Hindawi
Journal of Healthcare Engineering
Volume 2022, Article ID 8990907, 13 pages
https://doi.org/10.1155/2022/8990907

Research Article

Early Warning of Infectious Diseases in Hospitals Based on Multi-Self-Regression Deep Neural Network

Mengying Wang,[1] Cuixia Lee,[2] Wei Wang,[3] Yingyun Yang,[1] and Cheng Yang[1]

[1]*State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China*
[2]*Peking University Third Hospital, Beijing, China*
[3]*Goodwill Hessian Health Technology Co, Ltd., Beijing, China*

Correspondence should be addressed to Cheng Yang; chy@cuc.edu.cn

Received 6 April 2022; Accepted 11 July 2022; Published 18 August 2022

Academic Editor: Yi-Zhang Jiang

Objective. Infectious diseases usually spread rapidly. This study aims to develop a model that can provide fine-grained early warnings of infectious diseases using real hospital data combined with disease transmission characteristics, weather, and other multi-source data. *Methods*. Based on daily data reported for infectious diseases collected from several large general hospitals in China between 2012 and 2020, seven common infectious diseases in medical institutions were screened and a multi self-regression deep (MSRD) neural network was constructed. Using a recurrent neural network as the basic structure, the model can effectively model the epidemiological trend of infectious diseases by considering the current influencing conditions while taking into account the historical development characteristics in time-series data. The fitting and prediction accuracy of the model were evaluated using mean absolute error (MAE) and root mean squared error. *Results*. The proposed approach is significantly better than the existing infectious disease dynamics model, susceptible-exposed-infected-removed (SEIR), as it addresses the concerns of difficult-to-obtain quantitative data such as latent population, overfitting of long time series, and considering only a single series of the number of sick people without considering the epidemiological characteristics of infectious diseases. We also compare certain machine learning methods in this study. Experimental results demonstrate that the proposed approach achieves an MAE of 0.6928 and 1.3782 for hand, foot, and mouth disease and influenza, respectively. *Conclusion*. The MRSD-based infectious disease prediction model proposed in this paper can provide daily and instantaneous updates and accurate predictions for epidemic trends.

1. Introduction

Infectious diseases are usually characterized by rapid transmission, high morbidity, and high uncertainty and are extremely dangerous. The global health sector is currently working to promote early warning and surveillance capabilities for infectious disease outbreaks. Medical institutions are the frontline units for detecting, reporting, and treating patients with infectious diseases, and they are also responsible for the routine diagnosis and treatment of noncommunicable diseases to ensure a harmonious society [1]. China has established a relatively well-developed national infectious diseases information monitoring system (NIDIMS), owing to which the earlier practice of cascading reporting of infectious diseases has been changed to direct reporting to the government [1].

Through the NIDIMS, infectious diseases are reported directly to the government [1, 2]. A total of 40 types of infectious diseases are covered, including two types of class A infectious diseases, 27 types of class B infectious diseases, and 11 types of class C infectious diseases [1]. As early as the end of 2006, the direct reporting network covered 100% of disease control centers, 95% of medical and health institutions at the county level and above, and 70% of rural health centers nationwide [2], making the reporting of infectious diseases 10 times faster [3]. In 2020, COVID-19 exposed the disadvantages of the system: it was not equipped with active alerting and intelligent analysis of predetermined warnings

[4], which impedes the timeliness of diagnosis and the accuracy of risk prevention and control, resulting in the possibility of underreporting or delayed reporting [3]. The existing infectious disease warnings are also mostly government warnings regarding the national situation. It is difficult to provide actual feedback on the prevailing situation in local medical institutions as it develops, and there is a lack of efficient early warning based on the data obtained from medical institutions. As a result, the existing system cannot support the deployment of hospitals for epidemic prevention and control, which will substantially weaken the role of general hospitals as frontlines for epidemic prevention and control.

2. Related Work

To address these issues, most of the existing studies adopt polynomial fitting, mathematical statistical models, SEIR methods, and machine learning. Dette et al. [5] applied a polynomial function to fit the curve, which is a traditional statistical method that uses only the data series of the number of confirmations. By increasing the number of polynomials, it is possible to fit a more complex function curve that cannot be predicted flexibly in practice according to the current time. At present, infectious disease dynamics models are also widely applied, such as SEIR model, susceptible-exposed-infected-recovered-die-hard-infected (SEIRD) model, and their variants. For example, Ghostine et al. [6], Wangping et al. [7], Wei et al. [8], Yang et al. [9], and Youssef et al. [10] adopted the SEIR method to predict the spread of COVID-19. Despite the small number of parameters in the kinetic model, the actual meaning of the parameters is explicit, and the level of interpretability is high. However, there must be multiple types of data available, including the latent population, the number of recovered patients, and other group data that are difficult to obtain. For example, Feng et al. [11] obtained the latent data by estimation, which affects accuracy, thus restricting the scope of application. In some studies, the autoregressive integrated moving average (ARIMA) model [12, 13], linear regression [14], moment estimation [15], hidden Markov model [16], and grey self-memory coupling model [17] were adopted. However, there is a limit to the scope of application for each of these mathematical models, which means they are often suited to a single or certain type of disease. Not only does the data used in the research have a single dimension, but it is also heavily reliant on the information obtained from epidemiological retrospective surveys. Erraguntla et al. [18] and Telarolli et al. [19] treated the data series of developmental trends exhibited by infectious diseases as random ones, and the autoregressive model was adopted to analyze the interdependence and autocorrelation between various random variables. However, stable time series are required for this method. In the modeling process, only the number of deaths and that of people who have recovered are used, while the information other than the numerical sequence is ignored.

Since the outbreak of the COVID-19 pandemic, there are many scholars [20–24] from various countries who have studied machine learning and neural network to predict the developmental trend of COVID-19 and other infectious diseases. However, this method is disadvantaged by long-time series and the lack of data regarding important influencing factors in the developmental trend of infectious diseases, such as environment and climate. As a result, the prediction results can only be obtained on a monthly basis and are prone to overfitting. Gu et al. [25] used three-layer long short-term memory (LSTM) to model the developmental trend of hand, foot, and mouth disease, taking into account various external factors such as wind speed and temperature, with the root-mean-square error (RMSE) reaching 0.71. However, for children with obvious group characteristics, factors such as opening and closure of schools is not considered. Liao et al. [26] applied a neural network to learn the parameters in the dynamic model, which accelerates the adjustment of the parameters used in the dynamic model. However, in the process of parameter learning, only the product coefficient can be learned when there are multiple coefficients used in the dynamic model. Therefore, it is difficult to decompose multiple coefficients, which affects the interpretability. Bedi et al. [27] used the SEIR dynamic model and LSTM to study COVID-19, with satisfactory prediction results. However, constraints such as difficult access to population and limited data dimensions are yet to be addressed.

Considering the advantages and disadvantages of existing methods, an MSRD-based approach is proposed in this study to predict the developmental trend of infectious diseases. While incorporating the information on multi-dimensional epidemiological features closely associated with infectious diseases, the proposed model simultaneously adopts LSTM as the building block to construct a recurrent neural network. In addition, temporal modeling and calculation are performed for the time-series data organized by means of self-regression learning, which addresses the challenges facing existing studies, such as the difficulty in obtaining quantitative data such as latent population, the overfitting of long time series, and the lack of consideration given to the epidemiological features of infectious diseases and only a single series of the number of patients. In addition, compared with the SEIR model, machine learning model, and neural network model in existing studies, the MSRD-based approach proposed in this study achieves better performance.

3. Methods

3.1. Sources of Data. The data comes from two sources. One is the official monthly public health scientific data of the national center for disease control and prevention (CDC) from 2012 to 2017. The second is the daily data of inpatient and outpatient medical records from Peking University Third Hospital from 2012 to 2020. The hospital data was obtained from the hospital data center. The hospital data center adopts Hadoop architecture and integrates a Hadoop distributed file system, HBase column database, and Hive data warehouse, which can easily perform data storage and analytical computations [28]. After in-depth mining of 110,000 historical data points on infectious diseases

accumulated in a large hospital over eight years, big data technology is applied to collect and clean the clinical data and then store and manage them in a centralized way, which provides the necessary basis for the training and application of an early-warning model for infectious diseases. In addition to the data from the data center, this study collected the daily temperature, humidity, wind, and other climate data from the website of the national meteorological data department, given the close relationship between some infectious diseases and climate factors [29, 30]. In addition to the aforementioned number of infectious diseases and environmental factors, the spread of infectious diseases is closely related to human activities. For example, infectious diseases are more likely to spread widely in human aggregated activities. Therefore, this factor is also used as a data feature for prediction in this study, which is reflected in the study of some infectious diseases.

The study was approved by the Medical Science Research Ethics Committee of Peking University Third Hospital (serial number: IRB00006761-M2020318). All methods were performed in accordance with the relevant guidelines and regulations.

3.2. MSRD Model. In this study, we constructed an MSRD model using an LSTM neural network and a sliding window as the core structure, as shown in Figure 1. The MSRD model takes the time-series data of multiple elements from multiple days as inputs to predict the number of confirmed infectious diseases on the next day. The trend of this number confirmed that infectious diseases can be generated through continuous prediction. LSTM consists of an input gate, a forgetting gate, and an output gate, which can be used to preserve information and calculate the features of data with time-series characteristics. This study uses daily data as the experimental metadata. Each metadata is an $m \times 1$ vector that includes the number of confirmed infectious diseases on a particular day, the current date, current climate data, and the expected social activities. The data in the experiment includes D-Day. The total data occupies an $\mathrm{m} * D$ matrix. MSRD uses a window of size $\mathrm{m} * w$ to slide to the right within the total data of $\mathrm{m} * D$. Each slide extracts fragment data of size $\mathrm{m} * w$ as the input data and takes the confirmed number of infectious diseases in the first column of the $m \times 1$ vector outside the window as the label data. At the end of sliding, $D - w$ pieces of $\mathrm{m} * w$ input data matrix and $D - w$ label data can be obtained, as shown in the sliding window in Figure 1. Then, the $D - w$ pieces of $\mathrm{m} * w$ data are fed into the LSTM neural network, and the LSTM can learn the mapping function between the $\mathrm{m} * w$ segment data and the corresponding label. Next, to make use of the output of the LSTM neural network and the characteristic information of the original data simultaneously, the input data of $\mathrm{m} * w$ is expanded in a one-dimensional form and horizontally spliced with the output of the LSTM neural network, followed by the prediction result of the diagnosis number being output through the feedforward neural network with a rectified linear unit (ReLU) function as the activation function. Based on the original fragment data containing multi-day data, the MSRD model uses the unique structural characteristics of LSTM to extract the time-series features in the data. It can simultaneously use the historical time-series data and the multi-dimensional feature data outside the series to predict the epidemic trend of infectious diseases. Compared with the typical LSTM that uses only the output of the last time step, MSRD makes full use of the output of each time step of the sequence structure based on the use of time windows to extract data and feeds the original first-order data (not computed by LSTM neurons) and second-order data (computed by LSTM) together into a feedforward neural network with multiple hidden layers, enhancing the crossover capability of the features as well as the model fitting ability. In addition, the introduction of the time window concept improves the flexibility of the model in predicting infectious diseases and avoids the overfitting problem caused by using long series to train the model, allowing the model to make more accurate predictions for the future using data from different date spans according to different application scenarios.

3.3. Evaluation Metrics. We evaluate the performance of the proposed MSRD method and compared it with the three other models using the MAE, which is the simplest measure of fitting and prediction accuracy describing the mean value of the difference between the model prediction results and the true results at each time in terms of the series as a whole. In addition, we also use the RMSE to measure the deviation of the observed values from the ground truth, which is calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |\mathrm{actual}(t) - \mathrm{predict}(t)|, \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} [\mathrm{actual}(t) - \mathrm{predict}(t)]^2}. \tag{2}$$

3.4. Statistical Software. We used Microsoft Excel 2016 to construct the original database and used *Python* v3.6.4, PyTorch v1.6.0, Numpy v1.14.1, and Scikit-Learn v0.19.1 for model building.

4. Experimental Results

4.1. Trends in the Incidence of Real Infectious Diseases in Hospitals. Based on real hospital outpatient data collected from the Peking University Third Hospital from 2012 to 2020, we analyze seven common infectious diseases, including hand-foot-and-mouth disease (HFMD), influenza, viral hepatitis, infectious diarrheal disease, scarlet fever, syphilis, and tuberculosis during in this study, as shown in Figure 2. Among them, influenza and viral hepatitis have a trend of slow growth year by year, while HFMD, infectious diarrheal disease, scarlet fever, syphilis, and tuberculosis have decreased year by year, which sufficiently demonstrates the effectiveness of overall infectious disease prevention and

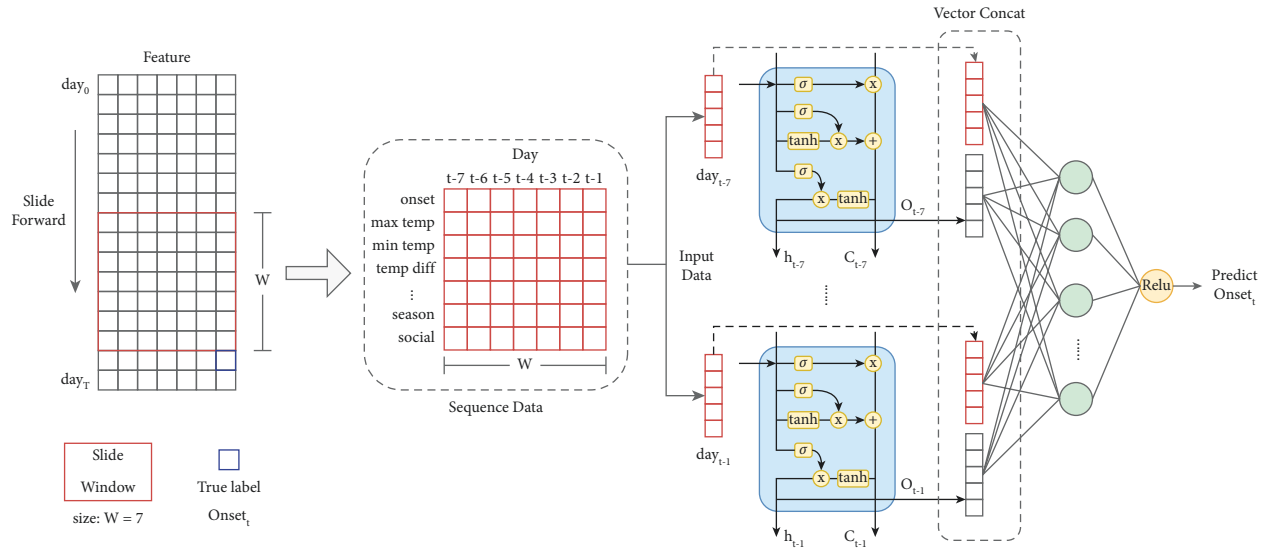

Figure 1: Multi self regression deep model.

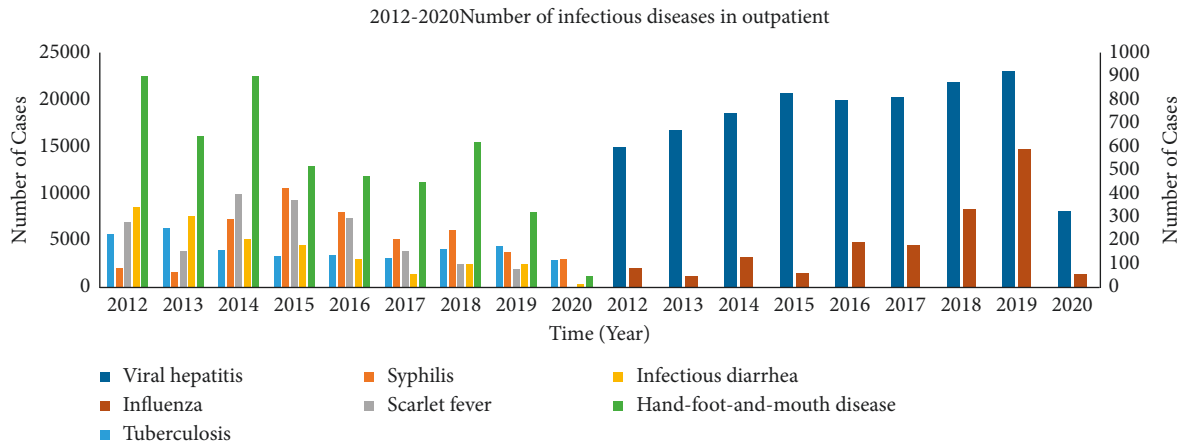

Figure 2: Number of infectious diseases in outpatients between 2012 and 2020.

control, but there is a trend of slow growth for influenza and viral hepatitis, and the government should take preventive measures in advance. The year 2020 is more affected by the COVID-19 epidemic, and the actual number of hospital visits has also decreased significantly. Based on the analysis of data between 2012 and 2020, we construct an MSRD model, where we tested multiple parameter combinations: the sliding window length w candidates (3, 5, 7, 9, 14), the number of LSTM neurons candidates (6, 8, 16, 32, 64, 128), the number of feedforward neural network neurons candidates (32, 64, 128, 256, 512), and multiple learning rates. The optimal values of the above parameters are as follows: the sliding window length is 7, the number of LSTM neurons is 16, the feedforward neural network neuron is 128, and the learning rate is 0.001. It was observed that as the sliding window length, the number of LSTM neurons and the number of feedforward neurons increased and the performance of the MSRD model in the training data improved, but the performance in the test set first increased and then decreased. The results demonstrate that the model complexity and the number of days of historical data used are positively correlated with the fitting ability of the model, while the generalization ability varies, into a trend of first increasing and then decreasing. Figure 3 shows the model trained using the optimal parameters described above to predict the prevalence of infectious diarrhea in 2021. The model predicts a continuous decrease in the number of infectious diarrheas in 2021. See Supplementary Material 1 for the prediction results of other infectious diseases.

4.2. Comparison of MSRD and SEIR Models. In the research of infectious disease prediction, SEIR kinetic model is quite common. The main idea is to divide the whole population into different groups in a closed system and design the population transfer coefficients among different groups according to the infectious disease transmission mechanism, so as to form the differential equations. According to the National Data Center for Public Health Sciences, we build the SEIR model for HFMD. As shown in Figure 4, the model has monthly granularity. We observe that the number of HFMD cases in the national data shows a consistently

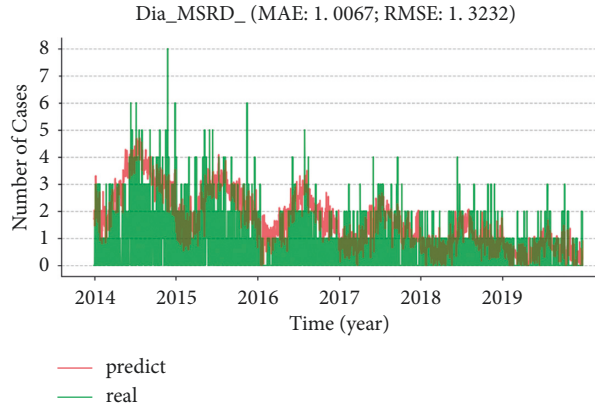

Figure 3: The MSRD model predicts the epidemic situation of infectious diarrhea in 2021. Overlapping contrast: the green line denotes the real data, and the red line shows the predicted data.

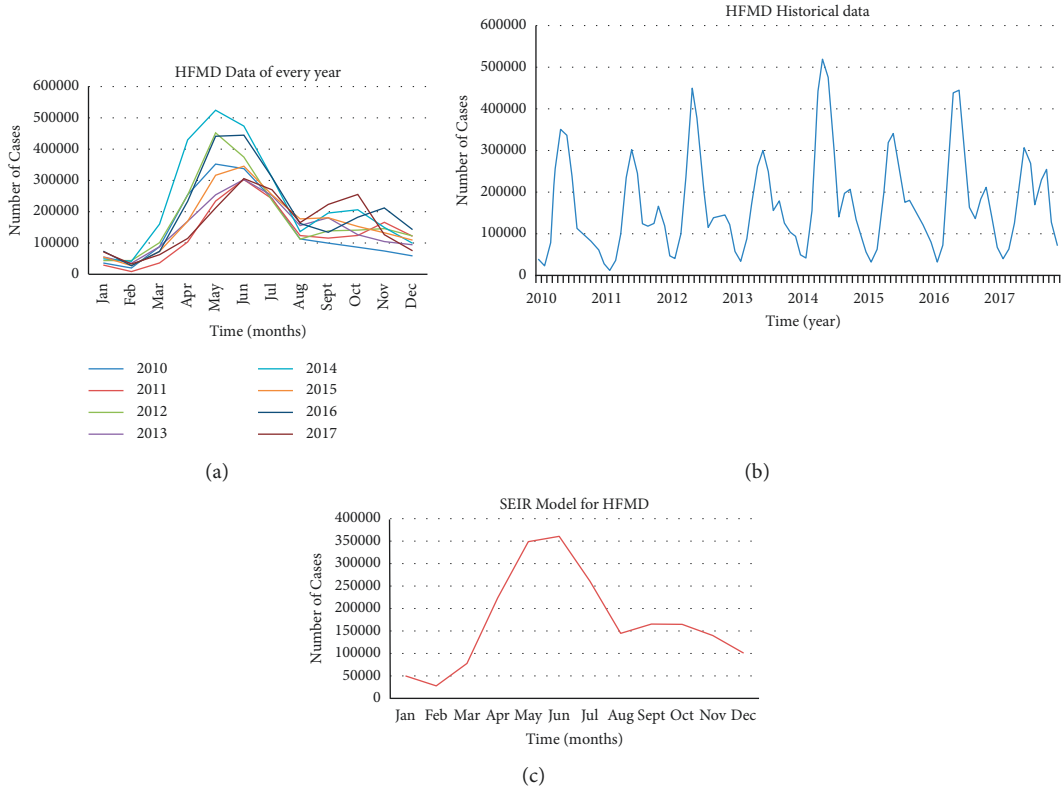

Figure 4: National monthly real data curve (a), continuous curve of monthly real data over the years in China (b), and SEIR model fitting curve (c).

increasing trend every year. The prediction trend after model learning is identical to the historical development trend. Therefore, it is feasible to use national data modeling. However, the national data alone has no daily data per month, while the fine-grained data are insufficient.

Based on the data of a hospital, here, we take HFMD as an example and use the "monthly granularity" for modeling, as shown in Figure 5. It is difficult to obtain the data of the exposed persons, which is necessary for SEIR. The SEIR model is used to model by stages, the results of which are given below. Figure 6 shows the daily data of real HFMD cases in the hospital.

It is clear that the SEIR model is poorly fitted to the infectious disease data of medical institutions for the following reasons: (1) there is an evident lack of continuity of infectious diseases in medical institutions. In the data of daily granularity, there are many months in which there are zero cases of HFMD. (2) The SEIR model requires population information such as susceptible groups, latent groups, isolated groups, infected groups, recovered groups, and dead groups, among which susceptible groups can be estimated; however, infected groups and recovered groups are very important, and these data are difficult to obtain in medical institutions. (3) For the infectious disease data of medical institutions with a

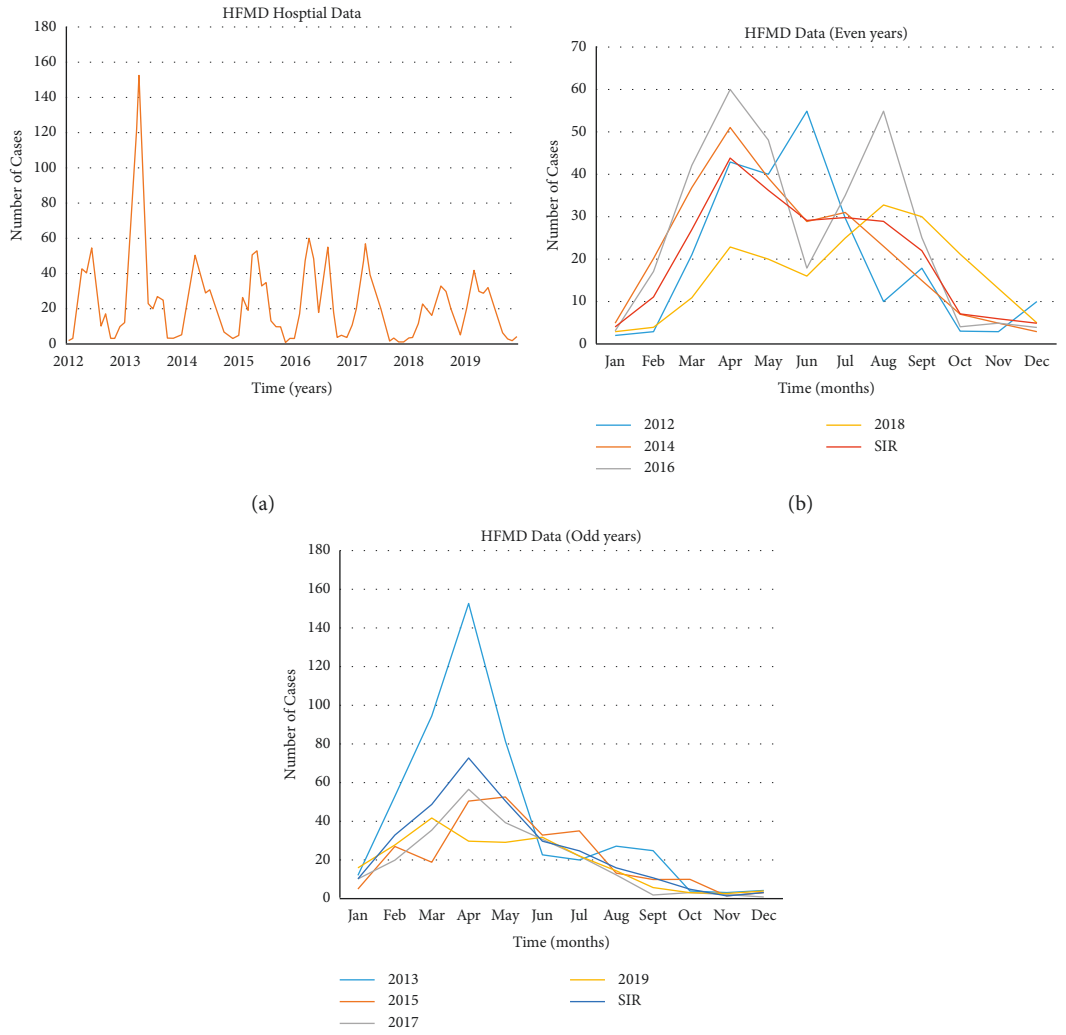

Figure 5: HFMD monthly curve of real data of the hospital (a), real and fitting data curve of even years and months (b), and real and fitting data curve of odd years and months (c).

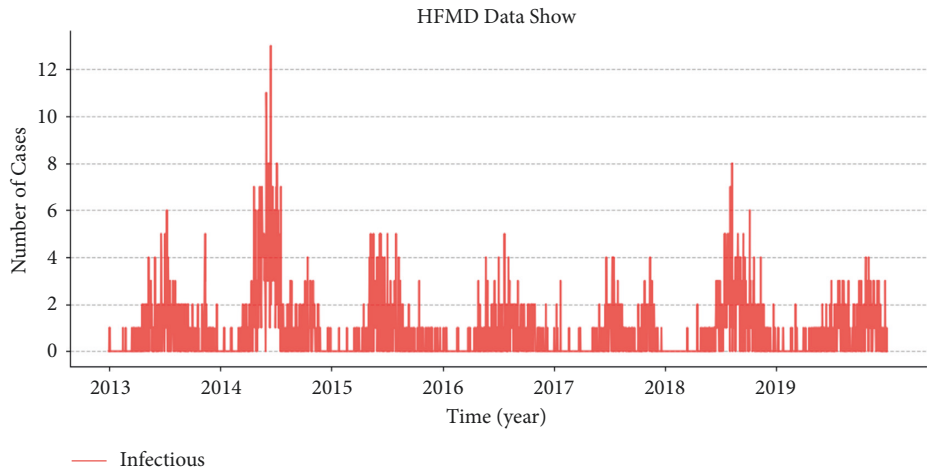

Figure 6: Daily data of real HFMD cases in the hospital.

small magnitude, the "daily granularity," according to the single-digit change, leads to significant fluctuations in the data curve. Therefore, the SEIR model is suitable for trend forecasting with a large order of magnitude and comprehensive population information and is not applicable to forecasting on the basis of medical institutions.

4.3. Prediction Effect of MSRD. Based on the MRSD model and careful feature engineering, it is eventually applied to the training of the model. Feature engineering is the processing of data for the purpose of maximizing the extraction of effective features from the raw data for use in algorithms and models. For the current study, the number of daily confirmed cases of each infectious disease from January 1, 2012, to December 31, 2019, was first counted from the hospital data center. After that, the current year, month, and season are extracted from them to identify the current year, month, and season based on the date. The year and month are characterized by numerical type, and the season is represented in the form of a unique heat code. Next, the collected climate data are processed to calculate the diurnal temperature difference of the day. Finally, considering that infectious diseases are closely related to human group activities, another feature of social factors is incorporated into the training data. For example, HFMD is a prevalent infectious disease among children and students, and the main group activity of adolescents is studying at school, so the feature of whether they are currently on vacation is added to the features to characterize the social gathering activity.

The features for training include the daily number of confirmed cases of the target infectious disease, the current year, the current month, the current season, the highest temperature of the day, the lowest temperature of the day, the temperature difference of the day, whether the students are currently on vacation, and so on. The above features are then constructed as time-based series data. The features we selected are based on the literature review [6, 18, 19, 25, 29, 31–33] and expert knowledge. The highest and lowest temperatures within a single day in conjunction with the season can describe the climate. The development trend of certain infectious diseases is clearly known to be related to the environment and climate change. Consequently, the introduction of these features can represent the prevalent environmental climate to a certain extent and thus correlate with the development trend of infectious diseases. The feature "intraday temperature difference" is introduced based on the disease characteristics of infectious diseases. For example, when the intraday temperature difference is large, people are more vulnerable to influenza. The feature "students are on holiday" represents social activity factors, in order to account for the fact that socializing can lead to the spread of infectious diseases. The features used in this study combine the climatic environment, the characteristics of infectious diseases, and social activity factors. In addition, in light of epidemics of infectious diseases that occurred in the past, comprehensive consideration from multiple perspectives can better help improve the prediction effect of the prediction model.

Finally, for the training of the MSRD model and the validation and evaluation of the model's effectiveness in predicting future infectious disease epidemic trends, data from October 28, 2013, to December 31, 2018, with a total of 1,890 time-series samples, accounting for 82.7% of the total time-series sample data, were used as training data in the study; a total of January 1, 2019, to January 31, 2020, 395 time-series samples were used as the test data. Figure 7 shows the test results of applying the MSRD model to predict the epidemic trends of HFMD and influenza. The MSRD model is selected with a sliding window length of 7; the number of LSTM neurons is 32; the number of feedforward neural network neurons is 128; and the model is trained at a learning rate of 0.001. The predicted infectious disease epidemic trends from the multi-dimensional autoregressive neural network model shown in the figure broadly match the actual trends. The MAE was used to evaluate the results in the regression prediction. In this study, the number of confirmed cases of HFMD and influenza were of different orders of magnitude, and the MAE of the multi-dimensional autoregressive neural network was 0.6928 and 1.3782 cases lower for the test data in the training of both, which means that the average difference between the number of diseases predicted by the model at each time and the real number of diseases on a corresponding day were 0.6928 and 1.3782 cases, respectively. The difference between the prediction results of HFMD and the trend of influenza is mainly attributable to the magnitude of the total number of cases being different and a higher number of breakpoints in the real data of HFMD. The highest number of single-day influenza cases is nearly 50, while that of HFMD is only 10. At the same time, there is not only a substantial discrepancy between the data of HFMD and influenza but also a large number of zero cases in the time series of HFMD data. As a result, the continuity of data is poor, thus making it more difficult for the model to capture the regularity of data.

5. Discussion

5.1. Analysis of Model Methods. The MSRD model proposed in this paper provides a better prediction result than the SEIR model. The SEIR dynamic model is suitable for the prediction of large orders of magnitude and the entire population, and the influence coefficient in the model needs to be set manually, so it is difficult to tune the parameters. Consequently, it cannot be used to flexibly predict based on the actual conditions. In addition, SEIR struggles to fit infectious diseases without obvious regularity. During the prediction of the hospital's own data, the proposed MSRD method is set to "day," and the features of multiple high correlation dimensions, including multiple infectious diseases, current environmental conditions, and transmission factors, are calculated with better nonlinear fitting ability. The MSRD model can flexibly predict real infectious diseases according to current conditions and factors at any time while obtaining better prediction results.

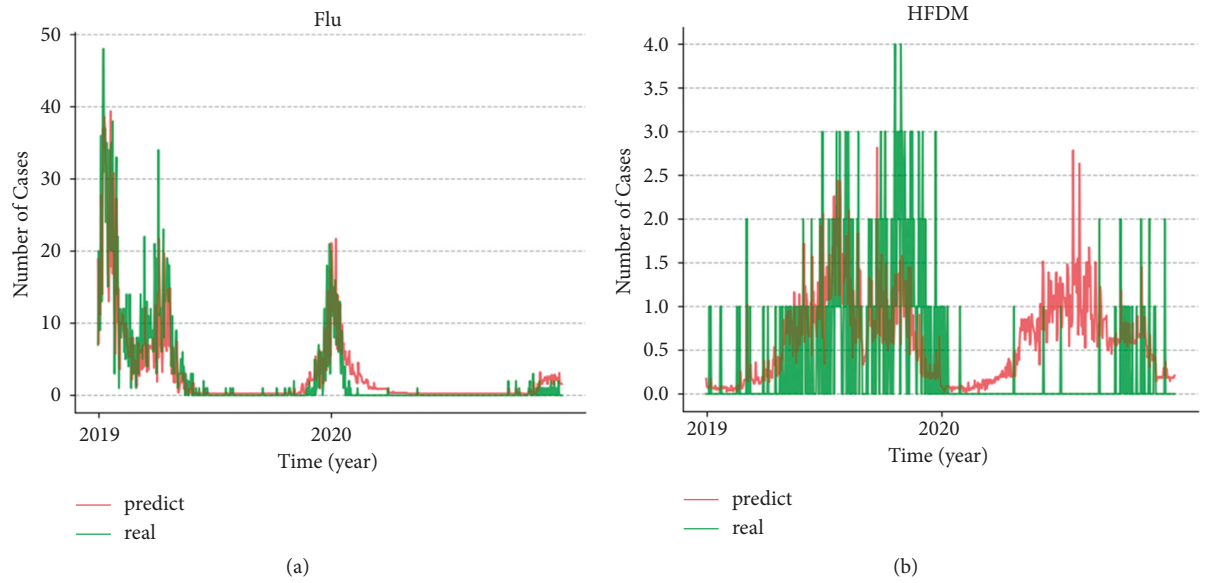

Figure 7: (a) Comparison of the real and predicted trends of influenza and (b) comparison of real and predicted trends in HFMD.

5.2. Model Performance Comparison. To verify the performance of the MSRD model, we compared the MSRD method with a variety of commonly used regression prediction models, such as support vector machine (SVM) [34], Lasso regression (lasso) [35], and Bayesian method (Bayesian) [36–38] aiming at the trend prediction of influenza and HFMD. Among them, the Bayesian regression method is less common, and this model assumes that the prior probability, likelihood function, and posterior probability are normally distributed. In this calculation, we need to maximize the marginal likelihood function to estimate the model parameters and regression coefficients. The above comparison models can be found in Python's scikit-learn library. In addition, the deep learning models that are used in the prediction of infectious disease trends are also compared: deep neural network (DNN) [11, 26], LSTM [20], bi-directional long short-term memory (Bi-LSTM), and gated recurrent unit (GRU). Among them, the MSRD model selects a sliding window length of 7, the number of LSTM neurons as 32, and the number of feedforward neural network neurons as 128. The model is trained at a learning rate of 0.001, with 223 epoch iterations being performed. The process of getting the best parameters of MSRD is provided in Supplementary Material 2. From Figure 8, it can be found that the performance of the proposed MSRD method in the two diseases is the best, making it an excellent model for practical applications. As can be seen from Table 1, the MSRD prediction results for different diseases had significant differences, and the data used in each model were identical. The reason is that the models have different learning and fitting ability to the data, and the essence is that each model has a different structure, complexity, and computational principle.

5.3. Impact of Infectious Disease Data Sources on Model Prediction. For hospitals, the probability of infectious diseases is lower than that of noninfectious diseases, and the amount of data is smaller. Therefore, this study found that determining the amount of disease data that can support model training requires attention to zero diagnosis days. Furthermore, the disease data set needs to satisfy the requirement that in the overall data, the number of days with zero confirmed cases is less than 40%. For example, 3 years of data consist of 1,095 days, so for any disease, if for at least $1{,}095 \times 60\% = 657$ days, the number of confirmed cases is not 0, then the disease can be selected as a model study object. The fewer days with zero confirmed cases should be better to represent the disease as more common and easy to capture its epidemic trend. This means that when a disease is more common, its epidemic trend is easier to predict. Extreme cases are similar to cholera. During 2012–2020, only one patient was confirmed to have cholera, while the other months had zero cases. Thus, the model for cholera cannot be learned and predicted. Therefore, this study used tuberculosis, viral hepatitis, syphilis, scarlet fever, other infectious diarrhea, influenza, and HFMD. Various model characteristics will directly affect the rationality and accuracy of the prediction results. Therefore, effective features should be selected flexibly as per the characteristics of different infectious diseases. For example, HFMD is an infectious disease among children and students [31]. The main group activity of teenagers is to attend school, so the feature of whether they are on holiday is added to describe social gathering activities. The main mode of transmission of tuberculosis is person-to-person respiratory transmission, so the feature of whether January and February contain Chinese New Year needs to be included in the infectious disease trend prediction consideration [32]. During

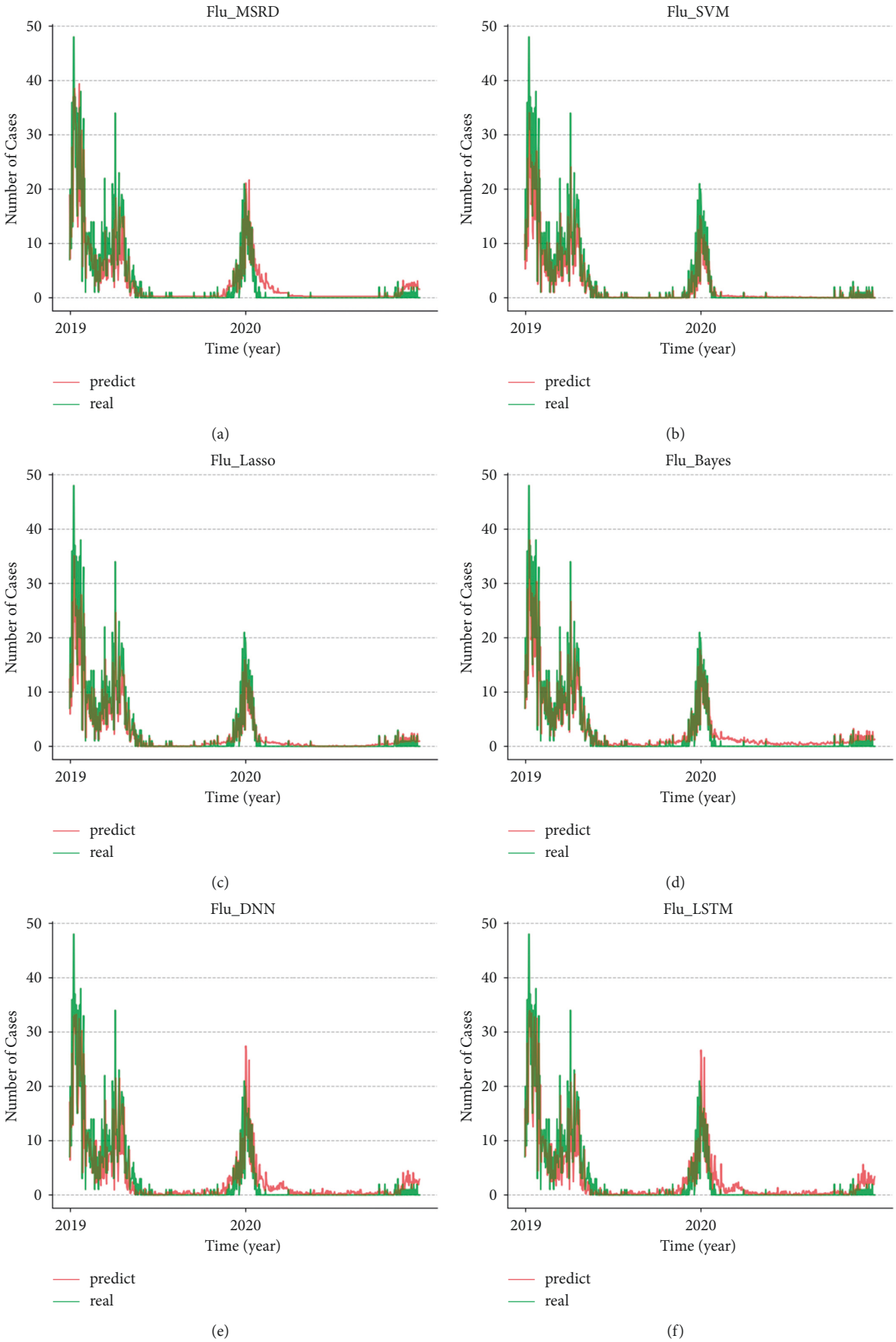

Figure 8: Continued.

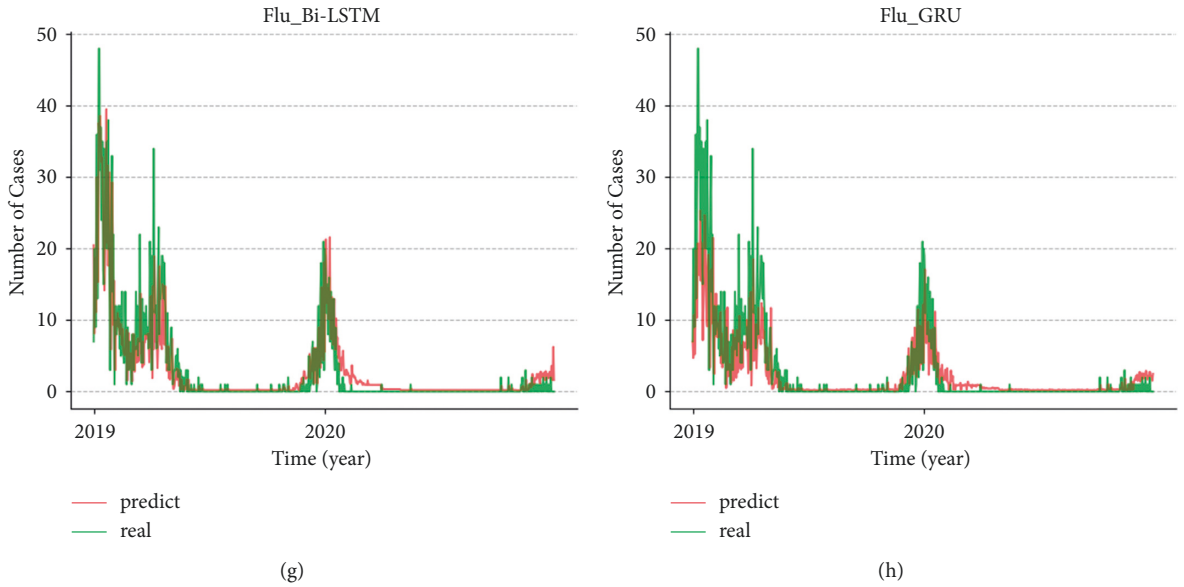

Figure 8: (a) MSRD model results, (b) SVM model results, (c) Lasso model results, (d) Bayesian model results, (e) DNN model results, (f) LSTM model results, (g) Bi-LSTM model results, and (h) GRU model results.

Table 1: Comparison of trend prediction results of infectious disease.

Comparison of trend prediction MAE results of infectious disease							
	Influenza	HFMD	Viral hepatitis	Tuberculosis	Syphilis	Scarlet fever	Infectious diarrhea
MSRD	1.6252	0.5650	7.6897	1.2754	0.4974	0.2198	0.9564
Bi-LSTM	1.8185	0.6018	8.6141	1.4270	0.5565	0.2459	1.0701
DNN	1.9585	0.6403	9.2669	1.5369	0.5894	0.2734	1.1525
LSTM	1.8859	0.6402	8.9231	1.4999	0.5771	0.2550	1.1098
GRU	1.8731	0.6075	8.8633	1.4698	0.5732	0.2533	1.1022
SVM	1.7034	0.8923	8.1596	1.3367	0.5213	0.2303	1.0024
Lasso	1.9342	0.8864	9.1515	1.5178	0.5920	0.2615	1.1382
Bayes	2.0652	0.6759	9.7713	1.6206	0.6320	0.2792	1.2152
Comparison of trend prediction RMSE results of infectious disease							
	Influenza	HFMD	Viral hepatitis	Tuberculosis	Syphilis	Scarlet fever	Infectious diarrhea
MSRD	3.8607	0.6069	9.1090	1.5705	1.7066	0.4459	1.253
Bi-LSTM	3.8588	0.8079	9.1744	1.5897	1.7177	0.4556	1.2663
DNN	3.8899	0.9209	9.3685	1.5924	1.7315	0.4492	1.2665
LSTM	3.8958	0.9170	9.2826	1.6048	1.7241	0.4499	1.2684
GRU	3.8763	0.8043	9.2361	1.5868	1.7155	0.4677	1.2620
SVM	3.8667	1.0531	9.4132	1.5829	1.8112	0.4565	1.2589
Lasso	4.5932	1.0859	10.9443	1.8903	2.0327	0.5305	1.4954
Bayes	3.9344	0.8462	9.3746	1.6106	1.7412	0.4544	1.2809

the Spring Festival, on account of the increased concentration of family gatherings, infection and transmission of tuberculosis potentially increase, and more patients tend to delay their treatment [33].

5.4. Limitations. The limitations of this study are mainly due to the small amount of data and the fact that the characteristics of the diseases were not considered in the modeling. On the one hand, this study focuses on early warning of trends in common and highly prevalent infectious diseases in hospitals. However, since some diseases such as plague, cholera, and Middle East respiratory syndrome are relatively rare (i.e., their annual incidence is less than 10 cases), their trends in individual hospitals are incidental and unpredictable. In the future, methods for predicting various types of infectious diseases can be developed based on real hospital medical record data in conjunction with the National Data Center for Public Health Sciences. Early warning and monitoring of emerging unknown infectious diseases are also worth exploring. On the other hand, it is important to note that the MSRD model in this study can calculate data characteristics such as the number of confirmed diagnoses, date, and temperature. However, the MSRD model cannot model the characteristics of the infectious disease itself or the characteristics of the infection, which still needs to be further explored.

It is also necessary to explain that the prediction of each day is based on the number of confirmed cases on the previous day and the meteorological forecast data for the next day. To predict results for Day $T+1$, we need to use the actual number of confirmed cases on Day T and the climate data of Day $T+1$. The mean value of historical temperature in the same period is used to replace the meteorological data of long-span continuous prediction. If we make a continuous prediction for the future, each prediction depends on the prediction results of the number of cases in the previous round. However, there are inevitable errors in each prediction. Thus, the continuous prediction will lead to error accumulation, and the longer the time span of continuous prediction, the larger the error accumulation and the lower the accuracy. In order to alleviate the issue of error accumulation, the historical average temperature of the same period is used in the prediction of a long-time span in the future because the actual temperature characteristics cannot be obtained, resulting in an a priori error. An alternative method with higher accuracy could be adopted to estimate the temperature. However, error accumulation occurs because the diagnosis case predicted in a certain time step needs to be used in the succeeding time step. As an alternative, the importance of the characteristics of the diagnosis number in the model construction could be reduced so that a minor fluctuation in the diagnosis number will not have a large impact on the prediction results.

6. Conclusions

In this study, we proposed an MSRD model to predict infectious disease trends in hospitals. Experimental results show that the proposed approach outperforms the SEIR model. The shortcomings of the SEIR model in hospital infection prediction were also elucidated. We also compare several neural network methods, such as DNN, LSTM, GRU, Bi-LSTM, and machine learning methods, such as SVM, Lasso regression, and Bayesian, and demonstrate that the MSRD method outperforms the above approaches. MSRD extracts the features of training data through a time window, avoiding the overfitting problem caused by long time series and the practical application of nonflexibility. In addition, the fitting ability of the model is improved by combining the output of each time step with the corresponding original input. The impact of common infectious diseases predicted by the model is consistent with the actual high prevalence of infectious diseases. The model combines hospital data with data from external data sources using a combination of medical record information, climate, and crowd gathering to provide information support for rapid response and decision-making and to assist hospitals in early warning and prediction of infectious diseases. The model can be extended to all types of hospitals for infectious disease surveillance, helping to advance infectious disease surveillance and prediction, promoting the standardization of infectious disease management, and contributing to dynamic early warning of infectious diseases.

Abbreviations

MSRD: Multi self-regression deep
MAE: Mean absolute error
SEIR: Susceptible-exposed-infected-removed
NIDIMS: National infectious diseases information monitoring system
SEIRD: Susceptible-exposed-infected-recovered-die-hard-infected
ARIMA: Autoregressive integrated moving average
LSTM: Long short-term memory
CDC: National center for disease control and prevention
HFMD: Hand-foot-and-mouth disease
SVM: Support vector machine
DNN: Deep neural network
Bi-LSTM: Bi-directional long short-term memory
GRU: Gated recurrent unit.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethical Approval

The study was approved by the Medical Science Research Ethics Committee of Peking University Third Hospital (serial no. IRB00006761-M2020318). Informed consent from the patients was exempt due to the retrospective nature of the study. The ethics committee (Peking University Third Hospital Medical Science Research Ethics Committee) approved the study abandoned informed consent. We confirm that all experiment protocols involving humans were in accordance with the guidelines of the national/international/institutional or Declaration of Helsinki in the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Cheng Yang and Yingyun Yang initiated the research and designed the experiments. Mengying Wang analyzed the data. Cuixia Lee and Wei Wang contributed to the data collection. Mengying Wang wrote the paper with the help of Cheng Yang and Yingyun Yang. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank the information management and big data center of Peking University Third Hospital for its data support for this project, Capital's Funds for Health Improvement and Research, CFH, 2021-1G-4091.

Supplementary Materials

Supplementary Material 1: Supplementary Figure 1: MSRD model predicts the epidemic situation of influenza in 2021; the green line is real data, and the red line is predicted data. Supplementary Figure 2: MSRD model predicts the epidemic situation of HFMD in 2021; the green line is real data, and the red line is predicted data. Supplementary Figure 3: MSRD model predicts the epidemic situation of viral hepatitis in 2021; the green line is real data, and the red line is predicted data. Supplementary Figure 4: MSRD model predicts the epidemic situation of tuberculosis in 2021; the green line is real data, and the red line is predicted data. Supplementary Figure 5: MSRD model predicts the epidemic situation of syphilis in 2021; the green line is real data, and the red line is predicted data. Supplementary Figure 6: MSRD model predicts the epidemic situation of scarlet fever in 2021; the green line is real data, and the red line is predicted data. Supplementary Material 2: Supplementary Figure 7: Process of getting the best parameters of MSRD. (*Supplementary Materials*)

References

[1] Standing Committee of the National People's Congress Prc, *Law on Prevention and Treatment of Infectious Diseases of 28 August 1998 (As Amended up to 2004)*International Labour Organization, Geneva, Switzerland, 2004.

[2] R. Wang, Y. Jiang, X. Guo, Y. Wu, and G. Zhao, "Influence of infectious disease seasonality on the performance of the outbreak detection algorithm in the China Infectious Disease Automated-alert and Response System," *Journal of International Medical Research*, vol. 46, no. 1, pp. 98–106, 2018.

[3] L. Zhang, W. Zhao, B. Sun, Y. Huang, and W. Glänzel, "How scientific research reacts to international public health emergencies: a global analysis of response patterns," *Scientometrics*, vol. 124, no. 1, pp. 747–773, 2020.

[4] A. Pan and T. Wu, "Wuhan COVID-19 data - an example to show the importance of public health interventions to fight against the pandemic," *Toxicology*, vol. 441, Article ID 152523, 2020.

[5] H. Dette, V. B. Melas, and A. Pepelyshev, "Optimal designs for estimating individual coefficients in polynomial regression - a functional approach," *Journal of Statistical Planning and Inference*, vol. 118, no. 1-2, pp. 201–219, 2004.

[6] R. Ghostine, M. Gharamti, S. Hassrouny, and I. Hoteit, "An extended SEIR model with vaccination for forecasting the COVID-19 pandemic in Saudi arabia using an ensemble kalman filter," *Mathematics*, vol. 9, no. 6, p. 636, 2021.

[7] J. Wangping, H. Ke, Y. Shanshan 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.

[8] Y. Y. Wei, Z. Z. Lu, Z. C. Du et al., "Fitting and forecasting the trend of COVID-19 by SEIR(+CAQ) dynamic model," *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi*, vol. 41, no. 4, pp. 470–475, 2020.

[9] Z. Yang, Z. Zeng, K. 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.

[10] H. M. Youssef, N. A Alghamdi, M. A Ezzat, A. A El-Bary, and A. M Shawky, "A new dynamical modeling SEIR with global analysis applied to the real data of spreading COVID-19 in Saudi Arabia," *Mathematical Biosciences and Engineering*, vol. 17, no. 6, pp. 7018–7044, 2020.

[11] S. Feng, Z. Feng, C. Ling, C. Chang, and Z. Feng, "Prediction of the COVID-19 epidemic trends based on SEIR and AI models," *PLoS One*, vol. 16, no. 1, p. e0245101, 2021.

[12] Q. Liu, Z. Li, Y. Ji et al., "Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu Province of China using advanced statistical time-series analyses," *Infection and Drug Resistance*, vol. 12, pp. 2311–2322, 2019.

[13] W. Wu, S.-Y. An, P. Guan, D.-S. Huang, and B.-S. Zhou, "Time series analysis of human brucellosis in mainland China by using Elman and Jordan recurrent neural networks," *BMC Infectious Diseases*, vol. 19, no. 1, p. 414, 2019.

[14] N. Shachar, A. Mitelpunkt, T. Kozlovski et al., "The importance of nonlinear transformations use in medical data analysis," *Jmir Medical Informatics*, vol. 6, no. 2, pp. e27–79, 2018.

[15] A. Singhal, P. Singh, B. Lall, and S. D. Joshi, "Modeling and prediction of COVID-19 pandemic using Gaussian mixture model," *Chaos, Solitons & Fractals*, vol. 138, p. 110023, 2020.

[16] R. E. Watkins, S. Eagleson, B. Veenendaal, G. Wright, and A. J. Plant, "Disease surveillance using a hidden Markov model," *BMC Medical Informatics and Decision Making*, vol. 9, no. 1, pp. 1–12, 2009.

[17] X. Guo, S. Liu, L. Wu, and L. Tang, "Application of a novel grey self-memory coupling model to forecast the incidence rates of two notifiable diseases in China: dysentery and gonorrhea," *PLoS One*, vol. 9, no. 12, Article ID e115664, 2014.

[18] M. Erraguntla, J. Zapletal, and M. Lawley, "Framework for Infectious Disease Analysis: a comprehensive and integrative multi-modeling approach to disease prediction and management," *Health Informatics Journal*, vol. 25, no. 4, pp. 1170–1187, 2019.

[19] R. Telarolli, L. C. M. Loffredo, and R. M. Gasparetto, "Clinical and epidemiological profile of tuberculosis in an urban area with high human development index in southeastern Brazil. Time series study," *Sao Paulo Medical Journal*, vol. 135, no. 5, pp. 413–419, 2017.

[20] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," *International Journal of Environmental Research and Public Health*, vol. 15, no. 8, p. 1596, 2018.

[21] H. T. Rauf, J. Gao, A. Almadhor, M. Arif, and M. T. Nafis, "Enhanced bat algorithm for COVID-19 short-term forecasting using optimized LSTM," *Soft Computing*, vol. 25, no. 20, pp. 12989–12999, 2021.

[22] Y. Guo, Y. Feng, F. Qu, L. Zhang, B. Yan, and J. Lv, "Prediction of hepatitis E using machine learning models," *PLoS One*, vol. 15, no. 9, Article ID e0237750, 2020.

[23] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals*, vol. 135, Article ID 109864, 2020.

[24] H. Mohammad, M. Elham, E. Mehraeen et al., "Identifying data elements and key features of a mobile-based self-care application for patients with COVID-19 in Iran," *Health Informatics Journal*, vol. 27, no. 4, Article ID 146045822110657, 2021.

[25] J. Gu, L. Liang, H. Song et al., "A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China," *Scientific Reports*, vol. 9, no. 1, Article ID 17928, 2019.

[26] Z. Liao, P. Lan, X. Fan, B. Kelly, A. Innes, and Z. Liao, "Sirvd dl: a COVID-19 deep learning prediction model based on time-dependent SIRVD," *Computers in Biology and Medicine*, vol. 138, Article ID 104868, 2021.

[27] P. Bedi, S. Dhiman, P. Gole, N. Gupta, and V. Jindal, "Prediction of COVID-19 trend in India and its four worst-affected states using modified SEIRD and LSTM models," *SN computer science*, vol. 2, no. 3, p. 224, 2021.

[28] W. Li, H. J. C. Ji, and H. I. Jo, "Management: exploration and practice of hospita's data utilization based on the hadoop," *Chinese Journal of Health Informatics and Management*, 2016.

[29] C.-Y. Yang, R.-J. Chen, W.-L. Chou, Y.-J. Lee, and Y.-S. Lo, "An integrated influenza surveillance framework based on national influenza-like illness incidence and multiple hospital electronic medical records for early prediction of influenza epidemics: design and evaluation," *Journal of Medical Internet Research*, vol. 21, no. 2, p. e12341, 2019.

[30] F. Yavari Nejad, K. D. J. B. M. I. Varathan, and D. Making, "Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 141, 2021.

[31] E. Y. Sapia, C. Maroni, C. Groisman et al., "Atypical hand-foot-mouth disease virus genotyping in a pediatric hospital in Buenos Aires city, Argentina," *Archivos Argentinos de Pediatría*, vol. 118, no. 2, pp. E199–E203, 2020.

[32] K. Liu, T. Li, A. Vongpradith et al., "Identification and prediction of tuberculosis in eastern China: analyses from 10-year population-based notification data in zhejiang province, China," *Scientific Reports*, vol. 10, no. 1, p. 7425, 2020.

[33] S. Cao, F. Wang, W. Tam et al., "A hybrid seasonal prediction model for tuberculosis incidence in China," *BMC Medical Informatics and Decision Making*, vol. 13, no. 1, p. 56, 2013.

[34] H. Bagheri, L. Tapak, M. Karami et al., "Forecasting the monthly incidence rate of brucellosis in west of Iran using time series and data mining from 2010 to 2019," *PLoS One*, vol. 15, no. 5, p. e0232910, 2020.

[35] P. P. Schneider, C. J. A. W. van Gool, P. Spreeuwenberg et al., "Using web search queries to monitor influenzalike illness: an exploratory retrospective analysis, Netherlands, 2017/18 influenza season," *Euro Surveillance*, vol. 25, no. 21, pp. 13–22, 2020.

[36] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.

[37] J. Zou, A. F. Karr, G. Datta, J. Lynch, S. J. B. M. I. Grannis, and D. Making, "A Bayesian spatio-temporal approach for real-time detection of disease outbreaks: a case study," *BMC Medical Informatics and Decision Making*, vol. 14, no. 1, pp. 108–118, 2014.

[38] A. Lizasoain, D. Mir, N. Martinez, and R. Colina, "Coxsackievirus A10 causing hand-foot-and-mouth disease in Uruguay," *International Journal of Infectious Diseases*, vol. 94, pp. 1–3, 2020.