Enhancing Online Epidemic Supervising System by Compartmental and GRU Fusion Model

Ma, J., Wang, X., Wang, Y., Wang, J., Chu, X. & Zhao, J.

Research Article

Enhancing Online Epidemic Supervising System by Compartmental and GRU Fusion Model

Junyi Ma [1,2] **Xuanliang Wang**[1,2] **Yasha Wang**[1,3] **Jiangtao Wang**[4] **Xu Chu**[5] **and Junfeng Zhao**[1,2]

[1]*School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*
[2]*Key Lab of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing 100871, China*
[3]*National Engineering Research Center for Software Engineering, Peking University, Beijing 100871, China*
[4]*Center for Intelligent Healthcare, Coventry University, Coventry CV1 5FB, UK*
[5]*Department of Computer Science, Tsinghua University, Beijing 100084, China*

Correspondence should be addressed to Yasha Wang; wangyasha@pku.edu.cn

The global pandemic, COVID-19, is an acute respiratory infectious disease caused by the 2019 novel coronavirus. Building the online epidemic supervising system to provide COVID-19 dynamic prediction and analysis has attracted the attention of the industry and applications community. In previous studies, the compartmental models and deep neural networks (DNNs) played important roles in predicting and analyzing the dynamics of the pandemic. Nevertheless, the compartmental model has limited ability to fit historical data and thus leads to unsatisfactory prediction accuracy due to the difficulty in parameter estimation. For DNNs, the lack of interpretability makes it difficult to explain the prediction results; thus, it cannot provide an in-depth understanding of the transmission mechanism of the pandemic. We propose a fusion model to leverage the merits of both models and resolve their shortcomings. The fusion model extracts epidemic-related knowledge from the state-of-the-art SEIDR compartmental model to guide the training of the GRU model, which can preserve the interpretability and achieve a good performance in predicting epidemic dynamics. This model can help to enhance the online epidemic supervising system by providing more accurate prediction results and deeper analysis. Our extensive experiments across multiple epidemic datasets from six European countries demonstrate that our model outperforms existing state-of-the-art baselines in predicting the active confirmed cases. More importantly, by analyzing the effective reproductive number, our method can reveal the risk of the second wave of the epidemic in Europe and justify the importance of social distancing to control the outbreak of the epidemic.

1. Introduction

COVID-19 is a respiratory infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of October 2021, more than 241 million cases have been reported worldwide, resulting in more than 4.9 million deaths. Europe is one of the most severely affected areas in this pandemic. As of October 2021, over 71 million cases have been reported, together with more than 1.3 million deaths [1]. All European countries are facing unprecedented health challenges because of this severe pandemic.

At this time, increasingly large amounts of epidemic-related data are released on the web, which has great potential for enabling better epidemic dynamic prediction and analysis [2]. These data open plans enable the industry and applications community researchers worldwide to collectively build the epidemic supervising system and conduct COVID-related data analytics or dynamics predictions to combat the deadly virus [3–6]. For example, the system called *COVID19-Projections* uses an AI-based model to make the COVID-19 dynamic predictions and has been cited by the Centers for Disease Control and Prevention (CDC) as a tool to help inform public health decision-

making [6]. In general, the main task of such supervising systems is to make the accurate COVID-19 prediction and provide analysis for formulating epidemic prevention policies.

Previous studies have revealed the significance and importance of predicting and understanding the future dynamics of the epidemic transmission patterns [7–10]. Firstly, the prediction of the pandemic dynamics could reveal the important time points in advance, thus providing sufficient preparation and response time for dealing with the thorny challenges brought by the epidemic. Secondly, analyzing transmission patterns, such as the effective reproductive (R) number, death rate, and recovery rate, can provide valuable insights into future epidemic prevention. Lastly, understanding the dynamics of the pandemic can assist in evaluating the effectiveness of existing epidemic prevention measures. For example, the high death rate may raise the risk of a shortage of medical resources, and the government may need to invest more related resources to control the outbreak of the epidemic.

Recent research works about the COVID-19 pandemic analysis or modeling can be categorized into two types: compartmental model and DNNs model. The compartmental model, including SI, SIR, SEIR, and SIS model [9–11], is a type of mathematical model that simulates how individuals in different populational *compartments* interact. The earliest compartmental model is the SIR model introduced in 1927 [11, 12]. SIR model first assigns the population to compartments with different labels (e.g., susceptible, infectious, or recovered) and then constructs transition equations to model the flow between different compartments. By estimating the parameters in transition equations (e.g., death rates, recovery rates, and infection rates), the compartmental model is able to predict some vital epidemic factors, such as how disease spreads or the total infection numbers. As for DNNs-based models, recent studies concentrate on the prediction of the future dynamics of the pandemic by fitting historical epidemic data (e.g., confirmed cases and recovered cases) with the aim of faster convergence and less deviation. Most of the solutions are based on various types of recurrent neural networks such as Bi-LSTM, LSTM, and GRU [13–16].

Although they have made some progress, the above methods are still far away from satisfactory performance on the analytical tasks for COVID-19 in terms of both *interpretability* and *prediction accuracy*.

For example, recent solutions based on compartmental models try to increase the model complexity such that they can model the transmission pattern more precisely by considering the factors such as the limited virus test capabilities and the quarantined population [17, 18]. However, due to the increased, complicated model structure, those solutions need to estimate a large number of parameters, which usually is very difficult in practice, thus leading to unsatisfactory accuracy in predicting results [19]. On the other hand, although DNNs-based solutions can predict the dynamics of the pandemic with high accuracy, they fail to provide interpretability on both the prediction result and the transmission mechanism [20]. As a result, some

transmission patterns, such as transition rates, cannot be inferred, which hinders an in-depth understanding of the COVID-19 pandemic. For example, existing DNNs-based solutions cannot predict the effective R number of COVID-19 with trustworthiness and interpretability [20, 21]. Besides, recent studies have also revealed that neural network models suffer from the overfitting issue as historical epidemic data are usually insufficient [19].

Despite the shortcomings of both solutions, their technical advantages are complementary. The compartmental models have the advantage of having more interpretability than the neural network models, which can provide explainable and trustworthy analytical results for the COVID-19 epidemic. The advance of neural network models lies in their strong representation-learning ability from the historical data, which can predict COVID-19's future dynamics more accurately. Thus, the key novelty of this article is to *leverage the complementarity of both the compartmental model and neural network model to design a highly accurate and interpretable fusion model* for COVID-19 prediction and analyses.

Specifically, we first introduced a state-of-the-art compartmental model called the SEIDR model, which can better model this COVID-19 pandemic by considering the partially reported COVID-19 infections and the quarantine status of the population [22, 23]. Based on the SEIDR model, we proposed a SEIDR-guided GRU model. This fusion model can extract epidemic knowledge from the compartmental model to guide the training of the GRU model, which can preserve the interpretability and reduce the overfitting issue. After that, we collect history epidemic data from six European countries in different regions, including daily confirmed cases, recovered cases, and deaths from the day of the first case in each country to September 30, 2020. By applying our proposed fusion model to these real-world epidemic datasets, we show that our proposed model significantly outperforms the state-of-the-art baselines such as the SEIR model. We also demonstrate, by the prediction of the active confirmed cases and the analysis of the effective reproductive numbers of six countries, that our method can reveal the risk of the second wave of the COVID-19 epidemic in Europe before it happens and justify the importance of the social distancing to control the outbreak of the epidemic.

The contributions of this article are summarized as follows:

(i) We proposed a SEIDR-guided GRU fusion model, which can extract epidemic knowledge from the SEIDR model to guide the training of the GRU model. This fusion model leverages both merits of the SEIDR model and the GRU model, which can preserve the interpretability, overcome overfitting, and achieve state-of-the-art performance in terms of predicting the future COVID-19 dynamics.

(ii) We intensively evaluate our fusion model on real-world epidemic data from six European countries. Our experiments demonstrate that our model significantly outperforms existing state-of-the-art baselines such as the SEIR model in terms of the

prediction of the active confirmed cases and the estimation of R numbers (reproductive number) in six countries.

(iii) Our analysis also successfully reveals the risk of the second wave of the epidemic in Europe before it happens and justifies the importance of social distancing to control the outbreak of the epidemic from the perspective of pandemic modeling.

2. Related Works

2.1. Compartmental Model-Based Methods. The compartmental model is the most widely used method to analyze and predict the dynamics of infectious diseases, also including this COVID-19 pandemic [8–10, 24–26]. A typical pipeline of using a compartmental model is to design a compartmental model to represent the transmission of the epidemic (e.g., SEIR model), infer parameters of the model (e.g., recovery rate), and then predict and analyze the dynamics based on inferred parameters. For example, Pandey et al. deployed the SEIR model to India's COVID-19 epidemic data and then predicted the dynamics of COVID-19 transmission in India [9]. Hou et al. modified the original SEIR model by introducing the influence of interaction between people. They collected epidemic data from different provinces or cities in China and conducted analyses to investigate how social distancing affects the dynamics of the COVID-19 epidemic [10]. Peng et al. and López and Rodo also investigated how different antiepidemic measures influence the dynamics of the COVID-19 pandemic by extending the SEIR model to consider more factors [24, 25]. Besides, Wangping et al. utilized a SIR model to compare the transmission patterns (i.e., compartmental model parameters) between Italy and Hunan province, which has a similar population [8].

In summary, existing works that use compartmental models to predict and analyze COVID-19 dynamics are mainly based on SEIR and SIR models and tried to extend the compartmental model to better depict patterns of the COVID-19 pandemic. Considering more situations makes the structure of compartmental models more complex and harder to solve their parameters, which also leads to the low accuracy of the epidemic dynamic prediction [19].

2.2. Deep Neural Network-Based Methods. Related studies tend to predict the future dynamic of the pandemic by fitting historical epidemic data (e.g., confirmed cases and recovered cases) with the aim of faster convergence and less deviation. Most of these methods are based on the models such as Bi-LSTM, LSTM, and GRU [13–16, 27–29]. For example, Arora et al. collected the epidemic data of India from March 14, 2020, to May 14, 2020, and then utilized the modified LSTM model to fit and predict the epidemic dynamics. Their results show that their model works well when predicting the future in a short time (around 1 to 3 days) [14]. Melin et al. proposed a fuzzy response aggregation method, which can ensemble different simple neural networks for COVID-19 epidemic prediction. They trained their model on Mexico's

epidemic data and demonstrated that the ensemble model outperforms single models [28]. Besides, Zeroua et al. trained multiple deep learning models, including simple RNN, LSTM, bidirectional LSTM, GRU, and VAE, on historical epidemic data. They further compared the performance of predicting pandemic dynamics between different models [15].

In a word, previous works using deep neural network methods to predict COVID-19 epidemic dynamics mainly focus on fitting historical data with faster convergence and less deviation. The interpretability of models is insufficient as transmission patterns of COVID-19, such as recovery rate and death rate, cannot be inferred. The lack of interpretability limits further understanding of the epidemic. For example, the dynamics of COVID-19's effective reproductive number is hard to analyze by existing neural network models.

3. Fusion Model Framework

In this section, we first introduce the overview of our proposed compartmental and GRU fusion model for COVID-19 epidemic analysis and prediction. Then, we present the details of the two components of the fusion model, respectively, including the SEIDR compartmental model and the SEIDR-guided GRU fusion model. Finally, we introduce the loss function of the fusion model.

3.1. Framework Overview. We proposed a fusion model framework, as shown in Figure 1, that consists of the following components.

SEIDR Compartmental Model. We introduced a state-of-the-art compartmental model to meet the transmission patterns of this COVID-19 pandemic, called the SEIDR model [24, 25]. Specifically, the SEIDR model separated people who have been confirmed by testing and are strictly quarantined (i.e., confirmed active case, called D compartment) from all infected populations. Then the D compartment can further represent the daily reported infections. More details about the proposed SEIDR model, such as parameters and differential equations, would be introduced in the following section.

SEIDR-Guided GRU Fusion Model. We proposed the SEIDR-guided GRU fusion model to keep the interpretability of prediction and reduce the overfitting by leveraging epidemic-related knowledge extracted from the SEIDR model. Specifically, we first utilize a GRU model to fit the historical epidemic data. The parameters of the GRU model are optimized by minimizing the errors between the input epidemic data and the output of the GRU model (i.e., prediction), which is similar to previous COVID-19 prediction works. Then, we extract epidemic-related knowledge from the SEIDR model, which includes the constraints between parameters and predicted results (e.g., the prediction of confirmed cases and deaths should satisfy the death rate as much as possible). These constraints are finally

(a)

(b)
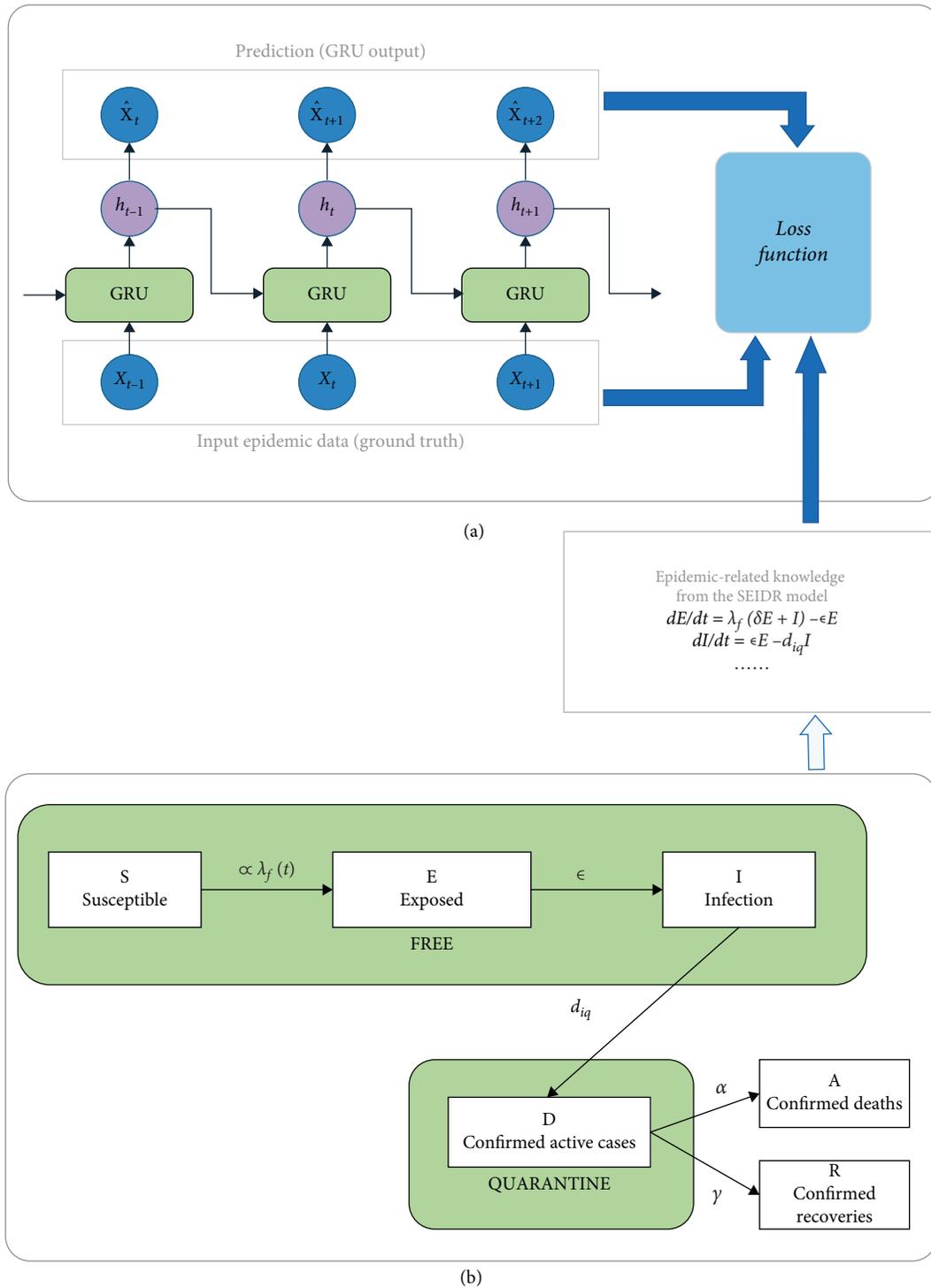
FIGURE 1: The framework overview of the SEIDR-guided GRU fusion model, which consists of the SEIDR-guided GRU fusion model (a) and the SEIDR compartmental model (b).

applied to the objective function for training the GRU model. More details about the SEIDR-guided GRU fusion model will be introduced as follows.

3.2. SEIDR Compartmental Model.

In this work, we introduce a SEIDR compartmental model, which satisfies two situations of the existing COVID-19 pandemic [24, 25].

First, to address the situation that the number of daily reported infections can not represent the true number of infections in the whole population, the SEIDR model separates out people who have been confirmed by testing and are strictly quarantined from all infected populations. These confirmed active cases can represent the daily reported infections and can be seen as a new compartment D of the model. Second, to address the challenge that a number of infections (i.e., confirmed active cases) are quarantined and cannot spread the virus, the SEIDR model divides all compartments into two parts: the quarantine part and the free part. People who are in the free part can infect susceptible individuals, whereas people who are in the quarantine part cannot. According to the antiepidemic measures in most countries, confirmed active cases are treated in hospitals or under strict home quarantine so that they can hardly contact other people. Thus, the model puts the compartment D into the quarantine part and other compartments into the free one. The reported deaths compartment A and the reported recoveries compartment R are not included in the above two parts because they would not interact with others.

Figure 1(b) presents the structure of the SEIDR model. In this figure, each rectangle represents one compartment, and the arrow between the rectangles indicates the transfer between different compartments. There are six compartments in the SEIDR model. The S one is those who are able to contract the disease. The E one is those who have been infected but are not yet infectious. The I one is those who are infected but are not confirmed by testing. The D one is those who have been confirmed and reported after testing. R and A are reported recoveries and reported deaths, respectively. The parameters on the arrow represent the transition probability between different compartments per unit time, which is equivalent to one day in this work. More details of all parameters are shown in Table 1.

It is worth mentioning that the number of susceptible cases is usually large, so the transition probability between susceptible individuals and exposed people is close to zero. This phenomenon results in difficulty in estimating the transition probability and also affects the estimation of other parameters in the model. Thus, we introduce the parameter λ_f as previous works did to present the number of cases an infected person can infect per day [30]. λ_f can be calculated as follows:

$$\lambda_f(t) = \frac{D(t+M)}{\sum_{k=1}^{M-1} D(t+k)}, \tag{1}$$

where M is the mean duration from onset to diagnosis. We set $M = 6$ according to the previous investigation [30]. Considering that people who are in the exposed compartment and those who are in the infection compartment have different abilities to infect susceptible individuals [31], we further introduce the parameter δ to represent the ratio of the infective capacity between the exposed population and the infection population. Thus, the population moving from S to E can be represented as $\lambda_f(\delta E + I)$.

Based on the structure and the flows between different compartments of the SEIDR model, we have the following differential equations that contain the constraints between the population in different compartments:

$$\frac{dE}{dt} = \lambda_f(\delta E + I) - \epsilon E,$$

$$\frac{dI}{dt} = \epsilon E - d_{iq} I,$$

$$\frac{dD}{dt} = d_{iq} I - (\alpha + \gamma)D, \tag{2}$$

$$\frac{dA}{dt} = \alpha D,$$

$$\frac{dR}{dt} = \gamma D.$$

3.3. SEIDR-Guided GRU Fusion Model.

In order to maintain the interpretability of the neural network model on the prediction of COVID-19 pandemic dynamics and reduce the overfitting problem, we propose a SEIDR-guided GRU fusion model which can leverage epidemic-related knowledge extracted from the SEIDR model. The training of this fusion model can be further divided into two phases.

The first phase is to fit the historical epidemic data by minimizing the errors between the input epidemic data and the output of the model. Specifically, we design a GRU model as shown in Figure 1(a). When epidemic data X_t on day t and the previous day's hidden state h_{t-1} input into the GRU model, the GRU cell will output the hidden state h_t of this day. Then h_t can be used to predict the epidemic state $\widehat{X}_{t+1}$ of the $t + 1$ day and also be input to the next GRU cell to predict the epidemic state of the $t + 2$ day. The input epidemic data X_t and the corresponding prediction $\widehat{X}_{t+1}$ can be represented as follows:

$$X_t = (D'(t), R'(t), A'(t)),$$
$$\widehat{X}_{t+1} = (D'(t+1), R'(t+1), A'(t+1)), \tag{3}$$

where $D'(t)$ means increased confirmed active cases compared with the $t - 1$ day; $R'(t)$ means increased recoveries cases compared with the $t - 1$ day; $A'(t)$ means increased confirmed deaths compared with the $t - 1$ day.

The second phase of training this fusion model is extracting pandemic-related knowledge from the SEIDR model and using this knowledge to guide the training of the GRU model in Figure 1(a). The differential equations of the SEIDR model indicate some constraints between the parameters of the SEIDR model and the population of different

TABLE 1: The notations of parameters in the SEIDR model and their meanings.

Notations	Meanings
S	Susceptible population
E	Exposed population
I	Infective population
D	Diagnosed population
A	Deaths
R	Recoveries
α	Death rate
γ	Recovery rate
ε	Transition probability: from exposed population to infective population
d_{iq}	Transition probability: from infective population to diagnosed population
λ_f	Cases an infected person can infect per day*
δ	Ratio of the infective capacity between E and I

*An infected person means a person who already have symptoms (i.e., I).

compartments. For example, equation (5) shows that the number of confirmed active cases and the number of deaths should satisfy the death rate α as much as possible. These constraints can be seen as epidemic-related knowledge to restrict the relationship between the GRU model's output (e.g., restrict the predicted value of A and D to conform equation (5)).

According to the differential equations of the SEIDR model, we can obtain constraint relations as follows:

$$C_1 = \frac{\mathrm{d}E}{\mathrm{d}t} - \lambda_f (\delta E + I) + \epsilon E,$$

$$C_2 = \frac{\mathrm{d}I}{\mathrm{d}t} - \epsilon E + d_{iq}I, C_3 = \frac{\mathrm{d}D}{\mathrm{d}t} - d_{iq}I + (\alpha + \gamma)D,$$

$$C_4 = \frac{\mathrm{d}A}{\mathrm{d}t} - \alpha D, \tag{4}$$

$$C_5 = \frac{\mathrm{d}R}{\mathrm{d}t} - \gamma D,$$

$$C_6 = N - (S + E + I + D + R).$$

To utilize these constraint relations as knowledge to keep the interpretability of prediction and reduce the overfitting problem, we first add the parameters of the SEIDR model to the fusion model and train these parameters with GRU parameters together. Then, we add the constraints represented by C_1 to C_6 to the loss function with the aim of getting them as close to zero as possible.

3.4. The Loss Function. The loss function of our proposed SEIDR-guided GRU fusion model contains two parts: the fitting loss $\mathscr{L}_f$ and the constraint loss $\mathscr{L}_c$.

The fitting loss is used to minimize the errors between the input epidemic data and the prediction data (i.e., increase in the number of confirmed active cases, recoveries cases, and deaths) for better fitting the historical data. It can be represented as follows:

$$\mathscr{L}_f = \sum_{t=1}^{T} \left[\left| D'(t) - \widehat{D}'(t) \right|^2 + \left| R'(t) - \widehat{R}'(t) \right|^2 + \left| A'(t) - \widehat{A}'(t) \right|^2 \right], \tag{5}$$

where T is the number of days of historical COVID-19 epidemic data available for training the model.

As for the constraint loss, this loss function is utilized to restrict the relationship between the fusion model's parameters and its output by leveraging knowledge extracted from the SEIR model. It can be represented as follows:

$$\mathscr{L}_c = \sum_{i=1}^{6} \sum_{t=1}^{T} \left| C_i(t) \right|^2, \tag{6}$$

where T is the number of days of historical COVID-19 epidemic data available for training the model and C_i means the constraint relations C_1 to C_6.

Only using the fitting loss often suffers from the overfitting issue as historical epidemic data are usually insufficient, which may limit the performance of prediction. Meanwhile, the constraint loss can extract knowledge from the compartmental model to further reduce the search space of model parameters and inhibit the overfitting problem. We finally make comprehensive utilization of the fitting loss and the constraint loss.

Composed of the fitting loss $\mathscr{L}_f$ and the constraint loss $\mathscr{L}_c$, the system for COVID-19 epidemic prediction is optimized in an end-to-end way. We minimize the following loss function:

$$\mathrm{Loss} = \mathscr{L}_f + \mathscr{L}_c. \tag{7}$$

4. Experiments and Results

In this section, we carry out experiments to evaluate our proposed fusion model for COVID-19 prediction and analyze the transmission patterns of six European countries based on the parameters inferred by our model. We aim to answer the following research questions:

(i) *RQ1*: Can the SEIDR-guided GRU fusion model outperform existing methods such as SEIR and GRU? (The answer is in Subsection 4.3.)

(ii) *RQ2*: How many days of data does the fusion approach need to train a model with good performance? Can the fusion model achieve good performance at the different stages of the pandemic? (The answer is in Subsection 4.4.)

(iii) *RQ3*: If the fusion model achieves better performance, can we conduct a deeper analysis of the COVID-19 pandemic based on this model? (The answer is in Subsection 4.5.)

(iv) *RQ4*: Can we obtain meaningful findings based on the analysis of different countries? (The answer is in Subsection 4.6.)

4.1. Data Collection. We use the JHU CSSE COVID-19 dataset as our source data [3]. This dataset contains time series data of confirmed cases, recovered cases, and deaths. The dataset is updated daily and reported the dynamics of the COVID-19 epidemic all over the world. We investigate European countries because of the following: (1) Europe has a well-developed medical system and relatively reliable data; (2) Europe has more countries with diverse data. Since it is "hard" to present the results of each country due to space limitations, six typical countries were selected according to geographic location and epidemic situation, including Italy in southern Europe, Germany in western Europe, Switzerland and Austria in central Europe, Denmark in northern Europe, and Russia in eastern Europe. As for the time span, we select the data just before the second wave epidemic as a training set (e.g., the day of the first case in each country to 2020.8.31) to predict active confirmed cases within the first month of the second wave(e.g., 2020.9.1 to 2020.9.30). Thus, we can analyze the risk of the second wave epidemic, which has attracted much attention in COVID-19 prediction works [32, 33].

4.2. Experiment Settings. Our SEIDR-guided GRU fusion model is implemented in PyTorch (https://pytorch.org/). We use the Adam optimizer [34] with an initial learning rate of 0.0005 because it can automatically adjust the learning rate during the training phase. We set the hidden state size to 256. Besides, consistent with previous works [31], the parameter $\lambda_f(t)$ used in the training phase is calculated with historical data by the equation (1), while this parameter used in the prediction phase is the average value of the last 10 days of the training data.

4.3. Prediction of Active Confirmed Cases. To answer the RQ1, we compare our model to existing methods, including the SEIR model, the SEIDR model, and GRU model, to explore whether our SEIDR-guided GRU fusion model outperforms existing methods. We deploy four models on historical epidemic data of six countries, respectively. For each country, the data spans from the day of the first confirmed case to August 31, 2020, is used for training models, while observed data spans from September 1, 2020, to September 30, 2020, is used to compare the performance between different methods.

Figure 2 demonstrates the dynamics of active confirmed cases (i.e., compartment D) in each country predicted by different models. Consistent with previous works [13], we further calculate the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2_score to evaluate the performance of different models when predicting COVID-19 epidemic dynamics. Table 2 presents different error measures of different models, from which we can make two observations. First, among different models for predicting epidemic dynamics, our proposed SEIDR-guided GRU fusion model obtains the best performance in all six countries. Second, among six different countries, the prediction of the dynamics of Russia is the worst one, while Italy gets the best performance.

Based on the epidemic dynamics predicted by the SEIDR-guided fusion model (shown in Figure 2 and the parameters of six countries inferred by the fusion model (shown in Table 3), we can make several observations for each country.

(i) In Italy, the daily death rate is about 0.67% and the recovery rate is about 2.7%. Around 36% of infected people can be confirmed per day. The predicted cases on September 30 are 51,000, while the ground truth is 51,263.

(ii) In Germany, the daily death rate is about 0.32%, while the recovery rate is around 6.1%. The active confirmed cases we predicted on September 30 is 25,800, while the ground truth is 26,557.

(iii) In Russia, the daily death rate is about 0.07% and the daily recovery rate is about 3.2%. The active confirmed cases we predicted on September 30 is 195,400, while the ground truth is 195,381.

(iv) In Switzerland, the active confirmed cases we predicted on September 30 is 8,300, while the ground truth is 8,508. Besides, the daily death rate in Switzerland is about 0.38%, while the daily recovery rate is about 5.6%.

(v) In Denmark, the daily death rate is about 0.35% and the daily recovery rate is about 6.6%. According to the prediction, the active confirmed cases in Denmark will increase rapidly and get around 6,450 cases on September 30, this number even higher than the previous peak on April 10, 2020.

(vi) As for Austria, the daily death rate is about 0.24% and the daily recovery rate is about 5.2%. Our model predicts that Austria's active confirmed cases can get about 9,700 on September 30, while the ground truth is 8,370.

We can also observe some differences in the parameters of different countries from Table 3. For example, Russia's α is $5x$ smaller compared to other places, and the reason can be summarized by referring to the news on the web: (1) Russians tend to see their doctor soon after symptoms appear (https://edition.cnn.com/2020/05/13/opinions/russia-low-covid-19-mortality-rate-sepkowitz/index.html); (2) Russia has far fewer elderly people, who are especially vulnerable to
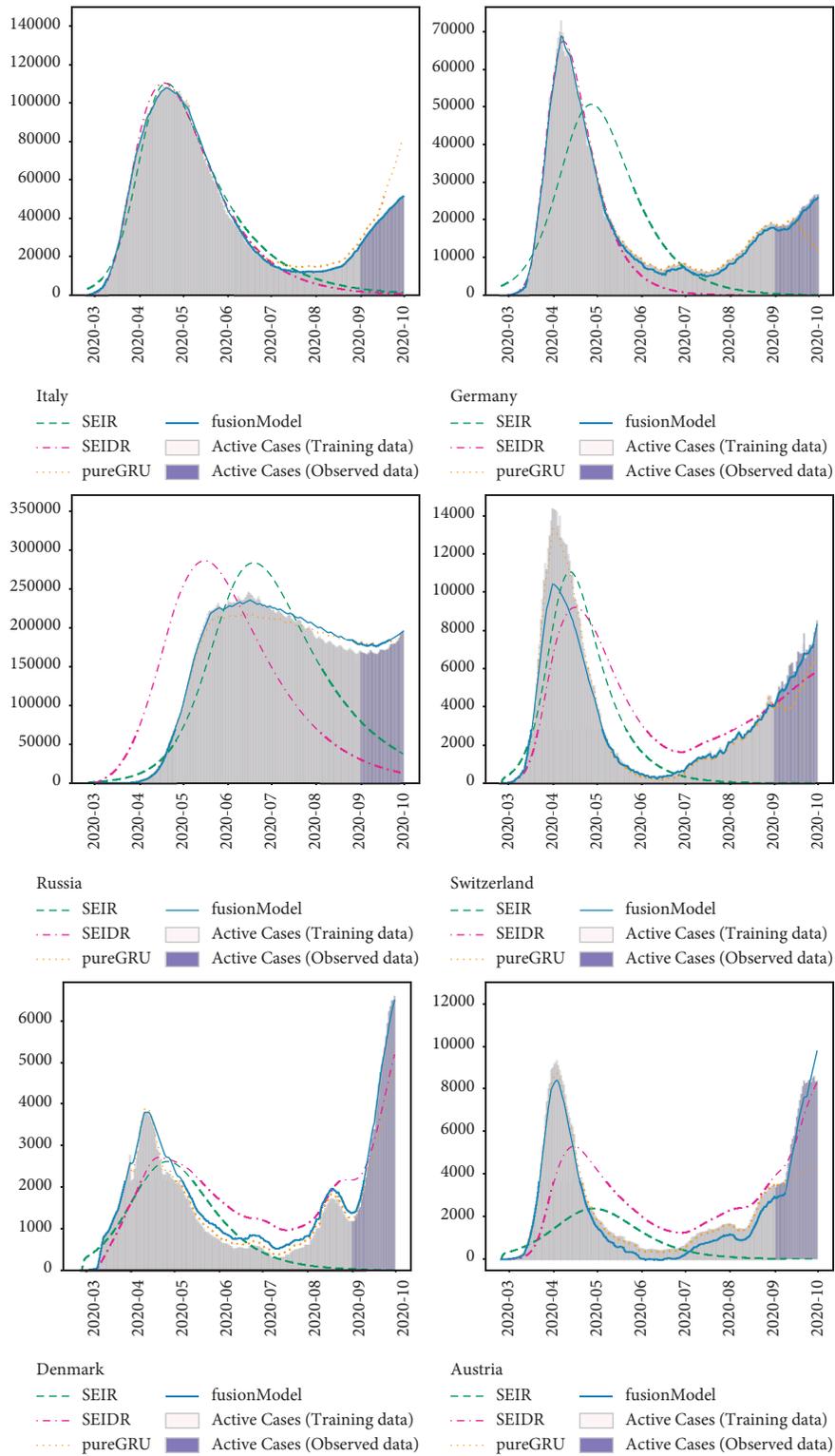
FIGURE 2: The comparison between our proposed fusion model and other baseline models about the prediction of active confirmed cases.

TABLE 2: Comparison of SEIR model, SEIDR model, GRU model, and our fusion model in terms of different error measures.

Countries	MAE				RMSE				r2_score			
	SEIR	SEIDR	GRU	Fusion	SEIR	SEIDR	GRU	Fusion	SEIR	SEIDR	GRU	Fusion
Italy	37616.4	38646.6	11228.4	202.5	38428.7	39396.5	14394.0	205.5	−26.9	−28.4	−2.9	**0.99**
Germany	21137.0	21404.1	4603.8	902.0	21359.1	21610.5	6700.1	913.7	−50.5	−51.7	−4.1	**0.91**
Russia	117079.8	152383.0	9033.7	8849.3	118800.6	152978.3	9684.4	9382.0	−172.5	−286.8	−0.2	**−0.08**
Switzerland	6212.2	1198.2	1422.2	310.6	6316.6	1361.1	1532.5	399.7	−29.7	−0.4	−0.8	**0.87**
Denmark	3640.5	855.2	830.4	155.3	4096.6	976.3	1090.0	175.2	−3.8	0.7	0.7	**0.98**
Austria	6151.6	755.6	2286.5	617.6	6483.2	861.7	2859.6	657.1	−9.1	0.8	−1.0	**0.89**

TABLE 3: The parameters of six European countries inferred by the SEIDR-guided GRU model.

Parameter	Meanings	Italy %	Germany %	Russia %	Switzerland %	Denmark %	Austria %
α	Death rate	0.67	0.32	0.07	0.38	0.35	0.24
γ	Recovery rate	2.7	6.1	3.2	5.6	6.6	5.2
ε	Transition probability: from exposed population to infective population	97	98	94	42	95	97
d_{iq}	Transition probability: from infective population to diagnosed population	36	38	32	98	23	97
δ	The ratio of the infective capacity between the exposed and the infected population	160	154	150	118	58	269

the virus (https://health.economictimes.indiatimes.com/news/industry/why-is-russias-coronavirus-death-rate-so-low/75748618); (3) Russia has a conservative count method; they attribute fatalities to the coronavirus only when death can be directly linked (https://www.scmp.com/news/world/russia-central-asia/article/3084458/why-russias-coronavirus-death-rate-so-low).

4.4. Performance of COVID-19 Prediction over Time.
To answer RQ2, we conduct two experiments to investigate how time factors influence the prediction performance of our fusion approach. First, we conduct the experiment to test how many days of data our approach needs to train a model with good performance. Specifically, we select the data 30, 45, 60, 90, and 180 days before the second wave epidemic as the training set (e.g., 30, 45, 60, 90, and 180 days before 2020.9.1) to predict active confirmed cases within the first month of the second wave (e.g., 2020.9.1 to 2020.9.30). Figure 3 presents the predicted active confirmed cases curves of models with the different training sets. The result shows that our fusion approach can achieve satisfying performance when the training data is longer than 60 days, which is comparable to the state-of-the-art COVID-19 prediction methods [9, 10, 14, 15].

Second, we conduct the experiment to test whether our fusion model can achieve satisfying performance at the different stages of the pandemic (e.g., in the beginning, at the peak, and in the reducing period of the pandemic). Specifically, we select 4 time points representing the stage before the pandemic (2020.7.15), at the beginning (2020.9.1), at the peak (2020.9.15), and in the reducing period of the pandemic (2020.6.1). Given that our dataset does not include the reducing phase of the second wave pandemic, we use the reducing period of the previous pandemic instead). For each time point, we select the data just before this time as the training set and predict active confirmed cases within 30 days. Figure 4 presents the predicted active confirmed cases curves of models with the different predicting points. The result shows that our fusion approach can achieve satisfying performance at the different stages of the COVID-19 pandemic. It means that our fusion model can accurately predict the dynamics of the COVID-19 epidemic over time.

4.5. Effective Reproductive Number Analysis.
To answer RQ3, we further analyzed the COVID-19 effective reproduction number of six selected European countries. The effective reproduction number (as known as R_t) represents the average number of new COVID-19 infections caused by an infectious individual. The effective reproduction number is often used to analyze the severity of the spread of a pandemic in an area [35]. In general, $R_t > 1$ means the epidemic is spreading, while $R_t < 1$ means the epidemic will disappear someday in the future. In this work, we calculated the curve of the effective reproduction number over time for these countries based on the parameters inferred by the SEIDR-guided GRU fusion model. The effective reproduction number can be calculated as the following equation (35):

$$R_t = \frac{1}{\epsilon} \times \delta \times \lambda_f(t) + \epsilon \times \frac{1}{d_{iq}} \times \lambda_f(t). \quad (8)$$

Figure 5 shows the curve of the effective reproduction number overtime for six selected countries, and some observations can be found from these curves:

(i) R_t of Italy started at a peak of 1.46 and fell below 1 after March 12. However, it tends to increase again in recent days and fluctuate at the position near $R_t = 1$, which indicates that the epidemic may recur.

(ii) R_t of Germany reached a peak of 2.35 on February 28 and fell below 1 after March 24. However, R_t of
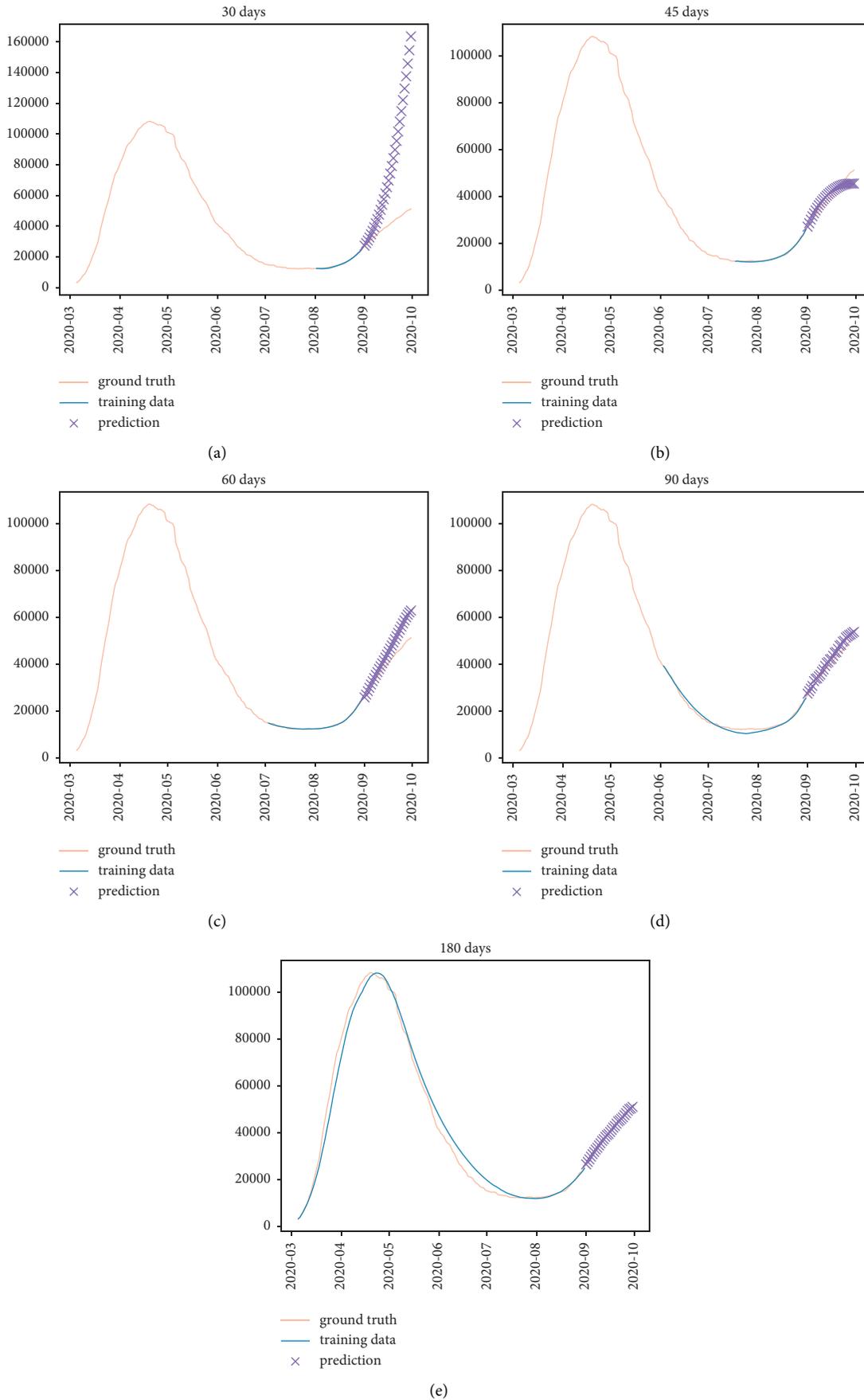
FIGURE 3: The performance of using training sets with different lengths to train the prediction model. (a) 30 days. (b) 45 days. (c) 60 days. (d) 90 days. (e) 180 days.
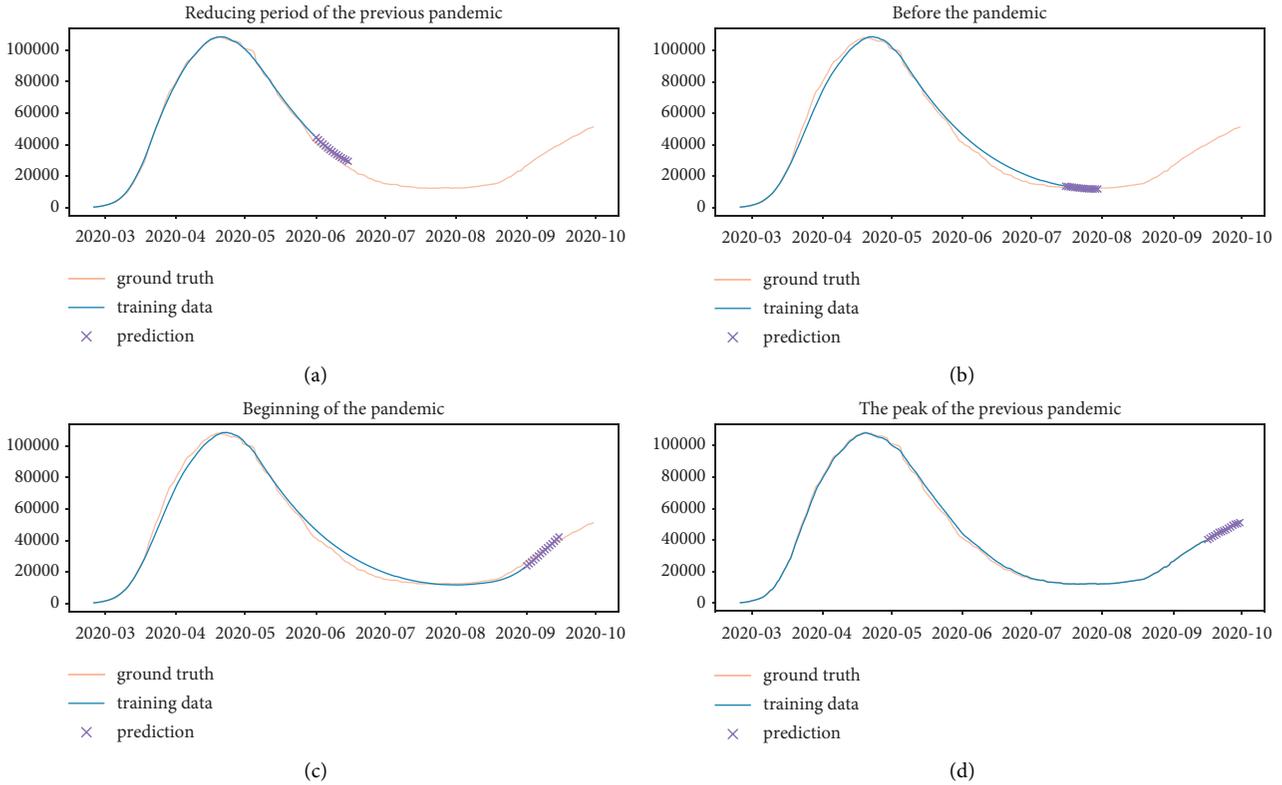
FIGURE 4: The performance of predicting active confirmed cases at the different stages of the COVID-19 pandemic. (a) Reducing period of the previous pandemic. (b) Before the pandemic. (c) Beginning of the pandemic. (d) The peak of the pandemic.

Germany started to increase after June and now it tends to fluctuate at the position near to the $R_t = 1$. This result reveals that the pandemic in Germany has a high risk of a second wave.

(iii) R_t of Russia reached a peak of 9.97 on February 28 and fell below 1 after May 8. Now it tends to remain stable at the position near to the $R_t = 1$. This result shows a high risk of the second wave of Russia's COVID-19 pandemic.

(iv) R_t of Switzerland started at a peak of 2.19 on February 25 and fell below 1 on March 6. After that, R_t of Switzerland increased again and exceeded 1 in mid-June. Now it tends to fluctuate at the position $R_t = 1$. This result indicates that the pandemic in Switzerland may have recurred.

(v) R_t of Denmark reached a peak of 6.37 on March 4 and fell below 1 on March 26. After that, R_t kept fluctuating at the position $R_t = 1$ and tends to increase, which indicates that the pandemic of Denmark may have recurred.

(vi) R_t of Austria started at a peak of 2.19 on February 25 and fell below 1 after March 23. Now it shows large fluctuations around $R_t = 1$ and tends to increase. This result reveals that the pandemic in Austria has a high risk of a second wave.

4.6. Findings and Implications. To answer RQ4, we extract two findings from the above prediction and analysis of active confirmed cases and effective reproductive number, which can be summarized as follows:

4.6.1. Revealing the Risk of the Second Wave of the Epidemic in Europe. Our proposed SEIDR-guided GRU fusion model reveals the risk of the second wave of the epidemic in Europe by predicting the active confirmed case and analyzing the effective reproductive number of six European countries. On the one hand, our fusion model predicts that the active confirmed cases of six countries will all increase in the future. For example, we predict that the active confirmed cases in Switzerland will increase rapidly and surpass the peak reached six months ago in the near future. On the other hand, effective reproductive number analysis based on our proposed model shows that the R_t of Italy, Germany, Switzerland, Denmark, and Austria are greater than 1 again and tend to keep fluctuating around the position $R_t = 1$. Although Russia's R_t is less than 1, it tends to remain stable at the position near to $R_t = 1$. All these results reveal the high risk of the second wave of the COVID-19 epidemic.

4.6.2. Indicating the Importance of the Social Distancing to Control the COVID-19 Epidemic. Our proposed SEIDR-
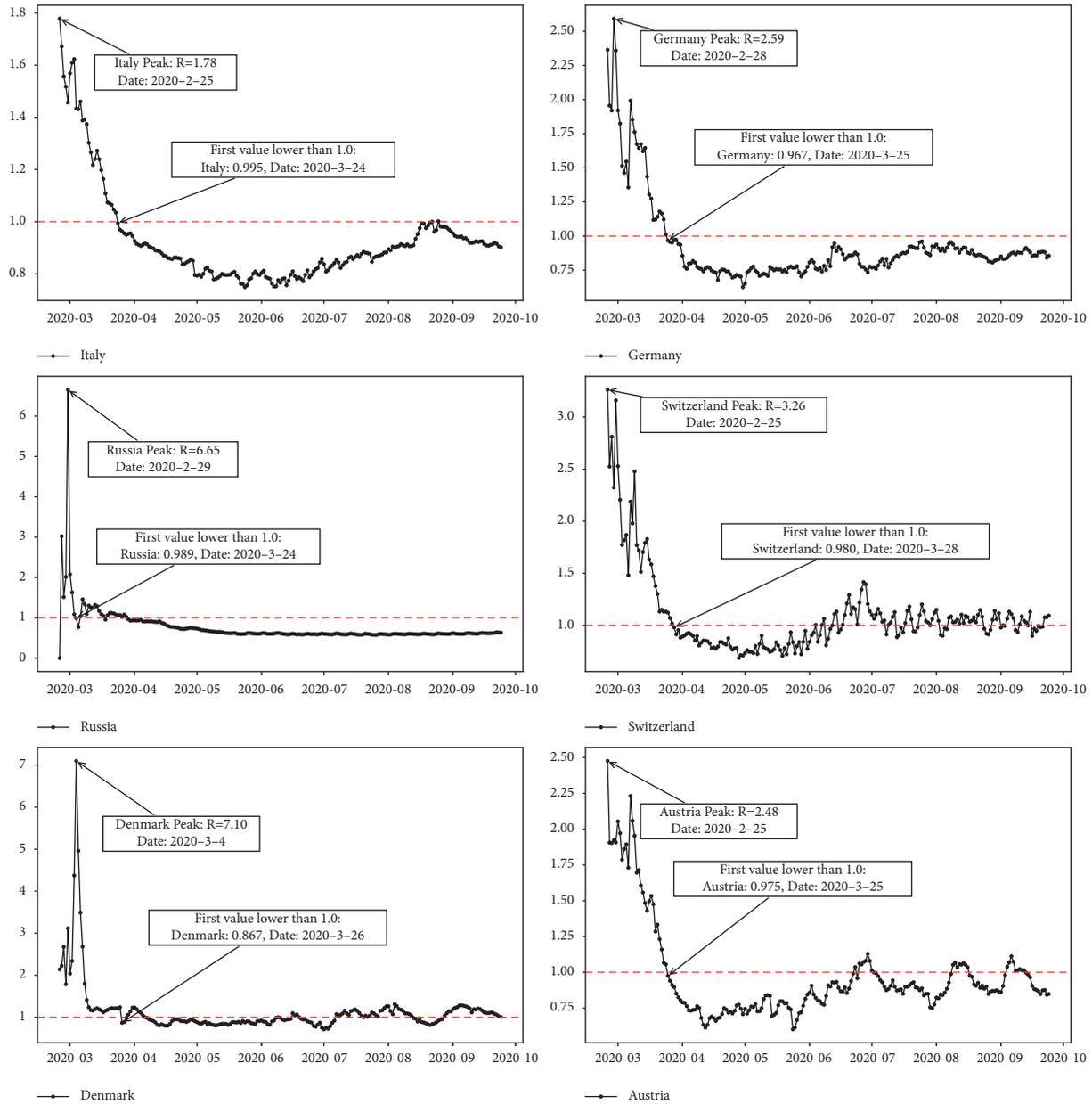
FIGURE 5: The curve of the effective reproduction number over time for six countries. The red dotted line marks $R_t = 1$.

guided GRU fusion model also indicates the importance of social distancing to control the COVID-19 epidemic. In this work, the SEIDR model divides infected people who are not in the quarantine part into the exposed population E (i.e., asymptomatic infected population) and infective population I (i.e., symptomatic infected population), and the parameter δ is introduced to represent the ratio of the infective capacity between E and I. Previous works show that E has less infective capacity than I due to fewer virus carries, so that δ can be less than 1 [31]. However, our SEIDR-guided GRU fusion model finds that δ of most countries is greater than 1 by fitting historical COVID-19 epidemic data. For example, δ of Austria and Russia is around 2.7 and 1.5, respectively. This result indicates that E has a more infective capacity than I.

Some recent works about COVID-19 get a similar finding. For example, Yang et al. found that the I population may be influenced by symptoms and reduce their actives, while the E population may not [7]. E may contact with more people, which leads to a higher infective capacity. This result indicates that it is hard to control the COVID-19 pandemic without reducing E's interaction with others. Especially, the asymptomatic infected population represented by E is a very high proportion, according to a recent investigation [17, 18].

We further investigate whether reducing E's mobility can slow down the spread of the epidemic by testing how the parameter δ influences the SEIDR model's prediction. Specifically, we turn down δ for Austria and Switzerland's SEIDR model and compare the active case prediction before
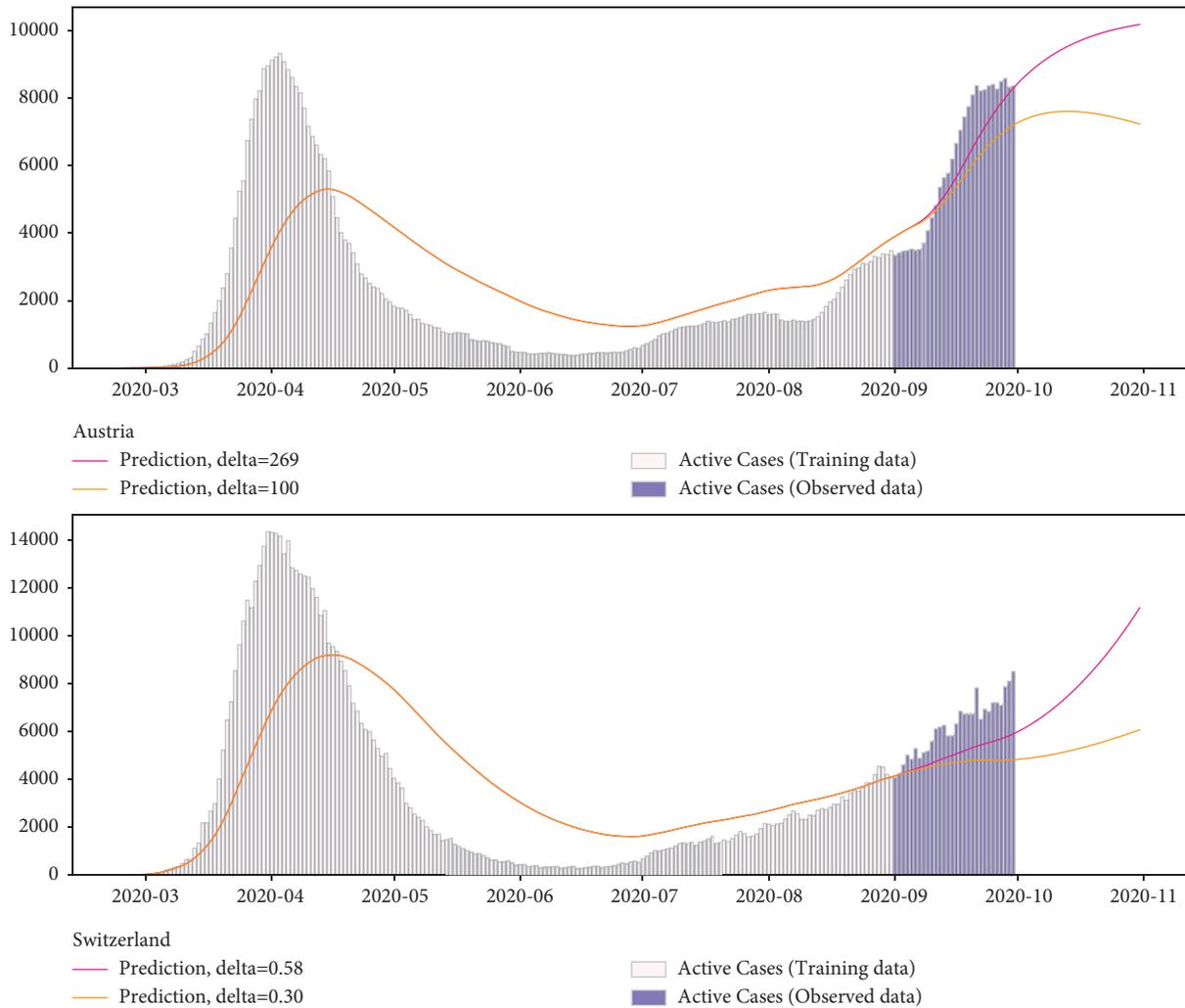
FIGURE 6: The active confirmed case prediction of Austria and Switzerland before and after turning down the parameter δ.

and after adjusting the parameter. As shown in Figure 6, the SEIDR model's prediction of future infections was reduced after adjusting the parameters. This result indicates that reducing E's mobility is meaningful for controlling the COVID-19 pandemic. Thus, we recommend taking more actions to reduce E's mobility, such as conducting social distancing or increasing the number of tests to filter out the asymptomatic infected population.

5. Discussion

In this section, we point out some limitations and present promising research directions for future work.

Evaluation in Other Continents. Through this study on historical epidemic data of six European countries, we propose a fusion model to predict and analyze the dynamics of COVID-19 by extracting epidemic-related knowledge from SEIDR compartmental model to guide the training of the GRU model. However, different continents in the world have conducted various antiepidemic measures according to their economic,

cultural, and medical conditions during this pandemic, which leads to different transmission patterns. A question may arise as to if this SEIDR-guided GRU fusion model can successfully predict and analyze the COVID-19 pandemic in more countries. Hence, one promising research direction is to investigate the generalizability of this model by collecting more data on other countries in the world and then study if the findings and model can be transferred or if any other interesting findings can be discovered.

Multicountry Learning Approach. In the current study, we trained the SEIDR-guided GRU model separately for each European country. This is because the model's parameters may be quite different in different countries, which can be related to their medical resources, populations, and antiepidemic measures. Compared to training a set of parameters for all countries, training parameters for each country individually may help get a more accurate prediction and analysis. However, common knowledge of different countries still exists (i.e., different countries have the same basic

TABLE 4: The MAE of different models when using the epidemic data between 2020.6.1 and 2020.8.31 for training.

Countries	SEIR	SEIDR	GRU	Fusion
Italy	24836.92	108980.43	11227.81	**207.23**
Germany	1701.73	10291.74	4626.28	**895.90**
Russia	44457.38	29481.63	9027.09	**8859.54**
Switzerland	721.54	1115.90	1403.00	**310.15**
Denmark	1629.13	2211.38	828.36	**162.34**
Austria	1270.93	1245.14	2292.55	**618.89**

reproduction number). Thus, another promising research direction combines epidemic data from multiple countries to improve the prediction and analysis of the COVID-19 dynamics. For example, we can utilize multitask learning to handle prediction tasks of different countries simultaneously.

Over Different Periods. In this study, we select the time span just before the second wave epidemic as training data to predict active confirmed cases within the first month of the second wave outbreak in order to analyze the risk of the second wave outbreak. However, the effectiveness of our model over different periods is still worth studying, which can be seen as another promising research direction. To address this problem, we use the data between 2020.6.1 to 2020.8.31 for training, and the results shown in Table 4 indicate that the fusion model is still the best among all methods.

Enhancing Online Epidemic Supervising Systems. In this article, we proposed a fusion model that can predict the dynamic of the COVID-19 pandemic well while maintaining sufficient interpretability. Based on good prediction and interpretability, our approach can enhance the online epidemic supervising systems as follows. First, our work can improve the epidemic supervising system by providing more accurate prediction, which is the function that users most care about. Second, our work can enhance the epidemic supervising system by providing COVID-19 transmission patterns (i.e., interpretable parameters of our fusion model such as the effective reproductive number R_0 and death rate). These patterns may offer valuable insights into future epidemic prevention and help public health experts better understand the epidemic. Third, our model provides the ability to simulate epidemic development in the supervising system. Users can explore the possible effects of an epidemic prevention measure by adjusting different model parameters and observing the prediction results.

6. Conclusions

In this study, we propose a fusion model, which can extract epidemic knowledge from the state-of-the-art compartmental model to guide the training. Doing so can preserve the interpretability of the fusion model and further reduce the overfitting problem. To evaluate our model, we collected historical epidemic data from six European countries from the day of the first case appearing in each country to October 30, 2020. Based on these epidemic data from the web, we demonstrate that our model can achieve state-of-the-art performance on several tasks, including the prediction of the active confirmed cases, and the analysis of the effective reproductive numbers. Our results reveal the risk of the second wave of the COVID-19 epidemic in Europe and justify the importance of social distancing to control the outbreak of the epidemic from the perspective of mathematical modeling.

Data Availability

The authors used the JHU CSSE COVID-19 dataset as the source data, which contains time series data of daily confirmed, recovered, and deaths all over the world. The link to this dataset is https://coronavirus.jhu.edu/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] European Centre for Disease Prevention and Control, "Covid 19 Situation Update Worldwide," [Online]. Available: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases, 2020.

[2] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[3] J. Hopkins University, "Covid-19 Dashboard by the center for Systems Science and Engineering," 2021, https://coronavirus.jhu.edu/map.html.

[4] Worldometer, "Covid-19 Coronavirus Pandemic," 2021, https://www.worldometers.info/coronavirus/.

[5] World Health Organization, "Coronavirus Disease (Covid-19) Data," 2020, https://www.who.int/data/gho/.

[6] COVID-19-Projections.com, "Covid-19 Projections Using Machine Learning," 2021, https://covid19-projections.com/.

[7] Z. Yang, Z. Zeng, Ke Wang et al., "Modified seir and ai prediction of the epidemics trend of covid-19 in China under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, pp. 165–174, 2020.

[8] J. Wangping, H. Ke, S. Yang et al., "Extended sir prediction of the epidemics trend of covid-19 in Italy and compared with hunan, China," *Frontiers of Medicine*, vol. 7, p. 169, 2020.

[9] G. Pandey, P. Chaudhary, R. Gupta, and S. Pal, "Seir and Regression Model Based Covid-19 Outbreak Predictions in india," 2020, https://arxiv.org/abs/2004.00958.

[10] C. Hou, J. Chen, Y. Zhou et al., "The effectiveness of quarantine of wuhan city against the corona virus disease 2019 (covid-19): a well-mixed seir model analysis," *Journal of Medical Virology*, vol. 92, no. 7, pp. 841–848, 2020.

[11] F. Brauer, C. Castillo-Chavez, and C. Castillo-Chavez, *Mathematical models in population biology and epidemiology*, Vol. 2, Springer, Berlin/Heidelberg, Germany, 2012.

[12] W. Ogilvy Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epi- demics," *Proceedings of the Royal Society of London - Series A: Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927.

[13] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm," *Chaos, Solitons & Fractals*, vol. 140, 2020.

[14] P. Arora, H. Kumar, and B. K. Panigrahi, "Prediction and analysis of covid-19 positive cases using deep learning models: a descriptive case study of India," *Chaos, Solitons & Fractals*, vol. 139, 2020.

[15] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, "Deep learning methods for forecasting covid-19 time-series data: a comparative study," *Chaos, Solitons & Fractals*, vol. 140, 2020.

[16] C Vinay Kumar Reddy and L. Zhang, "Time Series Forecasting of Covid-19 Transmission in canada Using Lstm Networks," *Chaos, Solitons & Fractals*, vol. 135, Article ID 109864, 2020.

[17] M. Day, "Covid-19: identifying and isolating asymptomatic people helped eliminate virus in Italian village," *BMJ*, vol. 368, 2020.

[18] K. Mizumoto, K. Kagaya, A. Zarebski, and G. Chowell, "Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yoko- hama, Japan, 2020," *Euro Surveillance*, vol. 25, no. 10, Article ID 2000180, 2020.

[19] M. Raissi, N. Ramezani, and P. Seshaiyer, "On parameter estimation approaches for predicting disease transmission through optimization, deep learning and statistical inference methods," *Letters in Biomathematics*, vol. 6, no. 2, pp. 1–26, 2019.

[20] S. Uhlig, K. Nichani, C. Uhlig, and K. Simon, *Modeling Projections for Covid-19 Pandemic by Combining Epidemiological, Statistical, and Neural Network Approaches*, Cold Spring Harbor Laboratory Press, Long Island, New York USA, 2020.

[21] C. Distante, I. G. Pereira, L. M. Garcia Goncalves, P. Piscitelli, and A. Mi- ani, *Forecasting Covid-19 Outbreak Progression in Italian Regions: A Model Based on Neural Network Training from Chinese Data*, Cold Spring Harbor Laboratory Press, Long Island, New York USA, 2020.

[22] A. Godio, F. Pace, and A. Vergnano, "Seir modeling of the Italian epidemic of sars-cov- 2 using computational swarm intelligence," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, p. 3535, 2020.

[23] A. E Fahmy, M. M El-desouky, and M Ahmed SA, *Epidemic Analysis of Covid-19 in egypt, qatar and saudi arabia Using the Generalized Seir Model*, Cold Spring Harbor Laboratory Press, Long Island, New York USA, 2020.

[24] L. Peng, W. Yang, D. Zhang, C. Zhuge, and H. Liu, "Epidemic Analysis of Covid-19 in china by Dynamical Modeling," 2020, https://arxiv.org/abs/2002.06563.

[25] L. López and X. Rodo, "A modified seir model to predict the covid-19 outbreak in Spain and Italy: simulating control scenarios and multi-scale epidemics," *Results in Physics*, vol. 21, 2020.

[26] S. He, Y. Peng, and K. Sun, "Seir modeling of the covid-19 and its dynamics," *Nonlinear Dynamics*, vol. 101, no. 3, pp. 1667–1680, 2020.

[27] H. Bouhamed, "Covid-19 cases and recovery previsions with deep learning nested sequence prediction models with long short-term memory (lstm) architecture," *Int. J. Sci. Res. in Computer Science and Engi- neering*, vol. 8, no. 2, 2020.

[28] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo, "Multiple ensemble neural net- work models with fuzzy response aggregation for predicting covid-19 time series: the case of Mexico," *Healthcare*, vol. 8, no. 2, p. 181, 2020.

[29] N. Singh Punn, S. Kumar Sonbhadra, and S. Agarwal, *Covid-19 Epidemic Analysis Using Machine Learning and Deep Learning Algorithms"*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N Y, USA, 2020.

[30] Q. Li, X. Guan, P. Wu et al., "Early transmission dynamics in wuhan, China, of novel coronavirus– infected pneumonia," *New England Journal of Medicine*, vol. 382, no. 13, pp. 1199–1207, 2020.

[31] J. Zhang, J. Lou, Z. Ma, and J. Wu, "A compartmental model for the analysis of sars transmis- sion patterns and outbreak control measures in China," *Applied Mathematics and Computation*, vol. 162, no. 2, pp. 909–924, 2005.

[32] M. Renardy, M. Eisenberg, and D. Kirschner, "Predicting the second wave of covid-19 in washt- enaw county, mi," *Journal of Theoretical Biology*, vol. 507, 2020.

[33] G. Cacciapaglia, C. Cot, and F. Sannino, "Second wave covid-19 pandemics in europe: a temporal playbook," *Scientific Reports*, vol. 10, no. 1, p. 15514, 2020.

[34] D. P. Kingma and Ba Jimmy, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[35] K. M. Gostic, L. McGough, E. B. Baskerville, A Sam, and J Keya, "Practical Considerations for Measuring the Effective Reproductive Number, Rt," *PLoS computational biology*, vol. 16, 2020.