Results of transmission rate (β)

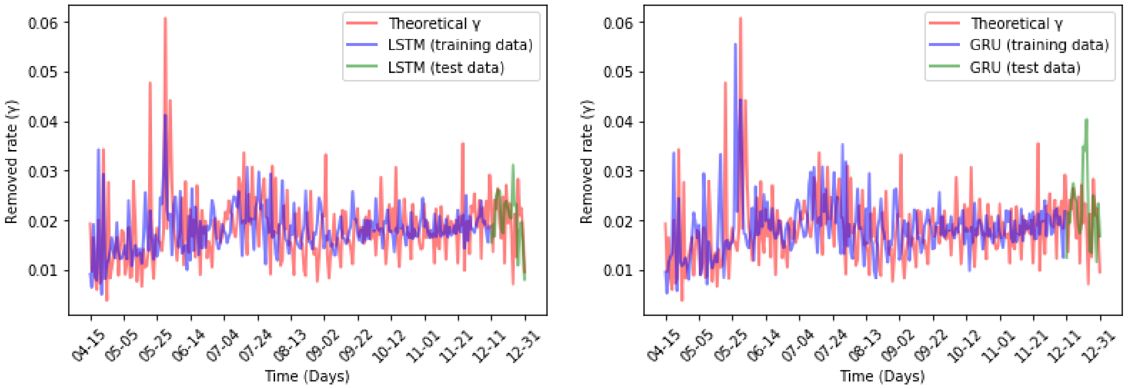


(a)

(b)

Figure 9. Variation of parameter β for USA cases between April 15, 2020 and December 31, 2020, as obtained via the (a) LSTM and (b) GRU networks. The red curve presents the theoretical values of β based on the SEIR model; the blue curve presents the predicted values for the training dataset of β; and the green curve presents the predicted values for the test dataset of β.

3.2.2. Results of removed rate (γ)

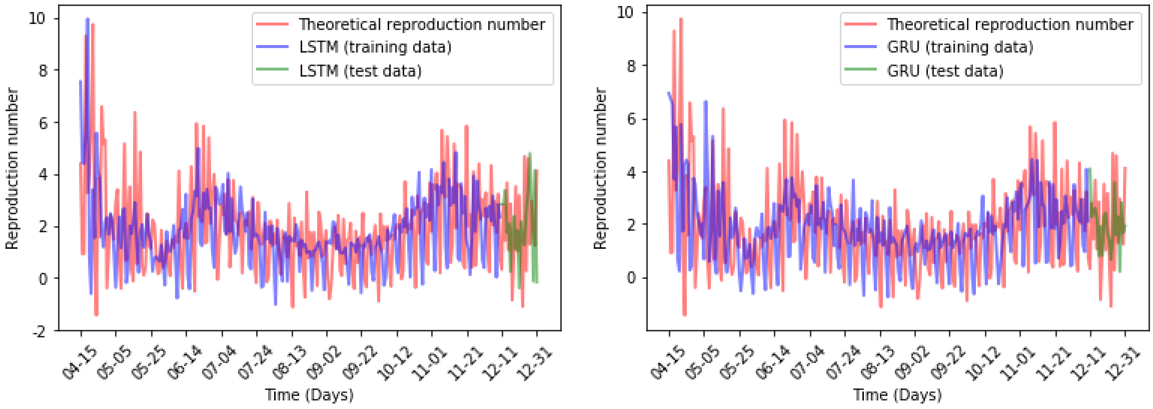


(a)

(b)

Figure 10. Variation of parameter γ for cases in the USA between April 15, 2020 and December 31, 2020, as obtained via the (a) LSTM and (b) GRUs. In these two graphs, the red curve presents the theoretical values of γ based on the SEIR model; the blue curve presents the predicted values for the training dataset of γ; and the green curve presents the predicted values for the test dataset of γ.

3.2.3. Results of time-varying reproduction number (R_t)



(a)

(b)

Figure 11. Variation of the reproduction number R_t for USA infections between April 15, 2020 and December 31, 2020, as obtained via the (a) LSTM and (b) GRU networks. In these two graphs, the red curve presents the theoretical values of R_t based on the SEIR model; the blue curve presents the predicted values for the training dataset of R_t; and the green curve presents the predicted values for the test dataset of R_t.

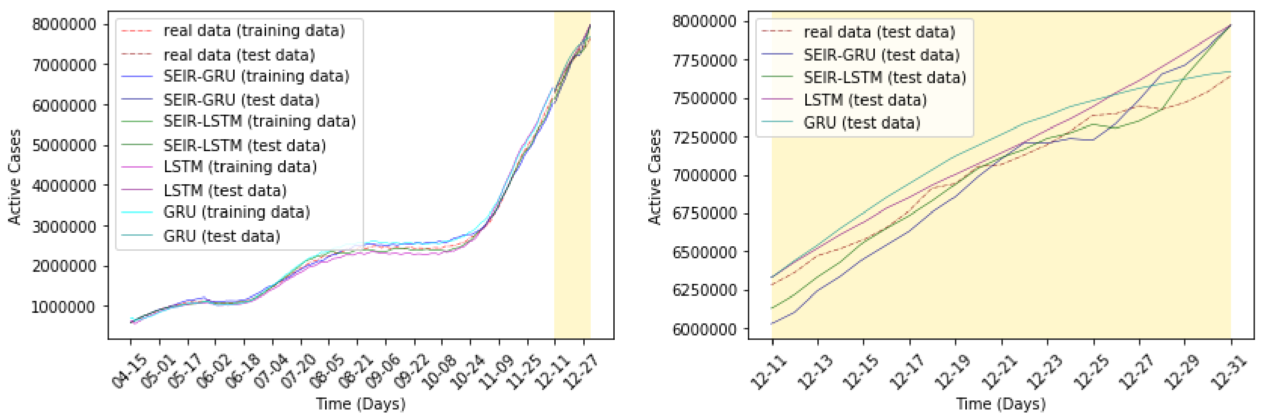
3.2.4. Additional statistical information

Table 3. Mean and variance results for parameters of different SEIR models.

Parameter	Data type	Mean	Variance
β_t	Theoretical	0.0306957	0.0005028
	Prediction by LSTM	0.0311974	0.0004248
	Prediction by GRU	0.0315283	0.0003758
γ_t	Theoretical	0.0182474	4.5075719×10^{-5}
	Prediction by LSTM	0.0182590	2.2022324×10^{-5}
	Prediction by GRU	0.0188130	3.5919904×10^{-5}
R_t	Theoretical	1.9042522	2.7817986
	Prediction by LSTM	1.8133258	1.9023700
	Prediction by GRU	1.8251087	1.7981305

Table 3 shows the mean and variance of each parameter obtained via the three methods. The variance of the predicted parameters was smaller than the theoretical variance. It can be seen that the mean and variance of the parameters predicted using LSTM were closer to the theoretical values than those predicted using the GRU.

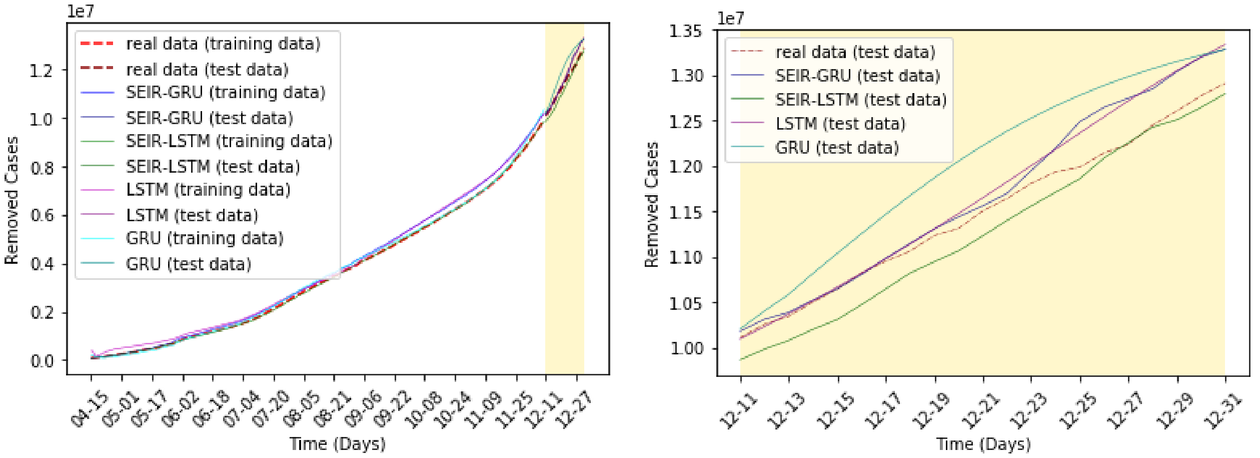
3.3. Solutions of active cases and removed cases using the time-varying parameters



(a)

(b)

Figure 12. Prediction of active cases for the USA for (a) all data and (b) the test data using the four methods, LSTM, GRU, SEIR-LSTM and SEIR-GRU.



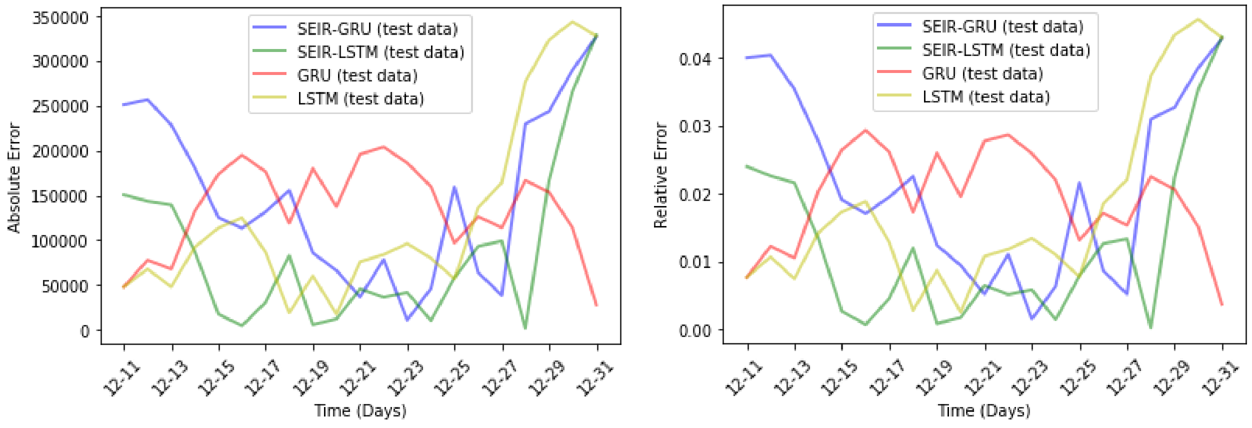
(a)

(b)

Figure 13. Prediction of removed cases for the USA for (a) all data and (b) the test data using the four methods, LSTM, GRU, SEIR-LSTM and SEIR-GRU.

Figures 12 and 13 present the predictions of the active cases and removed cases for the USA using the four methods, LSTM, GRU, SEIR-LSTM and SEIR-GRU.

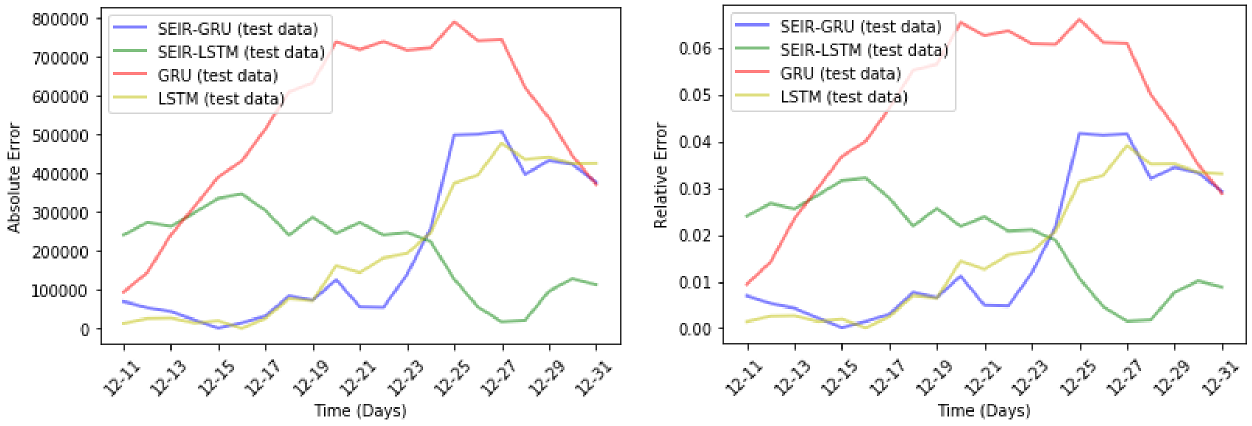
3.4. Results of errors



(a) Absolute error

(b) Relative error

Figure 14. Comparison of the (a) absolute error and (b) relative error for the test data of active cases for the USA using the four methods, LSTM, GRU, SEIR-LSTM and SEIR-GRU.



(a) Absolute error

(b) Relative error

Figure 15. Comparison of the (a) absolute error and (b) relative error for the test data of removed cases for the USA using the four methods, LSTM, GRU, SEIR-LSTM and SEIR-GRU.

Figures 14 and 15 show the error graphs for the test data of I and R that were respectively obtained using the four methods, LSTM, GRU, SEIR-LSTM and SEIR-GRU.

Table 4. Validation metrics for active cases and removed cases of COVID-19 forecasting using LSTM, GRU,SEIR-LSTM and SEIR-GRU models.

Variable	Model	RMSE	MAPE	r_2 score
Active cases (I)	LSTM	164687.0886764	0.0174709	0.8460605
	GRU	148185.4351031	0.0193679	0.8753644
	SEIR-LSTM	125023.1652022	0.0122529	0.9112820
	SEIR-GRU	178906.9787563	0.0213152	0.8183290
Removed cases (R)	LSTM	271232.1812481	0.0164587	0.9005331
	GRU	590706.5333376	0.0462272	0.5282199
	SEIR-LSTM	237405.1637538	0.0188187	0.9237962
	SEIR-GRU	279495.5196168	0.0164521	0.8943801

4. Discussion

4.1. Parameter discussion

Figure 9 presents the prediction results for β for the USA between April 15, 2020 and December 31, 2020, as obtained using LSTM and GRU networks. These figures show that the transmission rate experienced several upward and downward movements over the time frame of our study. Combined with our previous research on the different time periods of the USA epidemic under the influence of various measures [20], we interpreted its trend as follows: 1) because of the home isolation policies implemented in various states in March, there was a downward trend of β from April 15 to early May; 2) after that, due to the widespread return to work in May, the β showed an upward trend again; 3) until the beginning of July, β dropped again due to the increased awareness of prevention and policies such as encouraging masks in public areas and the closing of non-essential businesses; and 4) after an increase, the transmission rate showed a downward trend in winter under the influence of temperature. Overall, the change is not significant. From Table 3, we can observe that its mean is about 0.031 while the variance is about 5.0×10^{-5}.

It is not difficult to find from Figure 10 that there was no significant change in the value of γ, and that there was only a small upward trend in the early stage. From Table 3, we can see that its mean was about 0.018 and its variance was about 4.5×10^{-5}.

The results of the time-varying reproduction number R_t are presented in Figure 11. Because R_t is dependent on the ratio of β_t and γ_t, under the premise that β is relatively stable, the changes of its trend come primarily from γ. Because of the home isolation policies implemented in various states in March, the data shows a downward trend of R_t from April 15 to early May. After that, employees began to return to work in May, and the R_t reflects this with an upward trend again. Until the beginning of July, R_t dropped again due to the increased awareness of prevention and policies such as encouraging masks in public areas and closed businesses. After an increase, the R_t showed a downward trend in the winter months. During the time we studied, the change of R_t was not monotonically increasing or decreasing but exhibited an upward or downward trend during certain time intervals due to external factors. The change was an iterative and relatively smooth change, excluding the outliers. From Table 3, we can see that the mean of R_0 was about 1.9, i.e., each infected individual infects around 1.9 new individuals, which implies the disease will continue to spread unless there is an intervention (e.g., the vaccine).

4.2. Method discussion

In order to better illustrate the advantage of the proposed model, we compared four models: LSTM, GRU, SEIR-LSTM and SEIR-GRU models.

Figures 12 and 13 present the respective predictions for the active cases and removed cases for the USA, which were obtained byusing the four methods. In order to more intuitively show the differences in the predictions of the four methods for the two variables I and R, we respectively presented the absolute error and relative error graphs for the test data in Figures 14 and 15. The absolute and relative error graphs are similar in shape, but their scales and units are different. Because the scale of the real data was relatively large, the gap of 350,000 only accounts for 4% of the real data. As shown in Figure 14(b), the prediction results for the four methods are in the range of 0 to 0.04. Among them, the LSTM and SEIR-LSTM models fluctuated around 0.01 in most cases, while the other two methods are slightly higher. Based on the results, LSTM and SEIR-LSTM are more suitable for this data set from the USA than the other two models. Figure 15(b) verifies what we observed in Figure 13(b), which is that the GRU method has weak prediction ability for removed cases.

In order to evaluate the performance of the four models, we calculated the RMSE, MAPE and r_2 scores for the test data of the two variables I and R in Table 4. The RMSE, MAPE, and r_2 score of the SEIR-LSTM model for the prediction of the active cases were 125023.1652022, 0.0122529 and 0.9112820, respectively. The RMSE, MAPE and r_2 score of the SEIR-LSTM model for the prediction of the active cases were 237405.1637538, 0.0188187 and 0.9237962, respectively. Whether it is for the prediction of I or R, the SEIR-LSTM model outperforms the other models because it has the lowest RMSE and MAPE values, and it has the r_2 score closest to 1.

4.3. Limitation discussion

The method proposed in this paper suffers from two major limitations. The first is the instability of the model due to gradient descent. Because our method is based on LSTM and GRU and we use the Adam optimization algorithm for the adaptive learning rate, during the learning process it can converge to a local minimum instead of the full minimum. In this case, we need to repeat the experiment many times for each set of data to ensure that the optimal solution can be found, which makes the computation time potentially long. Another limitation is that our method is not one-size-fits-all, as it is based on LSTM and GRUs. If new data comes in, we need to retrain the model to capture the latest data features to improve its prediction results.

4.4. Application to other countries

The method we propose is not only applicable to data from the USA, but it can be applied to the data from other countries. We have made data-driven forecasts for Italy with good results. It is worth mentioning that for the USA data, we chose 240 data points as our training data. It is possible to improve the results and reduce the risk of overfitting by reducing the training dataset, because the characteristics of the data in the short-term are relatively stable. Additional results and code will be posted on the GitHub repository: https://github.com/Lin3829/Data-driven-time-varying-SEIR-LSTM-GRU-algorithms-for-the-spread-of-COVID-19.

5. Conclusions

In this study, we brought the parameters learned by LSTM and GRU networks to the SEIR model for the simulation of active cases and removed cases for the COVID-19 virus in the USA. The RMSE, MAPE, and r_2 scores were used to evaluate four models: LSTM, GRU, SEIR-LSTM and SEIR-GRU models. The results show that the SEIR-LSTM model performs very well for the predictions of the USA data from December 11, 2020 to December 31, 2020.

At the beginning of the paper, we introduced the SEIR model and discussed the challenges in estimating the parameters of the model. To solve this problem, we employed the deep learning algorithms LSTM and GRU to learn and predict the parameters of the SEIR model; we then combined traditional statistical methods to analyze the prediction results. The results showed that the change of R_t was not monotonically increasing or decreasing during the time we studied, but exhibited a cyclical trend due to changes by external factors. We observed that the mean value of the reproduction number was about 1.9, i.e., each infected individual infects around 1.9 new individuals. This shows that the disease will continue to spread if no additional measures are taken (e.g., a vaccine).

We put the estimated parameters of the LSTM and GRU networks back into the SEIR model for the simulation and compared the true values and predicted values of active cases and removed cases. Finally, we used RMSE, MAPE, and r_2 scores to evaluate the performance of the LSTM, GRU, SEIR-LSTM and SEIR-GRU models. The results show that the SEIR-LSTM model has smaller RMSE and MAPE values, and that the r_2 score value is closest to 1, regardless of whether it is for active cases or removed cases. The minimum MAPE was as low as 1.23%, and the r_2 score was as high as 0.911. This fully illustrates the potential of LSTM and GRUs for predicting the COVID-19 epidemic trend.

The main contribution of this study is the development of a model that optimally combines the SEIR model and the LSTM/GRU algorithms and generates a time-varying infection rate, removed rate, and reproduction rate. It analyzed the relationship between the time-varying reproduction number and the epidemic trend of COVID-19 and entailed the application of four models to forecast the time series of the number of active cases and removed cases for the USA.

We focused on the most basic compartmental model in epidemiology, the SEIR model, as the basis, and expanded around its parameter estimation problem. In fact, there are many other models that are inherently more effective than SEIR models, such as SEIRD (where individuals are susceptible, exposed, infected, recovered or dead), or dynamic SEIR models. Further research could expand beyond the two machine learning methods used here. In future work, we will apply more machine learning methods to explore their potential application in various fields.

References

1. *BBCnews*, Coronavirus disease named COVID-19, 2020. Available from: https://www.bbc.com/news/world-asia-china-51466362.

2. S. Roychoudhury, A. Das, P. Sengupta, S. Dutta, S. Roychoudhury, A. P. Choudhury, et al., Viral pandemics of the last four decades: Pathophysiology, health impacts and perspectives, *Int. J. Environ. Res. Public Health*, **17** (2020), 9411. https://doi.org/10.3390/ijerph17249411

3. F. Brauer, *Compartmental Models in Epidemiology*, Springer Berlin Heidelberg, (2008),19–79. https://doi.org/10.1007/978-3-540-78911-6_2

4. F. Salvadore, G. Fiscon, P. Paci, Integro-differential approach for modeling the COVID-19 dynamics-impact of confinement measures in Italy, *Comput. Biol. Med.*, **139** (2021) 105013. https://doi.org/10.1016/j.compbiomed.2021.105013

5. O. Diekmann, J. A. P. Heesterbeek, J. A. J. Metz, On the definition and the computation of the basic reproduction ratio r0 in models for infectious diseases in heterogeneous populations, *J. Math. Biol.*, **28** (1990), 365–382. https://doi.org/10.1007/BF00178324

6. J. M. Heffernan, R. J. Smith, L. M. Wahl, Perspectives on the basic reproductive ratio, *J. R. Soc. Interface*, **2** (2005), 281–293. https://doi.org/10.1098/rsif.2005.0042

7. R. M. Anderson, R. M. May, *Infectious diseases of humans, dynamics and control*, Oxford University Press, 1991.

8. P. van den Driessche, Reproduction numbers of infectious disease models, *Infect. Dis. Model.*, **2** (2017), 288–303. https://doi.org/doi:doi.org/10.1016/j.idm.2017.06.002

9. A. Zeroual, F. Harrou, A. Dairi, Y. Sun, Deep learning methods for forecasting COVID-19 time-series data: A comparative study, *Chaos Solitons Fractals*, **140** (2020) 110121. https://doi.org/doi.org/10.1016/j.chaos.2020.110121

10. G. Fiscon, F. Salvadore, V. Guarrasi, A. R. Garbuglia, P. Paci, Assessing the impact of data-driven limitations on tracing and forecasting the outbreak dynamics of COVID-19, *Comput. Biol. Med.*, **135** (2021), 104657. https://doi.org/10.1016/j.compbiomed.2021.104657

11. S. Bentout, A. Chekroun, T. Kuniya, Parameter estimation and prediction for coronavirus disease outbreak 2019 (COVID-19) in Algeria, *AIMS Public Health*, **7** (2020), 306–318. https://doi.org/10.3934/publichealth.2020026

12. A. C. S. de Oliveira, L. H. M. Morita, E. B. da Silva, L. A. R. Zardo, C. J. F. Fontes, D. C. T. Granzotto, Bayesian modeling of COVID-19 cases with a correction to account for under-reported cases, *Infect. Dis. Model.*, **5** (2020), 699–713. https://doi.org/10.1016/j.idm.2020.09.005

13. J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, Recent advances and applications of machine learning in solid-state materials science, *NPJ Comput. Materials*, **5** (2019), 83. https://doi.org/10.1038/s41524-019-0221-0

14. K. Olumoyin, A. Khaliq, K. Furati, Data-driven deep-learning algorithm for asymptomatic COVID-19 model with varying mitigation measures and transmission rate, *Epidemiologia*, **2** (2021), 471–489. https://doi.org/10.3390/epidemiologia2040033

15. A. Zeroual, F. Harrou, A. Dairi, Y. Sun, Deep learning methods for forecasting COVID-19 time-series data: A comparative study, *Chaos Solitons Fractals*, **140** (2020), 110121. https://doi.org/10.1016/j.chaos.2020.110121

16. F. Shahid, A. Zameer, M. Muneeb, Predictions for COVID-19 with deep learning models of lstm, gru and bi-lstm, *Chaos Solitons Fractals*, **140** (2020), 110212. https://doi.org/10.1016/j.chaos.2020.110212

17. A. Fokas, N. Dikaios, G. Kastis, Mathematical models and deep learning for predicting the number of individuals reported to be infected with Sars-Cov-2, *J. R. Soc. Interface*, **17** (2020), 20200494. https://doi.org/10.1098/rsif.2020.0494

18. J. Long, A. Q. M. Khaliq, K. M. Furati, Identification and prediction of time-varying parameters of COVID-19 model: a data-driven deep learning approach, *Int. J. Comput. Math.*, **98** (2021), 1617–1632. https://doi.org/10.1080/00207160.2021.1929942

19. B. Ridenhour, J. M. Kowalik, D. K. Shay, Unraveling r0: Considerations for public health applications, *Am. J. Public Health*, **104** (2014), e32–e41. https://doi.org/10.2105/AJPH.2013.301704

20. Z. C. Chen, L. Feng, H. A. L. Lay, K. Furati, A. Khaliq, SEIR model with unreported infected population and dynamic parameters for the spread of COVID-19, *Math. Comput. Simul.*, **198** (2022), 31–46. https://doi.org/10.1016/j.matcom.2022.02.025

21. A. Hassan, I. Shahin, M. B. Alsabek, Covid-19 detection system using recurrent neural networks, in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, (2020), 1–5. https://doi.org/10.1109/CCCI49893.2020.9256562

22. G. Petneházi, Recurrent neural networks for time series forecasting, preprint, arXiv:1901.00069.

23. H. Hewamalage, C. Bergmeir, K. Bandara, Recurrent neural networks for time series forecasting: Current status and future directions, *Int. J. Forecast.*, **37** (2021), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

24. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

25. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., Learning phrase representations using rnn encoder-decoder for statistical machine translation, preprint, arXiv:1406.1078.

26. *Worlometer*, Coronavirus cases, 2021. Available from: https://www.worldometers.info/coronavirus/coronavirus-cases/

27. S. García, J. Luengo, F. Herrera, *Data Preprocessing in Data Mining*, Springer International Publishing, (2015), 1–17. https://doi.org/10.1007/978-3-319-10247-4_1

28. *ProgrammerSought*, General process and necessary steps of machine learning tasks, Available from: https://www.programmersought.com/article/98093557423/.

29. *M. Sharma*, Data preprocessing: 6 necessary steps for data scientists, Available from: https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa.

30. *D. Jain*, Data preprocessing in data mining, 2021. Available from: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/.

31. S. B. Kotsiantis, D. Kanellopoulos, P. E. Pintelas, Data preprocessing for supervised learning, *Int. J. Comput. Sci.*, **1** (2006), 111–117. https://doi.org/10.5281/zenodo.1082415

32. *Centers for Disease Control and Prevention*, Symptoms of COVID-19, 2021. Available from: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

33. S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, et al., The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application, *Ann. Intern. Med.*, **172** (2020), 577–582. https://doi.org/10.7326/M20-0504

34. J. A. Backer, D. Klinkenberg, J. Wallinga, Incubation period of 2019 novel coronavirus (2019-ncov) infections among travellers from Wuhan, China, 20–28 January 2020, *Eurosurveillance*, **25** (2020), 20–28. https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062

35. B. Everitt, A. Skrondal, *The Cambridge Dictionary of Statistics*, Cambridge University Press, 2010.

36. *S. Glen*, Mean absolute percentage error (MAPE), 2021. Available from: https://www.statisticshowto.com/mean-absolute-percentage-error-mape/.

37. R. G. D. Steel, J. H. Torrie, *Principles and procedures of statistics*, McGraw-Hill Book Company, 1960.

38. S. Glantz, B. Slinker, *Primer of Applied Regression and Analysis of Variance*, McGraw-Hill, 2001.

39. N. R. Draper, H. Smith, *Applied regression analysis*, John Wiley and Sons, 1998.

AIMS Press

© 2022 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)