# SEIR model combined with LSTM and GRU for the trend analysis of COVID-19

---

A Thesis

Presented to the Faculty of the Department of Mathematical Sciences

Middle Tennessee State University

---

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Mathematical Sciences

---

by

Lin Feng

Spring 2021

---

Thesis Committee:

Dr. Abdul Khaliq, Chair

Dr. Wandi Ding

Dr. Zachariah Sinkala

## APPROVAL

**This is to certify that the Graduate Committee of**

Lin Feng

met on the

[8th] day of [April, 2021].

The committee read and examined his/her thesis, supervised his/her defense of it in an oral examination, and decided to recommend that his/her study should be submitted to the Graduate Council, in partial fulfillment of the requirements for the degree of Master of Science in Mathematics.

*Dr. Abdul Khaliq*
Chair, Graduate Committee

*Dr. Zachariah Sinkala*

*Dr. Wandi Ding*

*Dr. James Hart*
Graduate Coordinator,
Department of Mathematical Sciences

*Dr. David Chris Stephens*
Chair,
Department of Mathematical Sciences

Signed on behalf of
the Graduate Council

*Dr. David Butler*
Dean,
School of Graduate Studies

## DEDICATION

Dedicated to my father Jianjun, my mother Qiuping, and my boyfriend Ziren for their love and support.

## ACKNOWLEDGMENTS

I would like to thank Prof. Abdul Khaliq for his guidance, encouragement and support over these years. I wish to thank Prof. James Hart, the director of the mathematical science program, for his help and guidance for my courses. I would like to thank Prof. Khaled Furati and Dr. Harold A. Lay, Jr., for their time on reviewing my thesis. I am grateful for my committee Prof. Zachariah Sinkala and Prof. Wandi Ding, who spent valuable time attending my thesis defense.

I would like to thank Prof. Rebecca Calahan for her patience and care.

I am very grateful to my family for their support and encouragement. I thank my father Jianjun and my mother Qiuping for their love and teaching. I wish to thank my boyfriend Ziren for his love and company.

## ABSTRACT

Coronavirus disease (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 Susceptible-Exposed-Infected-Removed (SEIR) model and deep learning methods, including Long-Short-Term-Memory (LSTM) and Gated Recurrent Units (GRU), have been employed for the analysis and prediction of COVID-19. The purpose of this thesis is to analyze and predict the epidemic trend of COVID-19 in different countries by combining the SEIR model with the classic LSTM and GRU methods, and to explore the application potential of LSTM and GRU in COVID-19 epidemic trend prediction.

The core content of this thesis consists of two parts. The first part is about the learning and prediction of dynamic parameters. The parameters in the SEIR model, including infection rate and recovery rate, are constantly changing over time, and can be considered as a time series. We learn and predict the dynamic changes of these two parameters over time using LSTM and GRU and find the constantly changing reproduction rate which is closely related to them. Then, we discuss and analyze the relationship between the reproduction number and the epidemic trend of COVID-19 by simple linear fit. In the second core part, we employ LSTM, GRU and SEIR models with the dynamic parameters that were learned and predicted by LSTM and GRU to do the prediction of the epidemic trend of COVID-19 for the United States. We utilize three common error metrics, Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and $r_2$ score, to compare and study the results and explore the application potential of LSTM and GRU in COVID-19 prediction. Mathematical software, like Python, are used in this investigation.

# CONTENTS

# List of Tables

# List of Figures

# CHAPTER 1

# INTRODUCTION

7In early December 2019, the first case of Coronavirus 2019 (COVID-19 [1]) was reported in Wuhan, Hubei Province of China. Then 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. COVID-19 is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus Type 2 (SARS-CoV-2). The COVID-19 poses a continuous threat to human health with its high transmission rate, serious infection consequences, and changing genetic makeup.

With the continuous spread and mutation of COVID-19, a big challenge of researchers has been witnessed in several science areas to help slowdown or avoid the increasing trends of its spread. Various models, estimation methods, and forecasting approaches have been introduced to help people understand and manage this pandemic. [3] Susceptible-Exposed-Infectious-Recovered model(SEIR) is one of the most commonly used and convincing mathematical methods. However, due to the continuous mutation of the virus and the differences in the response measures of people and governments in different periods, the parameter estimation problem of the SEIR model has become a major problem faced by many researchers. Nevertheless, many parameter estimation methods for SIR/SEIR model have been proposed and applied to COVID-19 data. For example, Bentout et al. [4] mentioned in their article about COVID-19, they use least squares to estimate the epidemic parameter and the basic reproduction number $R_0$. Oliveira et al. [5] mentioned in their article that the Bayesian method (MCMC) is used to estimate the parameters of the SIR model. These are all statistical methods, which are already relatively mature systems that are often used.

In recent years, as people continue to explore the field of machine learning, they have discovered that machine learning can be applied in many fields. Some facts have

proved that machine learning has superhuman capabilities in many fields. [6] With an attitude of continuous exploration and innovation, many scholars have adopted machine learning methods to analyze and predict the epidemic trend of COVID-19. Some RNN methods such as LSTM and GRU are the most commonly used and well-performing machine learning methods, because COVID-19 related data belongs to time-related sequential data, which is what we often call 'time series'. In their article, Zeroual et al. [7] compared five common machine learning methods, including LSTM and GRU, to study and predict the number of new and recovered cases. In this article by Shahid et al. [8], five machine learning methods including LSTM and GRU are compared and evaluated through time series forecasting of population, death and recovery in ten major countries affected by COVID-19.

In this investigation, we learn and predict the dynamic changes of these two parameters over time by LSTM and GRU, and find the constantly changing reproduction rate which is closely related to them. Then, we discuss and analyze the relationship between the reproduction number and the epidemic trend of COVID-19 by simple linear fit. We use LSTM, GRU, SEIR model with dynamic parameters to predict the active cases and removed cases of COVID-19. We use three common evaluation indicators, RMSE, MAPE and $R^2$, to compare and study the four results obtained and explore the shortcomings and application potential of LSTM and GRU in COVID-19 prediction.

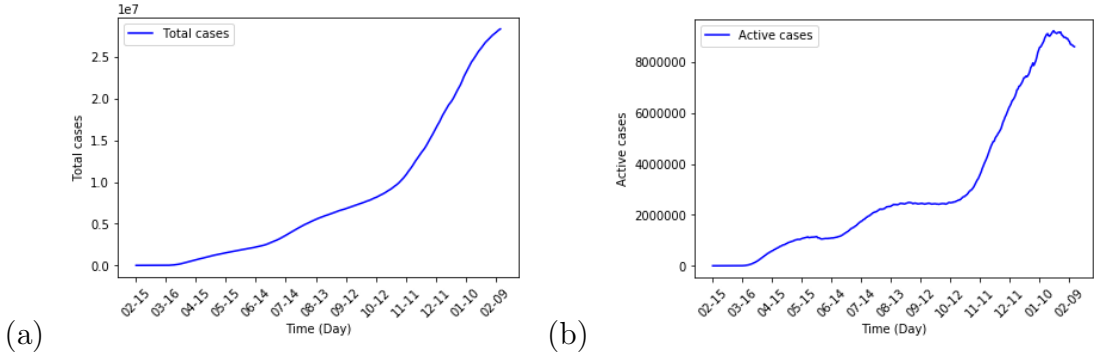
Chapter 2 introduces the data to be used in this thesis and its processing; Chapter 3 systematically introduces the SEIR model; Chapter 4 introduces the two deep learning methods of LSTM and GRU and focuses on the dynamic parameter learning and prediction of SEIR model by LSTM and GRU; Chapter 5 presents the numerical solutions for SEIR model with the dynamic parameters and gives an evaluation of the four models and methods; Chapter 7 is the conclusion of the thesis and the discussion of future research directions.

# CHAPTER 2

# DATA

## 2.1 Data collection

Our data comes from the website *'worldometer'*. We download the total cases and currently infected cases data from February 15, 2020 to February 14, 2021 for the United States (USA), which are shown in Figures 1 (a) and (b), respectively.




(a) (b)

Figure 1: (a) Total cases; and (b) Active case for the USA between Feb 15, 2020 and Feb 14, 2021.

## 2.2 Data preprocessing

Data preprocessing is a data mining technique of great importance to data scientists in the process of their projects [9] [10] [11] [12]. Data preprocessing is usually used to transform the original data into a more efficient and useful data format, which is conducive to better realization of the subsequent data analysis process. The original data we get may have missing values or may contain a lot of noise, which is very unfavorable for the training of the model. Therefore, data preprocessing is a necessary step before any data analysis. Moreover, sometimes different data preprocessing

methods are needed in order to meet different purposes or different algorithms. Data preprocessing has a significant impact on the generalization performance of LSTM and GRU that are used in this thesis. [13]

Next, we perform a simple pre-processing on the raw data to make it initially meet our algorithm requirements.

### 2.2.1 Left censoring

It has been observed that no matter which country it is, there is a phenomenon of data censorship in the early stage of data reporting.

We think that at the beginning of COVID-19, the monitoring and reporting system was not complete/perfect, which led to incomplete information collection or censored information. In order to reduce the impact of information left censorship 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 assume that the monitoring and reporting system will be more complete two months after the outbreak starts, so we choose April 15, 2020 as the new start time for each country.

### 2.2.2 Right censoring

With the introduction of the COVID-19 vaccine, the epidemic situation in many places has been brought under control. But not all are effective controls, because the vaccine at this stage is immature, and the virus has not stopped its changes and aggression.

We regard the vaccine as a changing factor. It will affect our data and cause right censorship. In order to reduce the impact of right censoring, we decided to abandon the data after the vaccine was produced. Most countries began to popularize vaccines from mid to late December. The accumulation rate is not very high due to the influence of factors such as early production and effect. Therefore, we choose the

end of December (Dec $31^{st}$) as our stay time point for this study.

### 2.2.3 Derived removed data

Now, we have the data of total cases and active cases for the USA from April 15, 2020 to December 31, 2020. We still need the total removed cases (including the recovered and death) in this investigation. We know that the total cases at time $t$ is all infected cases from the outbreak of COVID-19 to time $t$ and the active cases is the currently infected cases. It is obviously that the difference between them is the individuals that who have been infected but removed now, which are the removed cases. That is,

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

The derived removed data are presented in Figure 2.

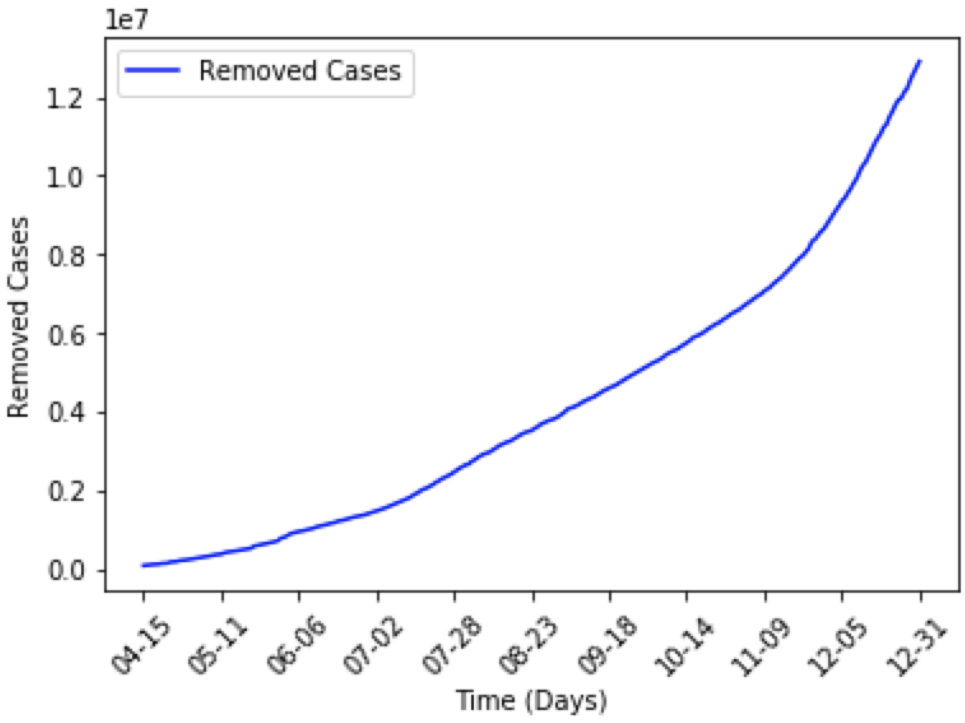



Figure 2: Derived removed cases for USA from April 15, 2020 to Dec 31, 2020

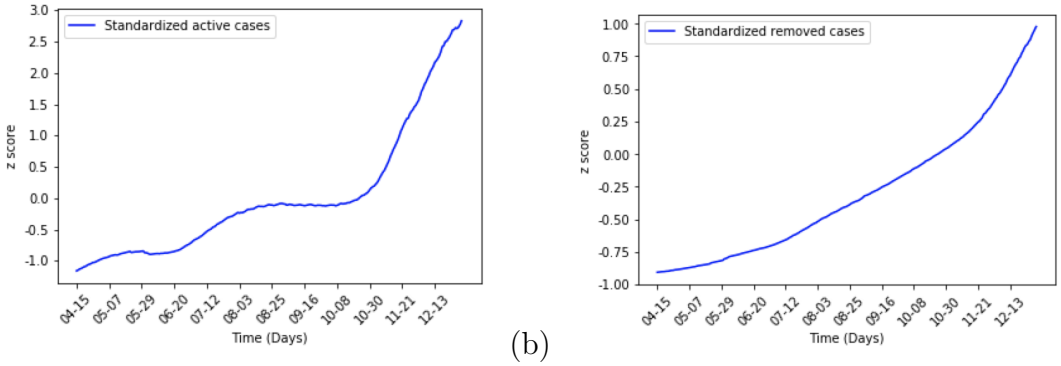
### 2.2.4 Data standardization

In machine learning, data standardization can indirectly avoid the impact of outliers and extreme values in the data on the training process in a centralized manner. Therefore, when there are outliers or a lot of noise in the data, we can reduce its impact through data standardization.

Here we choose the z-score standardization method to standardize the data. The mean and standard deviation of the processed data is 0 and 1, respectively. The data standardization formula is:

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

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




(a) (b)

Figure 3: Standardized (a) active cases; and (b) removed cases for USA between April 15, 2020 and Dec 31, 2020.

Z-score standardization is also called standard deviation standardization. It is also easy to understand the meaning of its name from the conversion formula. If given the question: how many standard deviations are the data from the mean value of the whole data, then the data greater than the mean will have a positive standardized score, otherwise, the data that is less than the mean will receive a negative

standardized score. Figure 3 presents the data of active cases and removed cases that after standardization.

## 2.3 Training and test data selection

The epidemic situation in different countries is affected by many different external factors at different stages, such as the reporting rate in different periods, different measures to respond to the epidemic in different periods, population movement measures, etc., and we did not set these when we established and operated the model. Corresponding parameters of influencing factors, our model is not suitable for prediction of particularly long-term data. Given these factors, we do not intend to use the commonly used 80%-20% division method to establish training data and test data but instead use the first 240 sets of data as training data to train the model. We then use the obtained model to predict the nearly three weeks (21 days) of remaining observations.

# CHAPTER 3

# SUSCEPTIBLE-EXPOSED-INFECTIOUS-REMOVED (SEIR) MODEL

## 3.1 Introduction

As we all know, the most important thing is to understand and analyze the rate of spread and trend of a disease during a pandemic. Only when we have a sufficient understanding of the spread of the pandemic can we propose targeted measures to slow it. These have more or less impact on public health policies, such as isolation.

Mathematical modeling of epidemic diseases helps to better understand the underlying mechanisms that affect disease transmission. And in this process, corresponding control strategies will be provided according to the model and results.[14]

The susceptible-infectious-removed(SIR) model is one of the most popular mathematical models to estimate the spread of the pandemic, and the SEIR model is also particularly worthy of discussion.[15]

In this chapter, we will focus on the SEIR model, discuss its basic model structure and numerical solution, and give the expression and solution of the reproductive number based on this model.

## 3.2 Susceptible-Exposed-Infectious-Removed(SEIR) model

The SEIR model divides the population into four categories: susceptible individuals, exposed individuals, infectious individuals, and removed individuals, with the following assumptions:

1) The population dynamics such as birth, nature death, and mobility are not considered.

2) Removed individuals will not be infected again.

3) Exposed individuals can not be infectious. In another word, the infectious group is the only group that can be infectious.

At the very beginning of a pandemic, the number of susceptible individuals is highest because the number of infected individuals is very small at the beginning. On the other hand, the number of infectious individuals is at its lowest during the beginning of a pandemic. The number of susceptible individuals has decreased as the time goes by, but the number of infectious individuals has increased. The changes could be reflectd by the following differential equations and can be represented by the flow chart in Figure 4.

$$\frac{dS(t)}{dt} = -\frac{\beta S(t)I(t)}{N} \tag{1}$$

$$\frac{dE(t)}{dt} = \frac{\beta S(t)I(t)}{N} - \sigma E(t) \tag{2}$$

$$\frac{dI(t)}{dt} = \sigma E(t) - \gamma I(t) \tag{3}$$

$$\frac{dR(t)}{dt} = \gamma I(t) \tag{4}$$

where

$$S(t) + E(t) + I(t) + R(t) = N \tag{5}$$

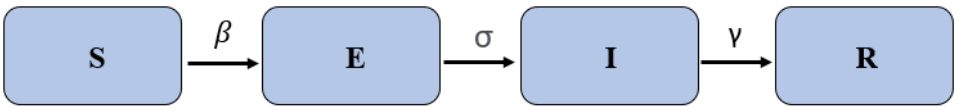



Figure 4: Flow chart of SEIR model

$S$ represents the susceptible individuals, which is the population that could be infected. At the beginning of outbreak, we can assume almost all the population is

susceptible, because the infectious individual is very small compared to the whole population at the initial break out time.

$E$ represents the exposed individuals, which is the population has been infected but does not show symptoms. It can be called an incubation period/latent period.

$I$ presents the infected individuals and is the infected population after the incubation period.

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

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

$\sigma$ is 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. In the SEIR model, $\sigma$ is the parameter that transport people from the exposed group $E$ to the infectious group $I$.

$\gamma$ is removed rate, which is the summation of the recovery rate and the death rate for the disease. In the SEIR model, $\gamma$ is the parameter that transports people from the infectious group $I$ to the removed group $R$.

$S(t)$, $E(t)$, $I(t)$, and $R(t)$ are the varying susceptible, exposed, infected, and removed individuals, respectively.

Furthermore, it is obviously for the model above, that

$$\frac{dN}{dt} = \frac{d(S + E + I + R)}{dt} = 0 \tag{6}$$

## 3.3 Basic reproduction number ($\mathbf{R}_0$)

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 infection and without the influence of external forces. In the SEIR model, $R_0$ can be calculated by [17]

$$R_0 = \frac{\beta}{\gamma} \tag{7}$$

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

**(1) The significance of the exploration of $R_0$**

There are three different conditions that indicate the possible transmission or decline of a disease based on the value of $R_0$:

1) $R_0 < 1$: each infected individual infects less than one new indiviudal, 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 will stay alive and keep in a stable status.

3) $R_0 > 1$: each infected individual infects more than one new individual, which implies the disease will keep transmitting between individuals, and it may cause an outbreak or epidemic.

**(2) The limitation of $R_0$**

From the definition, we can see that $R_0$ lets us know the average number of new infections from people who have the disease. It is suitable for people who have not previously been infected and have not been vaccinated. For example, if $R_0 = 15$ for some disease in an area, then a people who has been infected with the disease will transmit the disease to another 15 cases on average. The transmission will repeat in the area if no one has been vaccinated against and immunized against the disease.

Once the immune system is established or people's contact rate is reduced due to the influence of external forces, then $R_0$ will change. Therefore, the research and exploration of $R_0$ and the situation reflected are more suitable for early stages of the epidemic.

To summarize, the value of $R_0$ of the disease is only applicable when everyone in the population is completely susceptible to the disease. It can be done in the following situations:

1) No vaccine;

2) No one suffered from the disease;

3) No way to control the spread of the disease;

However, with the development of science and technology and the advancement of medical standards, the above-mentioned situation is rarely seen. This situation will be broken by external forces. So strictly speaking, $R_0$ only applies to the initial stage of the outbreak.

### 3.3.1 Current reproduction number($\mathbf{R}_t$)

To put it simply and understandably, effective reproduction number $R_t$ is the reproduction number at time $t$. The parameters of the traditional SEIR model are identified as a constant, and the initial reproduction number obtained from this is a constant that is only applicable to the initial stage of an epidemic. But in fact, the parameters of the SEIR model and reproduction number are time-related parameters. Assuming that $\beta_t$ and $\gamma_t$ are the value of $\beta$ and $\gamma$ at time $t$, respectively, then the current reproduction number in the SEIR model is expressed as

$$R_t = \frac{\beta_t}{\gamma_t} \tag{8}$$

## 3.4 Numerical solutions for SEIR model

### 3.4.1 Numerical solutions

Let the time step be one day, by the forward Euler's method, we can get the algorithm for the numerical solution of SEIR model as follows:

**Algorithm** Numerical solutions of SEIR model

**Input**
Local population of the surveyed area: $N$;
The number of days: $n$;
The initial value of variables: $S_0$, $E_0$, $I_0$ and $R_0$;
Parameters: $\beta$, $\gamma$ and $\sigma$;
**Output**
$S = \{S_0, S_1, ..., S_n\}$, $E = \{E_0, E_1, ..., E_n\}$, $I = \{I_0, I_1, ..., I_n\}$ and $R = \{R_0, R_1, ..., R_n\}$

**Procedure**
**For** $i$ in 0 to $n-1$
$S_{i+1} = S_i - \frac{\beta S_i I_i}{N}$
$E_{i+1} = E_i + \frac{\beta S_i I_i}{N} - \sigma E_i$
$I_{i+1} = I_i + \sigma E_i - \gamma I_i$
$R_{i+1} = R_i + \gamma I_i$
**end**

There are many other ODE dolver in Python, such as GEKKO Python and ODEINT function. The function ODEINT requires four inputs for the solution of SEIR model:

$$y = odeint(model, initialconditions, T, args) \tag{9}$$

'Model' gives the SEIR differential equation system that we generated by equation (1), (2), (3) and (4); 'initial conditions' give the initial value of the variables $S$, $E$, $I$ and $R$; 'T' is a sequence of time points for which to solve for the variables; 'args' is the extra arguments to pass to function, here is the vector that gives the value of the parameter $\beta$, $\sigma$ and $\gamma$ and the total local population $N$.

Here is an example of numerical solutions of SEIR model.

Assume that $N = 10000$, $\beta = 0.3$, $\sigma = 1/7$, $\gamma = 0.1$, $I_0 = 1$, $E_0 = 0$, $R_0 = 0$, then $S_0 = N - I_0 - E_0 - R_0$. Set the total time be $t = 300$ days, then plug these information to Euler's method or the odeint function, we can get the numerical solutions of the SEIR model as:

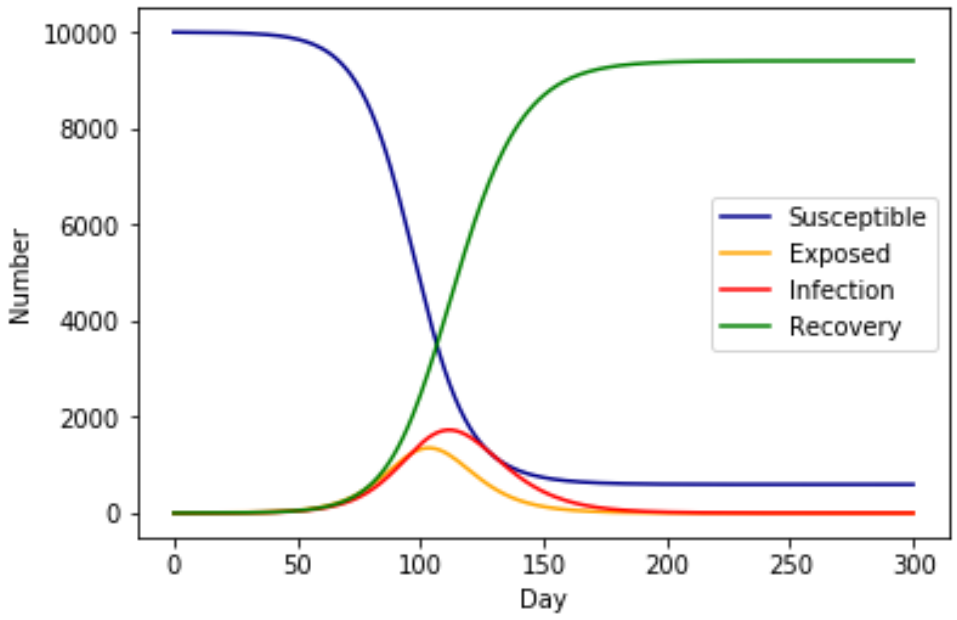



Figure 5: Numerical solutions for the SEIR example

The system of ordinary differential equations has been well developed, and there are many existing methods to find the numerical solution of the system of ordinary differential equations. The traditional SEIR model we mentioned is a relatively simple system of ordinary differential equations. For example, MuhammadFarman et al.[20] and BijilPrakash et al.[21] have provided good numerical solutions for the SEIR model in their articles.

### 3.4.2 Challenges of parameter estimation

In last section, we introduced several numerical solutions and examples of SEIR models in Python. We know that we need to enter the value of each parameter to solve a SEIR model. In the example, we randomly selected some parameter values to find out the numerical solution of the SEIR model, but when we apply the SEIR model to actual cases, such as the COVID-19, we need to find a method to estimate the parameters for solving the model, because we do not know the approximate value of the parameter.

Parameter estimation is very important for the numerical solution of SEIR model, because they can directly affect the accuracy of the results. As mentioned in introduction, we know that many researchers use statistical methods, such as least squares [4] and Bayesian method (MCMC) [5] to estimate the parameters of the SEIR model. These are already relatively mature systems that are often used. The parameters estimated by these methods are generally the optimal parameters that can meet the current epidemic trend, and they are all definite values.

In fact, the parameters of the SEIR model, including transmission rate $\beta$, incubation rate $\sigma$, and removed rate $\gamma$, are all time-dependent due to the effects of the factors such as reporting rate, government policies, and medical effects. In this case, the parameters that estimated using general estimation methods can only be applied to solve for the short-term changes in the epidemic, and the results of long-term simulations will deviate due to changes in parameters. We will propose a reasonable methods to improve this problem in the next chapter.

## 3.5 Summary

In this chapter, we introduce the SEIR model and its several numerical solutions in Python. It also briefly explained and analyzed regarding the changes in the reproduction number and the outbreak of the epidemic under the SEIR model. In section

3.4.2, we put forward the parameter estimation problem related to the SEIR model, and we give the corresponding solution in Chapter 4.

# CHAPTER 4

# DYNAMIC PARAMETER LEARNING AND PREDICTION OF SEIR MDOEL BY DEEP LEARNING METHODS: LSTM & GRU

## 4.1 Introduction

In order to solve the problem of SEIR parameters proposed in Chapter 3, in this chapter, instead of using statistical methods for parameter estimation, we propose a method of learning and predicting SEIR model parameters using LSTM and GRU.

In this chapter, we bring the real active cases and removed cases into the discretized SEIR model to solve the theoretical values of the parameters transmission rate $\beta$ and removed rate $\gamma$ at each moment. Then we apply the two methods of LSTM and GRU in deep learning to do the prediction of these two parameters and calculate the reproduction number$R_t$ corresponding to each moment. We briefly analyzed the relevant situation through the relative changes of transmission rate and removed rate. Based on the results, we conducted a systematic analysis of the epidemic trend of COVID-19 in the United States during the investigated period through the analysis of current reproduction rate $R_t$.

### 4.1.1 Artificial neural networks

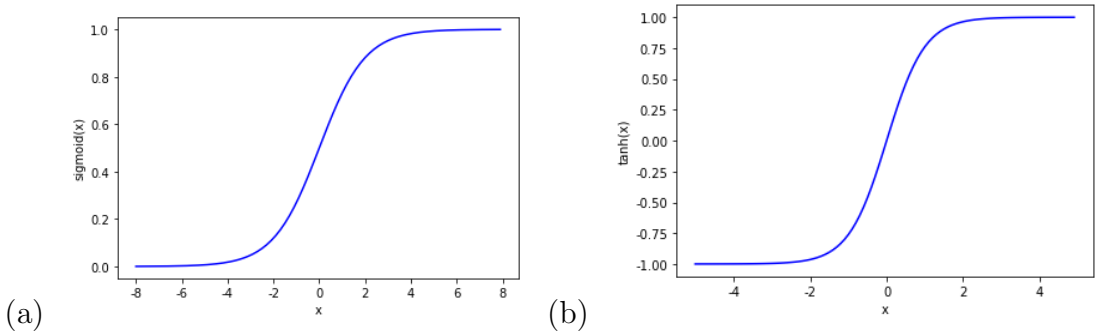
Artificial Neural Network (ANN) is a deep learning algorithm, which is based on the idea of the human brain's biological neural network. ANN tries to simulate the operation of the human brain. Its working principle is very similar to that of biological neural networks but not completely similar.

An activation function is used to introduce nonlinearities information combination in an artificial neural network. It allows us to model nonlinear relationships and helps us understand complex data. The activation functions used in this article are sigmoid

function and hyperbolic tangent function (tanh).

**Sigmoid function** The *sigmoid* activation function maps the input values to the range $(0, 1)$, which can be regarded as the probability of belonging to a certain category or the weight that reflects the importance of the information. Equation (10) presents the expression of the function, and Figure 6 (a) provides its image.

$$Sigmoid(x) = \frac{1}{1 + \exp(-x)} \tag{10}$$




(a) (b)

Figure 6: Image of (a) sigmoid; and (b) tanh activation function

**Hyperbolic tangent function (tanh)** The *tanh* activation function maps the inputs to the $(-1, 1)$ range. Compared with the sigmoid function, it provides a zero-centered output. Equation (11) presents the expression of the function, and Figure 6 (b) provides its image.

$$tanh(x) = \frac{2}{1 + \exp(-2x)} - 1 \tag{11}$$

### 4.1.2 Recurrent Neural Network(RNN)

RNN is a kind of Artificial Neural Networks with memory. The reason why RNNs are called recurrent neural networks is that they can learn and save the past information, and then use it for future predictions.

Figure 7 presents the repeating module in an RNN. $x_t$ 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 $t$; $W_{h(t)}$ 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 under the action of two different weight matrices. This vector contains the current input information and the previous memory, and the new hidden memory at time t is obtained under the activation of the activation function tanh. Then enter the next cell with the information at time t as input. The whole process can be represented by Equation (12), where $b$ is the bias.

$$h_t = tanh(W_{h(t)} * h_{t-1} + W_{x(t)*x_t} + b) \tag{12}$$

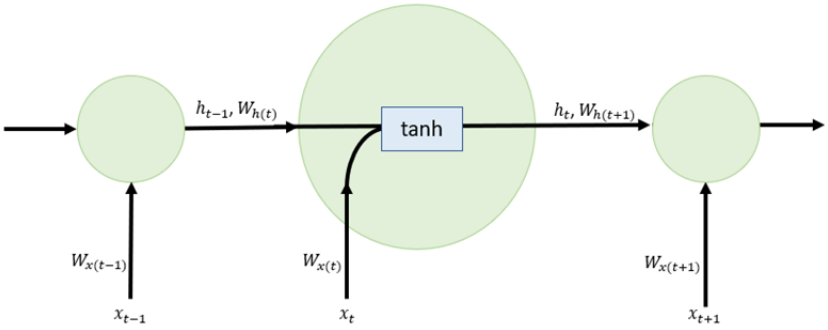



Figure 7: The repeating module in an RNN

RNNs are mainly used to do the sequential prediction problems.[22] [23] [24]. Thus, we use RNNs for time series processing.

However, due to the shortcomings of difficulty in training and difficulty in storing and obtaining long-term memory information, we usually do not use basic RNN methods when dealing with long sequence data. The most popular RNN methods that can solve these problems effectively are LSTM and GRU [26] [27].

## 4.2 Long-Short-Term-Memory (LSTM)

LSTM is a long-term short-term storage network used in the field of deep learning. It is a special recurrent neural networks (RNNs) 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. The cell state serves as a memory bank that runs through the entire sequence of processing. It can record relevant information during the entire sequence processing process and pass them on. It is responsible for storing and transferring the long-term information all the way down the sequence chain, which can be regard as the "memory" of the neural network. As the sequence processing progresses, new or old information are added or removed from the cell state via some gates. These gates can learn and decide what information can be added and stored or be forgot and removed during the training. Figure 8 presents the repeating module for an LSTM. There are total of four gates in the repeating module of an LSTM: forget, input, cell, and output gate, respectively.

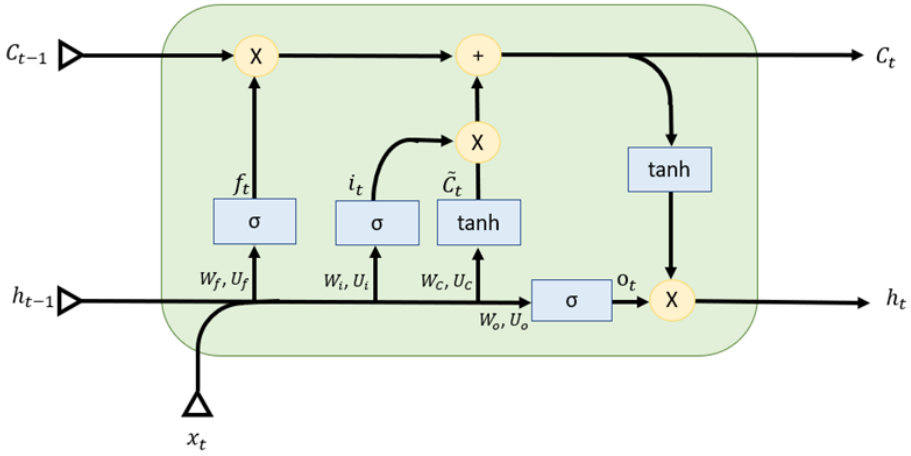



Figure 8: The repeating module(a cell) in an LSTM

The input of the repeating module of an LSTM at time $t$ includes the input of time $t(x_t)$, the output of last cell($h_{t-1}$), which brings short-term memory and the cell state($C_{t-1}$) from previous cell, which keeps the long-term memory. In the diagram

of the repeating module of the LSTM, the blue box represents the active 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" represents forget gate, input gate, cell gate and output gate, respectively. When $x_t$ and $h_{t-1}$ enters each gate, they will combine the information through the corresponding weighted matrix. For example, when $x_t$ and $h_{t-1}$ enters 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 a sigmoid layer called the "forget gate". The inputs of the gate are $h_{t-1}$ and $x_t$, and output is a weight(0-1) matrix of the cell state $C_{t-1}$, where '1' represents "completely keep" and '0' represents "completely get rid of".

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{13}$$

**(2) Input Gate**

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

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{14}$$

$$\widetilde{C}_t = tanh(W_C x_t + U_C h_{t-1} + b_C) \tag{15}$$

**(3) Cell State**

In this step, we update the old cell state $C_{t-1}$ by the information we got from 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 the preservation and discarding of the old information. 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 and record it in the cell state.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{16}$$

**(4) Output Gate**

Finally, we need to decide what are going to 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 a sigmoid layer, where "1" represents outputing all information and "0" represents nothing will be output. Then, we push the values of the cell state to be between $-1$ and 1 through a $tanh$ function and then multiply it by the weighted matrix to output the parts of the cell state ($h_t$) we decided to.

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{17}$$

$$h_t = o_t * tanh(C_t)) \tag{18}$$

## 4.3 Gated Recurrent Unit (GRU)

GRU is a variant of LSTM, it combines the forget and input gates into a single "update gate" and it also merges the cell state and hidden state, keeping the long-term and short-term information together. Therefore, GRU is more efficient compared with the traditional LSTM. For data learning and prediction capabilities, their performance will vary due to different data.

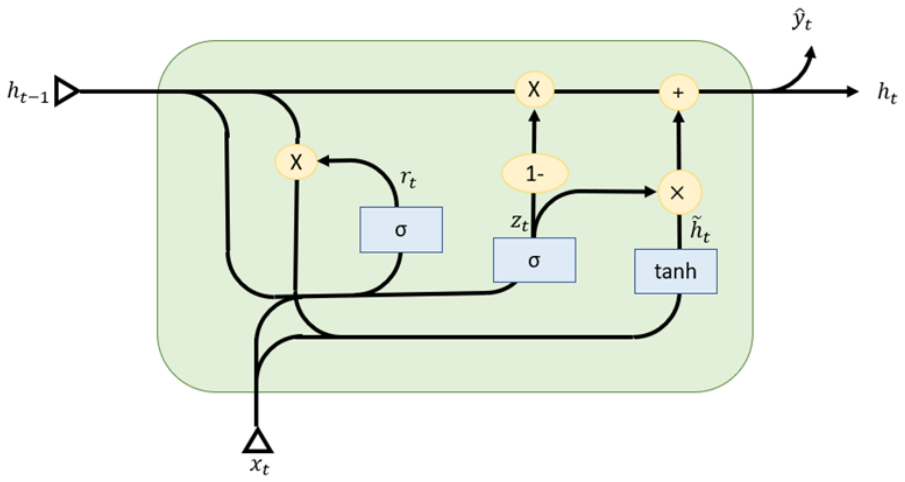



Figure 9: The repeating module in an GRU

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{19}$$

$$r_t = \sigma(W_r x_r + U_r h_{t-1} + b_r) \tag{20}$$

$$\widetilde{h_t} = tanh(W x_t + U(r_t * h_{t-1}) + b_{\widetilde{h}}) \tag{21}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \tag{22}$$

## 4.4 Implementation

### 4.4.1 Theoretical/Numerical solutions of dynamic parameters

As mentioned before in Chapter 3, the standard SEIR model can be expressed by the equations system (1), (2), (3) and (4). In a more real situation, in order to solve the parameter problem mentioned in Chapter 3, we consider the dynamic parameters $\beta(t)$ and $\gamma(t)$ instead of fixed value of $\beta$ and $\gamma$, we can drive the equation system of SEIR model with dynamic parameters as

$$\frac{dS(t)}{dt} = -\frac{\beta(t)I(t)S(t)}{N} \tag{23}$$

$$\frac{dE(t)}{dt} = \frac{\beta(t)I(t)S(t)}{N} - \sigma E(t) \tag{24}$$

$$\frac{dI(t)}{dt} = \sigma E(t) - \gamma(t)I(t) \tag{25}$$

$$\frac{dR(t)}{dt} = \gamma(t)I(t) \tag{26}$$

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

From our previous explorations, we know that the S(susceptible), E(exposed), I(infected), removed(R), $\beta$(infected rate), $\gamma$(recovered rate) and $\sigma$(incubation rate) are all time-dependent variables. Since our model is based on the SEIR model without vital dynamics, we know that the total population of the area is the summation of $S$, $E$, $I$ and $R$ at any time. To simplify the model, we set the incubation rate to be a constant. The incubation period for coronavirus disease 2019 is 2-14 days.[40] Stephen A. et. al. [41] concluded that 5.1 days (95% CI, 4.5 to 5.8 days) is the median incubation period and 97.5% of people will show symptoms within 11.5 days (CI, 8.2 to 15.6 days) of infection. Jantien A et. al. [42] used Weibull distribution to fit the data, getting the range of the incubation period is from 2.1 to 11.1 days with the mean 6.4 days (95% CI: 5.6 to 7.7 days). Here, we choose 6 days as the incubation period. Then the incubation rate $\sigma$ is 1/(incubation period)=$\frac{1}{6}$.

In order to obtain the reasonable values of the other parameters, we use the Forward Euler Method to discretize the ODE system of 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:

$$S_{t+1} = S_t - \frac{\beta_t S_t I_t}{N} \tag{27}$$

$$E_{t+1} = E_t + \frac{\beta_t S_t I_t}{N} - \sigma E_t \tag{28}$$

$$I_{t+1} = I_t + \sigma E_t - \gamma_t I_t \tag{29}$$

$$R_{t+1} = R_t + \gamma_t I_t \tag{30}$$

with the symbol interpretation in the below table:

| 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 $t$ |
| $R_t$ | individuals have been infected, then removed at time $t$ |
| $\beta_t$ | Transmission rate at time $t$ |
| $\gamma_t$ | Remove rate at time $t$ |
| $\sigma$ | Incubation rate |
| $N$ | the total local population |

Table 1: Symbol explanation for discrete SEIR model

From this system, it is not difficult to find that 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,

$$N = S_t + E_t + I_t + R_t \tag{31}$$

$$= S_{t+1} + E_{t+1} + I_{t+1} + R_{t+1} \tag{32}$$

At the same time, our basic assumption is verified, which is a non-dynamic basic SEIR model. A complete discrete SEIR model with dynamic parameters can be expressed by the equations (27), (28), (29), (30), (31) and (32).

And the current reproduction number at time $t$ is represents by

$$R_t = \frac{\beta_t}{\gamma_t} \tag{33}$$

Now, we have real data for all time of active cases and removed cases. We have determined the incubation rate based on the experiences. We also know the total population of the country we are studying. In other words, in this system of equations, we know $R_t$, $R_{t+1}$, $I_t$, $I_{t+1}$, $\sigma$ and $N$, and the unknown variables are $\beta_t$, $\gamma_t$, $S_t$, $S_{t+1}$, $E_t$ and $E_{t+1}$. For the six equations of this system, we have six unknowns, so we can get the theoretical values or numerical solutions of the parameters $\beta$ and $\gamma$ by solving the equation system. And the algorithm are showed below:

**Algorithm** Theoretical solutions of dynamic parameters of SEIR model

**Input:**
Local population of the urveyed area: $N$;
The number of iterations or days: $n$;
All sequencial value of the variable I and R from $t = 0$ to $t = n$:
$I = \{I_0, I_1, ..., I_n\}$, $R = \{R_0, R_1, ..., R_n\}$;
The 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** $i$ in 0 to $n-1$, do
    $\gamma_t = \frac{R_{t+1} - Rt}{It}$
    $E_t = \frac{I_{t+1} - I_t + \gamma_t I_t}{\sigma}$
    $S_t = N - E_t - I_t - R_t$
**For** $i$ 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** $i$ in 0 to $n-2$, do  $R_t = \frac{\beta_t}{\gamma_t}$

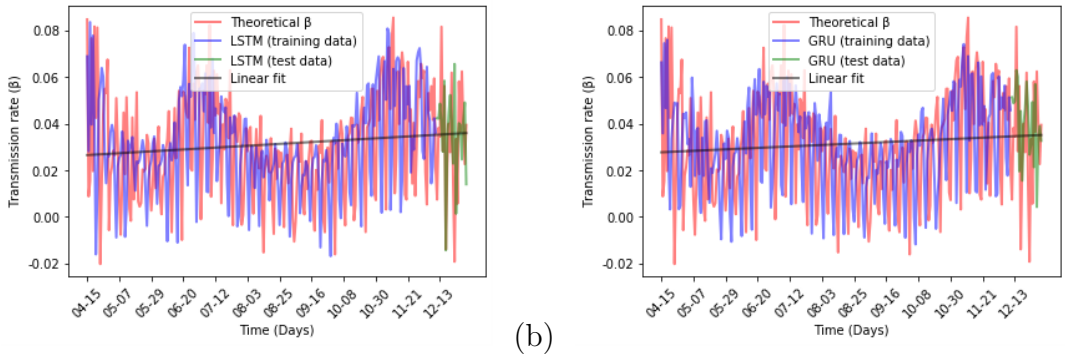
After obtaining the theoretical dynamic transmission rate and removed rate from April 15, 2020 to Dec 15, 2020 using the algorithm given above, separate the two obtained sequences into training data (the first 240 data) and test data (the remaining data of three weeks, that is, 21 days) . After that, put the training data as input for the LSTM and GRU models mentioned in section 4.2 and 4.3 and perform the

prediction for the next three weeks, by which we can get the predicted value of $\beta$ and $\gamma$. Then, we can use Equation (8) to find out the effective reproduction number ($R_t$) for each day from April 15, 2020 to Dec 31, 2020.

We can get the theoretical and predicted values of all parameters by the above process, and we combine the results for comparison and analysis. We can then perform regression analysis on the predicted results and theoretical results to compare and explore the changes and trends of these parameters over time. Furthermore, we explore and analyze the epidemic trend of COVID-19 for USA through the analysis of the changes of the parameters.

## 4.5 Results and analysis

### 4.5.1 Results of transmission rate ($\beta$)




(a) (b)

Figure 10: Track of $\beta$ for USA between April 15, 2020 and Dec 31, 2020 by (a) LSTM; and (b) GRU. In these two graphs, the red curve presents the theoretical values of $\beta$ based on SEIR mdoel; the blue curve presents the predicted values for training dataset of $\beta$; the green curve presents the predicted values for test dataset of $\beta$; the black line presents the linear regression of predicted values of $\beta$.

From Figure 10, we can see that GRU's prediction of $\beta$ is slightly better than that of LSTM, which is more obvious in the early and mid-term of this period. Consider it in conjunction with Table a, b and c for $\beta$, we know the slope of the linear regression for the theoretical $\beta$ is $2.824(10)^{-5}$, and the slope of the linear regression for the predicted $\beta$ by LSTM and GRU are $3.63(10)^{-5}$ and $2.85(10)^{-5}$, respectively. Obviously, compared with LSTM, GRU's prediction of $\beta$ reflects the actual situation better. The results show that the transmission rate of USA shows a downward trend, but not obvious from Apr 15, 2020 to Dec 31, 2020.

Table a: Linear regression analysis for theoretical $\beta$

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0270 | 0.003 | 9.770 | 0.000 | 0.022 | 0.032 |
| x1 | 2.824e-05 | 1.84e-05 | 1.534 | 0.126 | -8.01e-06 | 6.45e-05 |

| | | | |
|---|---|---|---|
| Omnibus: | 2.492 | Durbin-Watson: | 1.476 |
| Prob(Omnibus): | 0.288 | Jarque-Bera (JB): | 2.244 |
| Skew: | 0.137 | Prob(JB): | 0.326 |
| Kurtosis: | 2.638 | Cond. No. | 300. |

Table b: Linear regression analysis for the $\beta$ that predicted by the LSTM

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0266 | 0.003 | 10.565 | 0.000 | 0.022 | 0.032 |
| x1 | 3.63e-05 | 1.67e-05 | 2.170 | 0.031 | 3.36e-06 | 6.92e-05 |

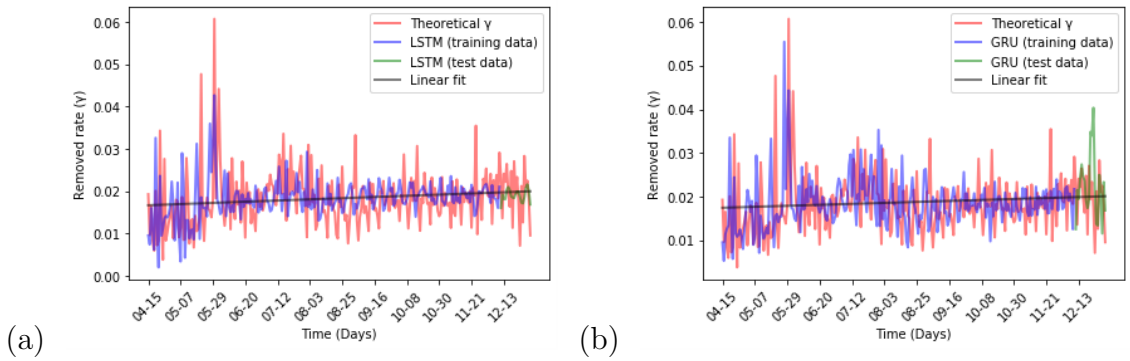
| | | | |
|---|---|---|---|
| Omnibus: | 0.994 | Durbin-Watson: | 1.195 |
| Prob(Omnibus): | 0.608 | Jarque-Bera (JB): | 1.009 |
| Skew: | 0.148 | Prob(JB): | 0.604 |
| Kurtosis: | 2.926 | Cond. No. | 300. |

Table c: Linear regression analysis for the $\beta$ that predicted by the GRU

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0277 | 0.002 | 11.622 | 0.000 | 0.023 | 0.032 |
| x1 | 2.85e-05 | 1.59e-05 | 1.794 | 0.074 | -2.79e-06 | 5.98e-05 |

| | | | |
|---|---|---|---|
| Omnibus: | 6.925 | Durbin-Watson: | 1.218 |
| Prob(Omnibus): | 0.031 | Jarque-Bera (JB): | 4.516 |
| Skew: | -0.161 | Prob(JB): | 0.105 |
| Kurtosis: | 2.442 | Cond. No. | 300. |

### 4.5.2 Results of removed rate ($\gamma$)




(a) (b)

Figure 11: Track of $\gamma$ for USA between April 15, 2020 and Dec 31, 2020 by (a) LSTM; and (b) GRU. In these two graphs, the red curve presents the theoretical values of $\gamma$ based on SEIR mdoel; the blue curve presents the predicted values for training dataset of $\gamma$; the green curve presents the predicted values for test dataset of $\gamma$; the black line presents the linear regression of predicted values of $\gamma$.

From Figure 11, we can easily see that compared to the GRU, the LSTM's prediction of $\gamma$ is much more stable than the theoretical value. Combining the linear regression analysis from Table d, e and f for $\gamma$, we know the slope of the linear regression for the theoretical $\gamma$ is $0.9943(10)^{-5}$, and the slop of the linear regression for the predicted $\gamma$ by LSTM and GRU are $1.267(10)^{-5}$ and $1.016(10)^{-5}$, respectively. Obviously, compared with LSTM, GRU's prediction of $\gamma$ reflects the actual situation better. The results show that the removed rate of USA shows a upward trend, but not obvious from Apr 15, 2020 to Dec 31, 2020.

Table d: Linear regression analysis for theoretical $\gamma$

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0170 | 0.001 | 20.507 | 0.000 | 0.015 | 0.019 |
| x1 | 9.943e-06 | 5.5e-06 | 1.807 | 0.072 | -8.93e-07 | 2.08e-05 |

| | | | |
|---|---|---|---|
| Omnibus: | 126.110 | Durbin-Watson: | 1.918 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 851.941 |
| Skew: | 1.818 | Prob(JB): | 1.01e-185 |
| Kurtosis: | 11.069 | Cond. No. | 300. |

Table e: Linear regression analysis for the $\gamma$ that predicted by the LSTM

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0167 | 0.001 | 30.861 | 0.000 | 0.016 | 0.018 |
| x1 | 1.267e-05 | 3.6e-06 | 3.516 | 0.001 | 5.57e-06 | 1.98e-05 |

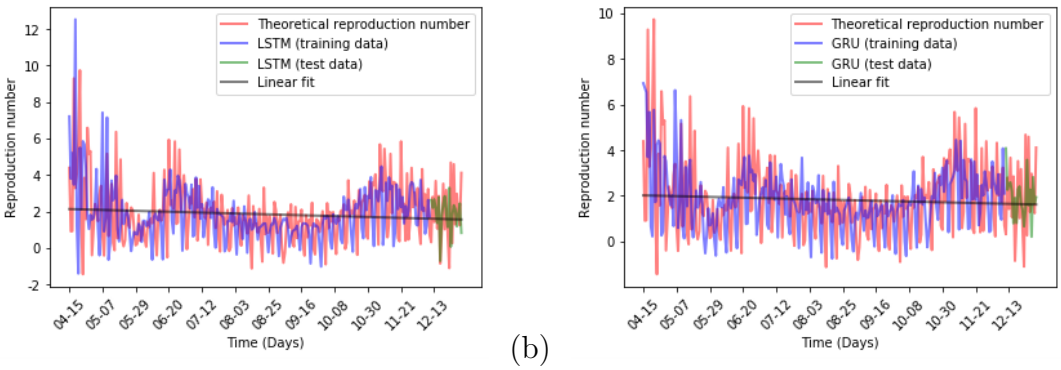
| | | | |
|---|---|---|---|
| Omnibus: | 87.155 | Durbin-Watson: | 1.537 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 556.819 |
| Skew: | 1.165 | Prob(JB): | 1.23e-121 |
| Kurtosis: | 9.765 | Cond. No. | 300. |

Table f: Linear regression analysis for the $\gamma$ that predicted by the GRU

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0175 | 0.001 | 23.698 | 0.000 | 0.016 | 0.019 |
| x1 | 1.016e-05 | 4.91e-06 | 2.069 | 0.040 | 4.89e-07 | 1.98e-05 |

| | | | |
|---|---|---|---|
| Omnibus: | 125.829 | Durbin-Watson: | 1.471 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 744.199 |
| Skew: | 1.870 | Prob(JB): | 2.51e-162 |
| Kurtosis: | 10.379 | Cond. No. | 300. |

### 4.5.3 Results of effective reproduction number ($R_t$)




(a) (b)

Figure 12: Track of $R_t$ for USA between April 15, 2020 and Dec 31, 2020 by (a) LSTM; and (b) GRU. In these two graphs, the red curve presents the theoretical values of $R_t$ based on SEIR mdoel; the blue curve presents the predicted values for training dataset of $R_t$; the green curve presents the predicted values for test dataset of $R_t$; the black line presents the linear regression of predicted values of $R_t$.

From Figure 12, we can easily see that compared to the GRU, the LSTM's prediction of $R_t$ is much more stable and lower than the theoretical value. Combining the linear regression analysis from Table g, h and i for $R_t$, we know the slope of the linear regression for the theoretical $R_t$ is $-0.0011$, and the slop of the linear regression for the predicted $R_t$ by LSTM and GRU are $-0.0022$ and $-0.0015$, respectively. Obviously, compared with LSTM, GRU's prediction of $R_t$ reflects the actual situation better. The results show that the effective reproduction number of USA shows a downwards trend, but it is not very obvious from Apr 15, 2020 to Dec 31, 2020. It shows that the epidemic in USA has improved during this time period.

Table g: Linear regression analysis for theoretical $R_t$

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 2.0493 | 0.206 | 9.928 | 0.000 | 1.643 | 2.456 |
| x1 | -0.0011 | 0.001 | -0.812 | 0.417 | -0.004 | 0.002 |

| | | | |
|---|---|---|---|
| Omnibus: | 56.791 | Durbin-Watson: | 1.874 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 113.836 |
| Skew: | 1.095 | Prob(JB): | 1.91e-25 |
| Kurtosis: | 5.381 | Cond. No. | 300. |

Table h: Linear regression analysis for the $R_t$ that predicted by the LSTM

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 2.1358 | 0.186 | 11.509 | 0.000 | 1.770 | 2.501 |
| x1 | -0.0022 | 0.001 | -1.776 | 0.077 | -0.005 | 0.000 |

| | | | |
|---|---|---|---|
| Omnibus: | 127.824 | Durbin-Watson: | 1.499 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 1012.668 |
| Skew: | 1.784 | Prob(JB): | 1.26e-220 |
| Kurtosis: | 11.966 | Cond. No. | 300. |

Table i: Linear regression analysis for the $R_t$ that predicted by the GRU

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 2.0227 | 0.166 | 12.198 | 0.000 | 1.696 | 2.349 |
| x1 | -0.0015 | 0.001 | -1.391 | 0.165 | -0.004 | 0.001 |

| | | | |
|---|---|---|---|
| Omnibus: | 31.101 | Durbin-Watson: | 1.360 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 44.427 |
| Skew: | 0.757 | Prob(JB): | 2.25e-10 |
| Kurtosis: | 4.338 | Cond. No. | 300. |

### 4.5.4 More statistical information

| Parameter | Data type | Mean | Variance |
|---|---|---|---|
| $\beta_t$ | Theoretical | 0.0306957 | 0.0005028 |
| | Prediction by LSTM | 0.0311968 | 0.0004248 |
| | Prediction by GRU | 0.0315283 | 0.0003758 |
| $\gamma_t$ | Theoretical | 0.0182474 | 4.5075719e-05 |
| | Prediction by LSTM | 0.0183418 | 1.9948614e-05 |
| | Prediction by GRU | 0.0188130 | 3.5919904e-05 |
| $R_t$ | Theoretical | 1.9042522 | 2.7817986 |
| | Prediction by LSTM | 1.8629908 | 2.3982446 |
| | Prediction by GRU | 1.8251087 | 1.7981305 |

Table 2: Mean and variance table for parameters of SEIR model in different situation

Table 2 shows the mean and variance of each parameter obtained in the three cases. It is obviously to see that the variance of the predicted parameters is smaller than the theoretical variance. From the perspective of the mean and variance, the mean and variance of the parameters predicted by LSTM are closer to the theoretical mean and variance.

## 4.6 Summary

In this chapter, we use LSTM and GRU to learn and predict the parameters of the SEIR model $\beta_t$, $\gamma_t$ and $R_t$, and combine regression analysis and other statistical methods to briefly explain and analyze the results. The results show that the variance of the parameter values predicted by LSTM and GRU is smaller than the theoretical value, that is, the predicted data fluctuates slightly less than the theoretical value. From the perspective of the overall trend, the prediction result of GRU is closer to the theoretical result; from the overall mean and variance, the prediction result of LSTM is closer to the theoretical value.

Through the analysis of the reproduction number, we also concluded that the epidemic situation in the United States during this period of time has eased, although

the magnitude is not large.

In this chapter, we briefly summarize the prediction results of LSTM and GRU by comparing the theoretical and predicted values of the parameters. But in fact, the theoretical values of the parameters we used here are calculated by the SEIR model, and there is no true value of the parameters in the strict sense. A summary based on this comparison is not necessarily convincing. But we know the values of real active cases and removed cases, and we can further evaluate the learning and prediction parameters of LSTM and GRU by comparing the real and predicted values of these two variables. More importantly, the learning and prediction of the parameters we did in this chapter are all serve for the SEIR model, and the solution of the SEIR model can really show that our method is good or not. We will continue to discuss about it in the next chapter.

# CHAPTER 5

# MODEL EVALUATION

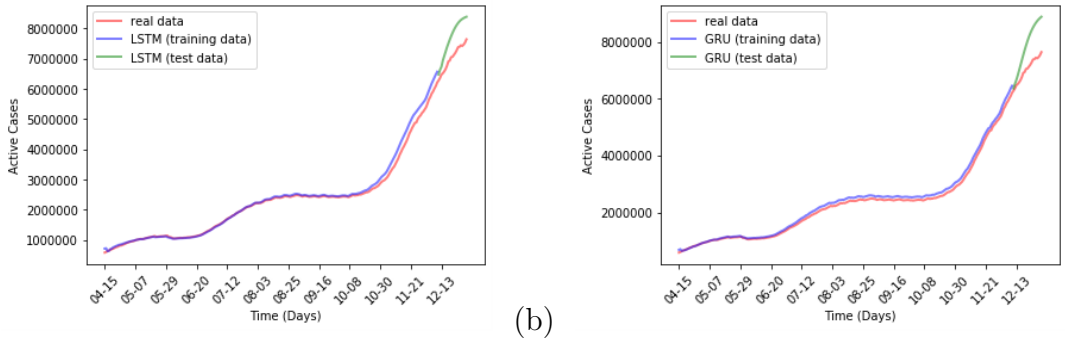
## 5.1 Introduction

Although training a model is essential in machine learning, it is equally important to judge the performance of a learned model on unseen data. Because we need to know whether the model we have learned is suitable for the data we are investigate and whether we can trust its predicted results. Model evaluation can help us judge the generalization accuracy of the model on out-of-sample data. [31] and help to find the best model that suitable for the data we focus on. Assessing the performance of a model with training data is usually not acceptable in data science because it is easy to generate overfitting models.[32]

In this chapter, we bring the parameters learned and predicted by the LSTM and GRU in the previous chapter back to the SEIR model to solve, and then the active cases and removed cases obtained are predicted using the value of these two variables. Then we utilize three common error metrics (RMSE, MAPE and $r_2$ score) to evaluate and analyze the true and predicted values of the test data part of active cases and removed cases to evaluate the performance of LSTM and GRU in parameter learning and prediction of SEIR model.

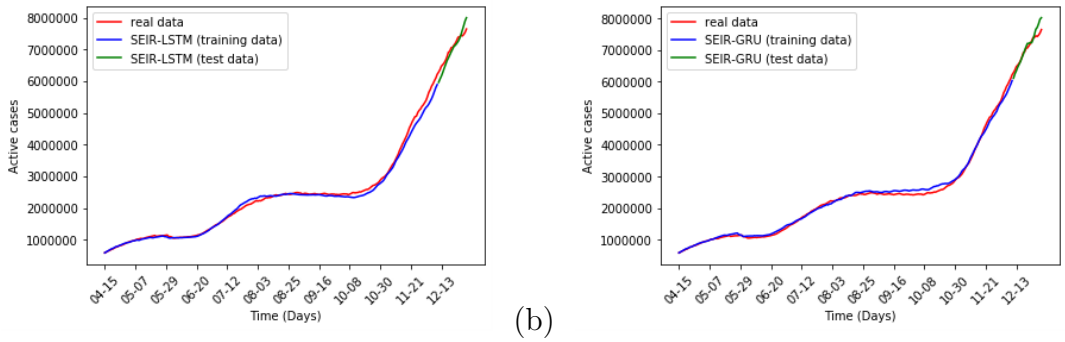
For the convenience of recording and thinking, we denote the SEIR model with dynamic parameters learned and predicted by LSTM as SEIR-LSTM; similarly, we denote the SEIR model with dynamic parameters learned and predicted by GRU as SEIR-GRU.

## 5.2 Numerical solutions for SEIR model using the dynamic parameters



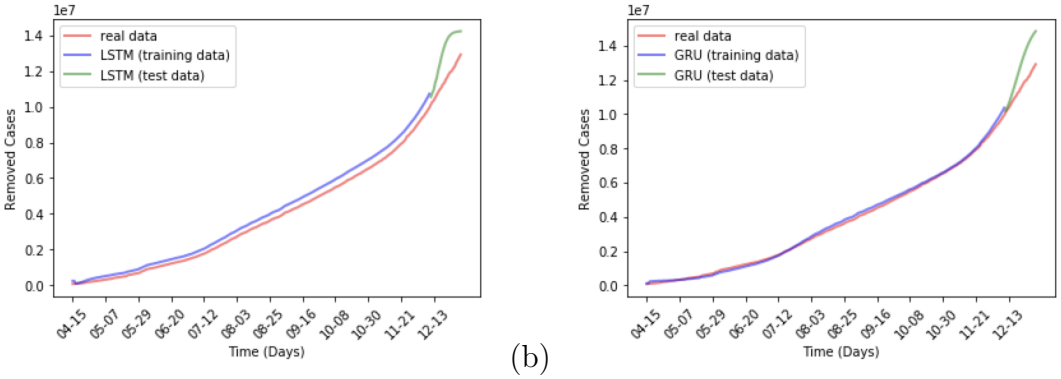

(a) (b)

Figure 13: Prediction of active cases for the USA by (a) LSTM; and (b) GRU between Apr 15, 2020 and Dec 31, 2020. The red curve presents the real data of active cases; the blue curve presents the prediction of the training data; the green curve presents the prediction of the test data.



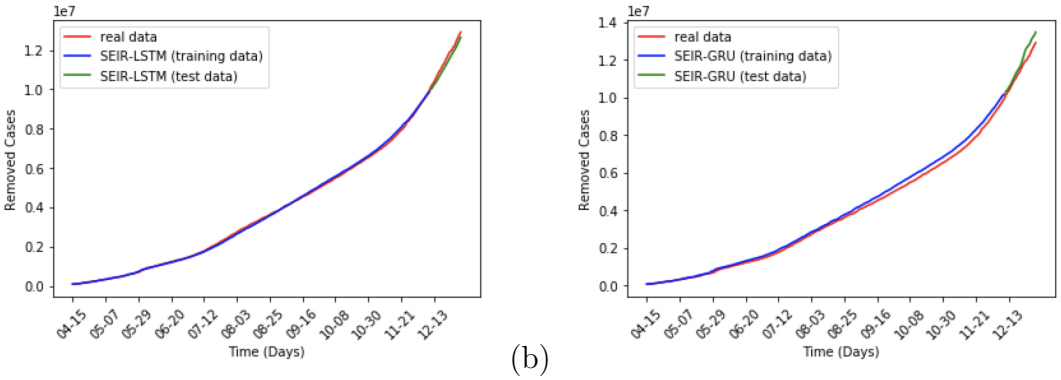

(a) (b)

Figure 14: Prediction of active cases for the USA by (a) SEIR-LSTM; and (b) SEIR-GRU between Apr 15, 2020 and Dec 31, 2020. The red curve presents the real data of active cases; the blue curve presents the prediction of the training data; the green curve presents the prediction of the test data.




(a) (b)

Figure 15: Prediction of removed cases for the USA by (a) LSTM; and (b) GRU between Apr 15, 2020 and Dec 31, 2020. The red curve presents the real data of active cases; the blue curve presents the prediction of the training data; the green curve presents the prediction of the test data.




(a) (b)

Figure 16: Prediction of removed cases for the USA by (a) SEIR-LSTM; and (b) SEIR-GRU between Apr 15, 2020 and Dec 31, 2020. The red curve presents the real data of active cases; the blue curve presents the prediction of the training data; the green curve presents the prediction of the test data.

## 5.3 Three common error metrics

We introduce three commonly used error metrics that used in this thesis in this section. Suppose there are a total of $N$ data, let $y_i$ 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:

### 5.3.1 Root-Mean-Square-Error (RMSE)

The RMSE is a commonly used error metric, which is the square root of the quadratic mean of the differences between predicted and actual values. It can more directly reflect the difference between the actual value and the predicted value. The RMSE is expressed mathematically as

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2} \tag{34}$$

The range of MSE is $[0, +\infty)$. The larger the RMSE is, the larger the error is, which implies the worse the model would be. Conversely, the smaller the RMSE is, the smaller the error is, which implies the better the model. When $RMSE = 0$, it means that the predicted value is completely consistent with the actual value, which implies a perfect model.

### 5.3.2 Mean-Absolute-Percentage-Error (MAPE)

The MAPE measures the error in percentage, which is the absolute mean of the ratio of the predicted error to the actual value. [35] [39] It reflects the relative error based on the actual data in the form of a ratio, and can effectively avoid the intuitive impact on the size of the error value caused by the difference in data scale.

There is an example that can help to understand this better. Assume the real data is one million, and the prediction by the model is one million and 10,000, then the error of this point is 10,000. Intuitively, "10,000" is a very large number, but we can't conclude the model is bad directly. Because ten thousand accounted for only 1%

of one million, which means that the prediction deviation is only 1%. If you consider the actual value be 10 and the prediction be 9.9, then you will see the 1% error is just 0.1, which is a very small number. From this point of view, the absolute error sometimes can not reflect the effect of the forecast truly. Therefore, here we regard the relative error as another important metric for the model evaluation. The MAPE is calculated as:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \tag{35}$$

The smaller the MAPE, the better the model. If the data has no extreme values and zero values, the evaluation effect of MAPE could be very good.

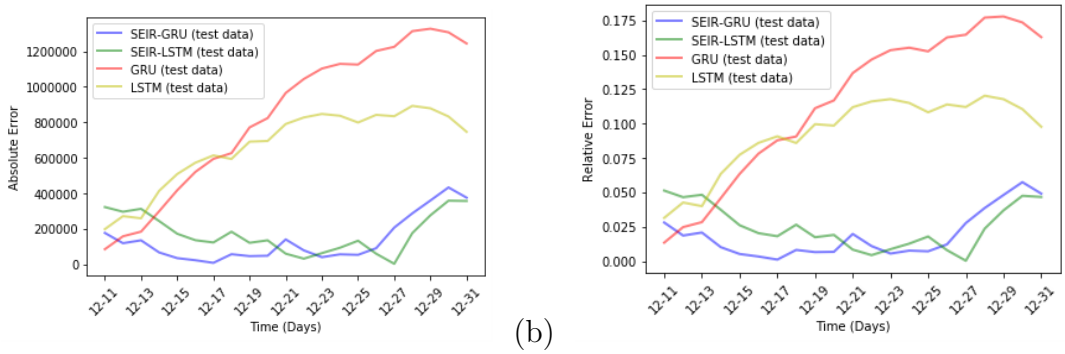
### 5.3.3 $r_2$ score

$r_2$ score [36] [37] [38] 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:

$$r_2(y, \hat{y}) = 1 - \frac{\sum_{t=1}^{N}(\hat{y}_t - y_t)^2}{\sum_{t=1}^{N}(y_t - \bar{y}_t)^2} \tag{36}$$

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

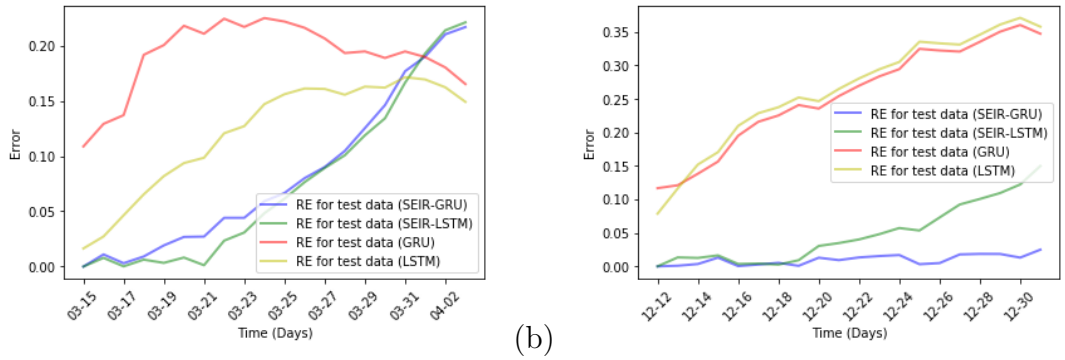
## 5.4 Results




(a) (b)

Figure 17: The (a) absolute error; and (b) relative error for the test data of active cases for the USA

Figure 17 and Figure 18 show the error graphs of the test data of I and R, respectively. From the results, we can see that the prediction results of the SEIR-LSTM and SEIR-GRU models are good regardless see from the absolute error or the relative error.




(a) (b)

Figure 18: The (a) absolute error; and (b) relative error for the test data of removed cases for the USA

| Variable | Model | RMSE | MAPE | $r_2$ score |
|---|---|---|---|---|
| Active cases (I) | LSTM | 714945.7666415 | 0.0932058 | -1.9011966 |
| | GRU | 951266.7168208 | 0.1154006 | -4.1361248 |
| | SEIR-LSTM | 251137.8920232 | 0.0251214 | 0.6420227 |
| | SEIR-GRU | 178906.9787563 | 0.0187947 | 0.8183290 |
| Removed cases (R) | LSTM | 3902973.6369046 | 0.1308772 | -19.5962081 |
| | GRU | 3723350.9183446 | 0.1019719 | -17.7440733 |
| | SEIR-LSTM | 391136.7520847 | 0.0325937 | 0.7931511 |
| | SEIR-GRU | 279495.5196168 | 0.0164521 | 0.8943801 |

Table 3: Validation Metrics for active cases and removed cases of COVID-19 forecasting using LSTM, GRU,SEIR-LSTM and SEIR-GRU models.

In order to have a more accurate evaluation of the performance of the prediction for the parameters of SEIR model of the four methods, we calculated the metrics RMSE, MAPE and $r_2$ scores for the test data of each parameter, and summarized them in the Table 3. From the above table, we can easily see that the SEIR-GRU model outperforms other models. Because among the four models, SEIR-GRU provids a better forecasting performance with lowest RMSE and MAPE values and the $r_2$ score closest to 1.

## 5.5 Summary

In this chapter, we bring the parameters learned and predicted by LSTM and GRU in the previous chapter back to the SEIR model for solution, and compare and analyze the obtained active cases and removed cases with real data. The RMSE, MAPE and $r_2$ scores were used to evaluate the models of LSTM, GRU, SEIR-LSTM and SEIR-GRU, respectively. The results show that the SEIR-LSTM and SEIR-GRU models perform very well for the predictions of the US data from April 15, 2020 to December 31, 2020.

# CHAPTER 6

# CONCLUSION

At the beginning of the thesis, we introduced the SEIR model and raised questions about 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, and then combined regression analysis and other statistical methods to analyze and discuss the prediction results. The results showed that the variance of the parameter predicted by LSTM and GRU is smaller than that of the theoretical values, that is, the fluctuation range of the predicted data is slightly smaller than the theoretical value. Judging from the overall trend, GRU's prediction results are closer to theoretical results. From the overall mean and variance, the prediction result of LSTM is closer to the theoretical values. In the meantime, we got the conclusion that the epidemic in the United States has eased during the period of our research, although the magnitude is not large. We put the learning and prediction parameters of LSTM and GRU back to the SEIR model for solutions, compared and analyzed the true values and prediction of active cases and removed cases. Finally, we used RMSE, MAPE and $r_2$ scores to evaluate the LSTM, GRU, SEIR-LSTM and SEIR-GRU models respectively. The results show that both the SEIR-LSTM and SEIR-GRU models have smaller RMSE and MAPE values, and the $r_2$score value closest to 1, regardless of whether it for active cases or removed cases. The minimum MAPE is as low as 1.6%, and the $r_2$ score is as high as 0.894. It fully illustrates the role and potential of LSTM and GRU in predicting the COVID-19 epidemic trend.

The main contribution of this paper is to solve the problem of the dynamic change of the parameters of the SIER model, put forward a set of relatively complete and effective SEIR model parameter learning and prediction ideas, and verify the application potential of LSTM and GRU in this field.

This article just selects the most basic compartmental models in epidemiology

model,SEIR model, as the basis, and expands around its parameter problem. In fact, there are many other models that are inherently more effective than SEIR models, such as SEIRD or dynamic SEIR models. These are all worth exploring. And only two machine learning methods are used here. In future work, we will apply more machine learning methods to explore its application potential in various fields.

All code of this thesis are written by Python. The core code has been uploaded to my GitHub: `https://github.com/zero3829/Lin_Feng_Master_Thesis_code/blob/main/USA`

## REFERENCE


[1] BBC News, *Coronavirus disease named Covid19*, (2020).

[2] S. Roychoudhury, A. Das, P. Sengupta, S. Dutta, S. Roychoudhury, A.P. Choudhury, A. B. Fuzayel Ahmed, S. Bhattacharjee, P. Slama, *"Viral Pandemics of the Last Four Decades: Pathophysiology, Health Impacts and Perspectives"*, International Journal of Environmental Research and Public Health. **17**(24): 9411, (2020).

[3] A. Zeroual, F. Harrou, A Dairi, Y. Sun, *Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study*, (2020).

[4] S. Bentout, A. Chekroun, T. Kuniya, *Parameter estimation and prediction for coronavirus disease outbreak 2019 (COVID-19) in Algeria*, AIMS Public Health, (2020), **7**(2): 306–318.

[5] Anderson Castro Soares de Oliveira, Lia Hanna Martins Morita, Eveliny Barroso da Silva, Luiz André Ribeiro Zardo, Cor Jesus Fernandes Fontes, Daniele Cristina Tita Granzotto, *Bayesian modeling of COVID-19 cases with a correction to account for under-reported cases*, Infect Dis Model, (2020) **5**: 699–713.

[6] Schmidt, J., Marques, M.R.G., Botti, S. et al., *Recent advances and applications of machine learning in solid-state materials science*, npj Comput Mater **5**, 83, (2019).

[7] A. Zeroual, F. Harrou, A. Dairi, Y. Sun, *Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study*, Chaos, Solitons and Fractals, **140**, (2020), 110121.

[8] F. Shahid, A. Zameer, M. Muneeb, *Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM*, Chaos, Solitons and Fractals, **140**, (2020), 110212.

[9] García S., Luengo J., Herrera F. *Introduction. In: Data Preprocessing in Data Mining*, Intelligent Systems Reference Library, **72**. Springer, Cham, (2014), 1-17.

[10] *General process and necessary steps of machine learning tasks*.

[11] *Data Preprocessing: 6 Necessary Steps for Data Scientists*, (2020).

[12] *Data Preprocessing in Data Mining*, (2019).

[13] S. B. Kotsiantis, D. Kanellopoulos, P. E. Pintelas, *Data Preprocessing for Supervised Leaning*, INTERNATIONAL JOURNAL OF COMPUTER SCIENCE, **1**, (2006), ISSN, 1306-4428.

[14] Brauer F., *Compartmental Models in Epidemiology*, In: Brauer F., van den Driessche P., Wu J. (eds) Mathematical Epidemiology. Lecture Notes in Mathematics, (2008), **1945**, 19–79.

[15] Abou-Ismail A., *Compartmental Models of the COVID-19 Pandemic for Physicians and Physician-Scientists*, SN Compr Clin Med, (2020), 1-7.

[16] van den Driessche P., *Reproduction numbers of infectious disease models. Infectious Disease Modelling*, (2017), **2**(3), 288–303.

[17] Ridenhour, B., Kowalik, J. M., Shay, D. K., *Unraveling R0: considerations for public health applications.*, American journal of public health, (2014), **104**(2), e32–e41.

[18] T. Harko, Lobo, Francisco S. N., Mak, M. K., *Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates*, Applied Mathematics and Computation. **236**: 184–194, (2014).

[19] R. Beckley, C. Weatherspoon, M. Alexander, M. Chandler, A. Johnson, Batt, Ghan S. *Modeling epidemics with differential equations*, Tennessee State University Internal Report. Retrieved July 19, 2020, (2013).

[20] M. Farman, M. Umer Saleemb, A. Ahmada, M.O.Ahmad, *Numerical solution of nonlinear fractional SEIR epidemic model by using Haar wavelets*, Ain Shams Engineering Journal, (2018), **9**, 4, 3391-3397.

[21] B. Prakash, A. Setia, D. Alapatt, *Numerical solution of nonlinear fractional SEIR epidemic model by using Haar wavelets*, Journal of Computational Science, (2017), **22** 109-118.

[22] A. Hassan, I. Shahin, M. Bader Alsabek, *COVID-19 Detection System Using Recurrent Neural Networks*, (2020).

[23] G. Petneházi, *Recurrent Neural Networks for Time SeriesForecasting*, (2018).

[24] H. Hewamalage, C. Bergmeir, K. Bandara *Recurrent Neural Networks for Time Series Forecasting: Current Statusand Future Directions*, (2019).

[25] Schafer, A. M., Zimmermann, H. G., *Recurrent Neural Networks Are Universal Approximators*, Proceedings of the 16th International Conference on Artificial Neural Networks -Volume Part I. ICANN'06. Springer-Verlag, Berlin, Heidelberg, pp. 632-640, (2006).

[26] Hochreiter, S., Schmidhuber, J., *Long short-term memory*, Neural Comput. **9**(8), pp. 1735–1780, (1997).

[27] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., *Learning phrase representations using RNN Encoder–Decoder for statisticalmachine translation*, Proceedings of the 2014 Conference on Empirical Methods in NaturalLanguage Processing (EMNLP). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1724-1734, (2014).

[28] Hethcote H, *The Mathematics of Infectious Diseases*, SIAM Review. **42**(4): 599–653, (2000).

[29] *List of Probability and Statistics Symbols* Math Vault, 2020-04-26, Retrieved 2020-09-12.

[30] *Mean Squared Error (MSE)*, www.probabilitycourse.com, Retrieved 2020-09-12.

[31] *Introduction to Machine Learning Model Evaluation*, (2019).

[32] *Model Evaluation*.

[33] Lehmann, E. L.; Casella, George, *Theory of Point Estimation (2nd ed.)*, New York: Springer, (1998).

[34] *MAE and RMSE — Which Metric is Better?*, (2016).

[35] Everitt, B. S.; Skrondal, A., *The Cambridge Dictionary of Statistics*, Cambridge University Press, (2010)

[36] Steel, R. G. D.; Torrie, J. H., *Principles and Procedures of Statistics with Special Reference to the Biological Sciences*, (1960).

[37] Glantz, Stanton A.; Slinker, B. K., *Primer of Applied Regression and Analysis of Variance*, (1990).

[38] Draper, N. R.; Smith, H., *pplied Regression Analysis. Wiley-Interscience*, (1998).

[39] *Mean absolute percentage error (MAPE)*, (2017).

[40] CDC, 2019 Novel Coronavirus, Wuhan, China: Symptoms. CDC. Available at *https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html*, January 26, 2020; Updated Feb, 2021.

[41] Lauer SA, Grantz KH, Bi Q, et al., *The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application*, (2020).

[42] J. Backer, D. Klinkenberg, J. Wallinga, *Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travelers from Wuhan, China*, Euro. Surveill., **25**(5), (2020), 20-28.