Research on Forecast of Global COVID-19 Development Trend

Yunyun Cheng

North University of China

Yanping Bai (✉ baipy666@163.com)

North University of China

Ting Xu

North University of China

Rong Cheng

North University of China

Hongping Hu

North University of China

Research Article

Keywords: COVID-19, Machine learning, Time series prediction, Neural network

Posted Date: October 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-948058/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Research on forecast of global COVID-19 development trend

Yunyun Cheng¹, Yanping Bai^{2*}, Ting Xu², Rong Cheng², Hongping Hu²

¹ Department of Mathematics, School of Science, North University of China, Taiyuan 030051, China.

² Department of Mathematics, School of Science, North University of China, Taiyuan 030051, China.

* Corresponding author

E-mail: baipy666@163.com

Abstract

The rapid spread of Corona Virus Disease 2019 (COVID-2019) has seriously threatened people's health and brought huge challenges to the medical systems of many countries. So it is necessary to predict the epidemic trend scientifically and accurately. In this paper, the development trend prediction of the global daily cumulative confirmed case, based on data from February 2, 2020 to August 14, 2021 using linear regression and nonlinear regression methods, which are auto regressive integrated moving average (ARIMA), wavelet neural network (WNN), support vector machine (SVM), recurrent neural network (RNN) and long short-term memory (LSTM). Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) were used to evaluate the prediction accuracy of the five models. The experimental results showed that the LSTM used in this study is more accurate for predicting the development trend of global cumulative confirmed cases. And its three evaluation indexes were 0.0936%, 0.1123%, and 0.0962% respectively, which were the smallest compared with the other four models. The LSTM model was used to predict the global cumulative confirmed cases in the next five days. The prediction results showed that the global cumulative confirmed cases will rise steadily and exceed 205 million. Therefore, the health departments of various countries should take appropriate prevention and control measures in advance.

Keywords: COVID-19, Machine learning, Time series prediction, Neural network

1. Introduction

COVID-19 has now become an epidemic infectious disease in the world, which is distributed in more than 220 countries and regions. As of August 25, 2021, the cumulative number of confirmed cases in the world has exceeded 200 million and the cumulative number of deaths has exceeded 4.6 million, which has had a great impact on the economy, people's lives and property of various countries, such as the delay of the Olympics and the European Cup, the deserted cities, and the degree of economic recession that exceeded the 2008 international financial crisis. Currently, the epidemic situation in some countries has been controlled.

However, due to the lack of monitoring and protective equipment, many countries still have a long time to go on the road of epidemic control. Therefore, understanding the development characteristics and laws of COVID-19 will help governments of various countries to improve the epidemic prevention and control system.

At present, the research on COVID-19 development trend prediction is mainly composed of dynamic model based on mechanism analysis and machine learning model based on time series data. Taking into account the characteristics of COVID-19 such as latency and mortality, the study of COVID-19 dynamic model takes SEIR as the main framework. Khan, M. A. et al. constructed the coupled SEIR dynamic model of bat and human, and gave the basic regeneration number and the numerical solution of the model [1]. Youssef, H. M. et al. established a SEIR dynamic model based on real data on the spread of COVID-19 in Saudi Arabia through copy number and extensive stability analysis [2]. He, S. B., et al. used particle swarm optimization algorithm (PSO) to estimate system parameters, and established a SEIR epidemic model of COVID-19 based on data from Hubei Province [3].

Considering that the spread of COVID-19 may be affected by factors such as season and location, Kissler, S. M. et al. referred to the seasonality, immunity, and cross-immunity characteristics of some coronaviruses similar to COVID-19, and estimated the parameters of the COVID-19 transmission model, and established an epidemic prediction model with the United States as the research area [4]. Li, R. Y. et al. used Tencent location request data in mainland China to establish a SEIR model based on a population movement network, and calculated that in the weeks before and after the “closing of the city” in Wuhan, China did not record the infectivity and number of infected people, explaining the COVID-19 reasons for rapid spread in China [5].

Some scholars have studied the development trend of the COVID-19 epidemic from the perspective of statistical modeling, such as machine learning [6-8], linear models [9-11], and exponential growth models [12, 13]. Oliveira, R. F. A. P. et al. established a classification model for the test results of suspected covid-19 patients and a classification model for the inpatients of patients with COVID-19 [14]. Abdulkareem, A. B. et al. developed three machine learning models (CNN, ADtree Classifier and Bayesnet) based on the important weather features selected by PCA [15]. El-dosuky, M. A. et al. presented a cockroach optimized deep neural network to detect COVID-19 and differentiate between COVID-19 and influenza types A, B, and C [16]. Tiwari, D. et al. proposed a meta-analysis based on artificial intelligence (AI) to predict the global COVID-19 epidemic trend [17]. Rustam, F. et al. established four standard prediction models (LR, least absolute shrinkage, LASSO, SVM and ES) to forecast the threatening factors of COVID-19 [18].

In the face of sudden public health events, how to use megadata to improve government governance capacity and governance efficiency is a new topic. Based on the above background, we explored new ideas for predicting the development trend of COVID-19. The main research process is as follows:

- (i) We combined the characteristics of the spread of COVID-19 and regarded the development trend of the epidemic as a long-term and dynamically changing time series forecasting problem.
- (ii) With the help of the superiority of artificial intelligence methods in the time series forecasting problem, we established linear regression model and nonlinear regression

model, including ARIMA, WNN, SVM, RNN and LSTM.

- (iii) The performance of the five methods was compared according to the MAE, RMSE, and MAPE indicators, and the LSTM model was found to be the best predictive model.
- (iv) LSTM model was used to predict the number of global cumulative confirmed cases in the next 5 days.

The main contributions of this paper are as follows: The development of COVID-19 is affected by many factors, and the relationship between its input variables and output variables is unknown, while linear and nonlinear can describe the different interactions between variables. Therefore, based on the previous multi-source data, this study proposed linear regression and nonlinear regression methods in machine learning to realize the COVID-19 trend prediction, analyzed the future epidemic development trend, and provided new methods and basic technical support for COVID-19 prevention and control.

2. Related work

2.1 COVID-19 data set source and analysis

The data set used in this study came from the real-time dynamics of the COVID-19 epidemic site searched by sogou, focusing on the global daily cumulative confirmed cases. We collected 560 sets of monitoring data and formed them into a time series. The trend of global cumulative confirmed cases is shown in Fig. 1. It can be seen that the global epidemic data shows a continuous growth trend, reflecting the increasingly serious epidemic problem. The previous historical statistics show that the longest incubation period of COVID-19 is 14 days, and the iteration interval is generally 5 to 7 days. Serial interval refers to the length of time for a new generation of cases, and there is a time difference of 1 to 2 days from the onset of symptoms to treatment. Therefore, one week is set as a prediction period in this study.

Fig. 1. Cumulative number of confirmed cases worldwide from February 2, 2020 to August 14, 2021. The cumulative number of confirmed cases has increased steadily, and more than 190 million people have been confirmed as COVID-19.

2.2 Time series prediction

In today's big data era, computer technology has developed rapidly. The research of machine learning combined with big data has played a more and more important role in the development and competition of various fields [19-21]. Applying machine learning to infectious disease early warning system can break through the limitations of traditional prediction system and achieve the purpose of real-time and dynamic early warning, thereby improving the effect of prevention and control [22, 23]. Therefore, machine learning is used for global COVID-19 epidemic trend prediction, which is a direction that can be studied and applied.

The transmission mechanism of COVID-19 has complex characteristics. At present, its trend prediction is based on the existing research results of infectious disease prediction, and improvements are made to meet current needs. In this paper, the COVID-19 forecasting problem is transformed into a time series forecasting problem for research. A time series is a set of random variables sorted by time. It is usually the result of observing a certain potential process at a given sampling rate in an equally spaced time period. At present, time series prediction has been widely used in the medical and health fields, such as clinical prediction [24-26], blood glucose concentration prediction [27-29], and health status prediction [30-32]. The time series forecast in this study is shown in Fig. 2. First, a time series model is constructed from historical data, and then the model is used to extrapolate to predict the future.

Fig. 2. Time series forecasting process. It can be described as predicting the first series value x_{n+k} in the future based on the continuous series values $x_k, x_{k+1}, x_{k+2}, \dots, x_{n+k-1}$ in the past.

2.3 Evaluation index

Appropriate evaluation indexes are used to compare the accuracy of the prediction methods in this paper. Three evaluation indexes, including MAE, RMSE and MAPE, are used to compare the prediction performance of COVID-19 prediction model in this study. However, there is no one evaluation index that can be applied to all situations. Considering the large value of the research data used in this paper, we redefined MAE and RMSE to solve their shortcomings of lacking consideration of the actual value range. Eqs. (1)-(3) are the calculation formulas of three evaluation indexes respectively, where n is the number of test sets, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

$$MAE(\%) = \frac{\frac{1}{n} \sum_{i=1}^n |y(i) - \hat{y}(i)|}{\frac{1}{n} \sum_{i=1}^n y(i)} = \frac{\sum_{i=1}^n |y(i) - \hat{y}(i)|}{\sum_{i=1}^n y(i)} \quad (1)$$

$$MAPE(\%) = \sum_{i=1}^n \left| \frac{y(i) - \hat{y}(i)}{y(i)} \right| \times \frac{100}{n} \quad (2)$$

$$RMSE(\%) = \frac{\sqrt{\frac{\sum_{i=1}^n (y(i) - \hat{y}(i))^2}{n}}}{\frac{\sum_{i=1}^n y(i)}{n}} \quad (3)$$

3. Linear regression prediction of time series

Linear regression prediction is a signal modeling method, which regards the current value of the signal as a linear combination of the sampled values at several previous times. It is mainly used in the fields of speech information coding, spectrum analysis and so on. The basic task of linear prediction is to predict the estimated value $\hat{x}_{n+1}$ of the current signal according to the known previous n observations. ARIMA model has the best linear filtering characteristics. When the signal and noise are input at the same time, it can extract the signal as accurately as possible at the output, and the noise is suppressed to the greatest extent.

3.1 Introduction to ARIMA

ARIMA is a classical time series model, which is often used to quantitatively describe the correlation between variables in time series, and is widely used in time series prediction research in various fields. ARIMA was selected as the linear prediction model for the COVID-19 trend, which was also compared with the nonlinear regression methods in this study.

Assuming that the target sequence is a stationary sequence after d times of difference processing, the expression of the ARIMA (p, d, q) model can be written in the form of Eq. (4):

$$\Delta x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \dots + \theta_q u_{t-q} + c \quad (4)$$

Where u_t is white noise sequence, c is a constant, ϕ_i is the autoregressive coefficient, θ_q is the moving average coefficient.

Before training ARIMA model, first test whether the sequence meets the requirements of stability. If not, make the sequence stationary through difference calculation. Next, the Bayesian information criterion (BIC) is used to determine the values of p and q. Eq. (5) is the calculation formula of BIC, where N is the number of parameters. The minimum BIC corresponds to p and q, which is the best ARIMA model. Then the residual of the model is tested. If the residual is a white noise sequence, the model is available.

$$BIC(p, q) = N \ln \sigma^2(p, q) + (p + q + 1) \ln N \quad (5)$$

3.2 ARIMA prediction

In this section, 560 sets of data are divided into two parts, 509 sets are training sets, and the remaining 51 sets are test sets. From the time series diagram, ACF and PACF diagram of the original data drawn by MATLAB, it can be judged that the time series sequence is a non-stationary sequence. The original data sequence and ACF and PACF are shown in Fig. 3 (a) and Fig. 3 (b). After two difference calculations, the data basically fluctuates around a fixed value. The differential data sequence and ACF and PACF are shown Fig. 3 (c) and Fig. 3 (d). At this time, the sequence becomes stable and can be initially applied to the ARIMA model.

In order to ensure that the order selection is more accurate and appropriate, the automatic recognition model function is used to respectively fit the various situations that can occur in ARIMA, and the following BIC values are obtained: 1.3987, 1.3846, 1.4065, 1.4069, 1.3966, 1.3867, 1.4070, 1.4073, 1.4048, 1.3870, 1.4071, 1.3858, 1.3851, 1.3850, 1.3873, 1.3832. When $p=4$, $q=4$, the value of BIC is the smallest, which is 1.3846. Therefore, ARIMA (4, 2, 4) is the optimal model.

Fig. 3. Comparison of the original data and the data after the difference calculation. (a) The development trend of raw data. (b) ACF and PACF diagrams of the original data. (c) The trend chart after two differential calculations on the original data. (d) After two difference calculations, the ACF and PACF diagrams of the sequence. The original data shows an exponential upward trend, and ACF and PACF have no tailing and truncation trend. It can be judged that the original data is a non-stationary series. After two differential calculations, the sequence trend fluctuates up and down at a certain value, and ACF and PACF are obviously truncated and tailed after order 2 and 7 respectively. It can be judged that the sequence is stable.

The test results of the ARIMA (4, 2, 4) model are shown in Fig. 4 (a-d). The standard residual plot and the trend plot of QQ PLOT show that the prediction residuals of the ARIMA (4, 2, 4) model obey normal distribution. The residual sequence is a white noise sequence, that is, almost all useful information in the sequence is extracted by this model. The ARIMA (4, 2, 4) model

can be used as a COVID-19 epidemic trend prediction model. The prediction result is shown in Fig. 4 (e).

Fig. 4. ARIMA model test and prediction results. (a) Standardized Residuals. (b) QQ Plot. (c) ACF plot of residual sequence. (d) PACF plot of residual sequence. (e) Predicted result.

4 Nonlinear regression prediction of time series

Various types of neural networks have become one of the main tools of nonlinear time series prediction because of their universal nonlinear approximation ability. In order to develop more COVID-19 prediction systems, four network models are used:

- Wavelet neural network
- Support Vector Machine
- Recurrent Neural Network
- Long Short-Term Memory Neural Network

4.1 Introduction to theory

A. Wavelet neural network

WNN (Wavelet neural network) is based on the topological structure of the BP neural network, and the wavelet basis function is used as the transfer function of the hidden layer nodes. The information transmission mode of wavelet neural network is signal forward

propagation and error back propagation.

The calculation formula of the hidden layer is as follows:

$$h(j) = h_j \left[\frac{\sum_{i=1}^k \omega_{ij} x_i - b_j}{a_j} \right], j = 1, 2, L, l \quad (6)$$

where $h(j)$ is the output of the hidden layer, b_j is the translation factor, a_j is the expansion factor. h_j is wavelet basis function. The wavelet basis function used in this paper is Morlet wavelet basis function, and the mathematical formula is Eq. (7):

$$y = \cos(1.75x) e^{\frac{-x^2}{2}} \quad (7)$$

B. Support Vector Machine

SVM (Support Vector Machine), like multilayer perceptron network and radial basis function, can be used for classification and regression prediction. The basic principle of support vector regression is to train all data samples so that they are distributed between two straight lines, and the total deviation from all points to the straight line is the smallest. Then find the maximum distance between the two lines at this time, which is the optimal hyperplane for support vector regression. It has the advantages of wide versatility, strong robustness, and simple calculation. Non-linear problems can be transformed into linear problems in high-dimensional space through the kernel function K of SVM.

Where K is kernel function. The kernel function used in this article is the Gaussian radial basis kernel Function, Eq. (8) is its calculation formula.

$$K(x_i, x_j) = \exp \left[-\frac{\|x_i - x_j\|^2}{\sigma^2} \right] \quad (8)$$

C. Recurrent Neural Network

RNN (Recurrent Neural Network) is a neural network that models sequence data, that is, the current output of a sequence is also related to the previous output. Compared with other traditional neural networks, RNN will memorize the previous information and apply it to the calculation of the current output. The structure of RNN is shown in Fig. 5. RNN introduces the concept of “memory”, and the input of the hidden layer includes not only the output of the input layer but also the output of the previous hidden layer.

Fig. 5. RNN network structure diagram. This is the structure of RNN unfolded in time series, and the entire neural network shares a set of parameters U , V , W .

Here, x_t is the input. h_t is the hidden layer state, which takes the network input of the current time and the hidden layer state of the previous time as the input of the hidden layer. y_t is the output. Eq. (9) is the calculation formula of hidden layer output. $f(*)$ is the activation function of the hidden layer, and $g(*)$ is the output layer activation function.

$$h_t = f(Wx_t + Uh_{t-1}) \quad (9)$$

$$y_t = g(Vh_t) \quad (10)$$

D. Long Short-Term Memory Neural Network

LSTM (Long Short-Term Memory Neural Network) is a special type of RNN that can learn long-term dependent information. Unlike RNN, LSTM adds a memory unit specifically to memorize the results of the previous round. LSTM has a complex network, which is composed of a sequence of cyclically connected memory units, each of which has one or more self-connected neurons and three gates that control the propagation of various types of information. Its network structure mainly includes forget gate, input gate and output gate.

The forget gate decides what information to discard and what information to keep from the state at the previous moment, which takes the input at the current moment and the output at the previous moment as input. There is a sigmoid layer in the structure of the forget gate, which outputs a vector between 0 to 1. Here 1 means all information is retained, 0 means all information is discarded.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (11)$$

The input gate decides how much new information is stored in the unit state, including the sigmoid layer and the tanh layer. The former decides which information is updated, the latter outputs a vector that can be added to the cell state.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (12)$$

$$C_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (13)$$

Then the cell state at the previous moment C_{t-1} is updated to the cell state at the current moment C_t .

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (14)$$

The output gate decides the output vector of the cell state h_t , which is calculated according to the cell state C_t . Some information in C_t needs to be filtered through the sigmoid layer.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (15)$$

$$h_t = o_t * \tanh(C_t) \quad (16)$$

Where h_t and x_t are input vector, f_t is output vector, W_f , W_i , W_c , W_o are weight vectors. b_f , b_i , b_c , b_o are bias, f_t , i_t and o_t are the control states of the forget gate, input gate and output gate, respectively.

4.2 Experimental predictions analysis

The ultimate goal of the experiment is to apply the trained model to the real environment, hoping that the trained model can get good prediction effect on the real data. COVID-19 data set is divided into training set and test set in this study, where 502 groups are training sets and 51 groups are test sets. The forecasting process in this paper is shown in the Fig. 6. The data of the training set were used to train the model, and then the trained model was applied to the test set.

Fig. 6. Forecast flow chart. The prediction process mainly includes three parts: data preprocessing, network training and prediction and model evaluation.

A. Wavelet neural network prediction

The cumulative confirmed cases at a certain moment are related to the cumulative confirmed cases in the previous time periods. The COVID-19 iteration interval is generally 7 days. According to the propagation characteristics of COVID-19, the wavelet neural network with structure of 7-6-1 is used in this study. The number of nodes in the output layer is 1, that is, the

cumulative number of confirmed cases at a certain moment. The number of nodes in the input layer is 7, that is, the cumulative number of confirmed cases at the first seven time points. The hidden layer node is 6, which is composed of wavelet function.

WNN is iterated 100 times with 502 groups of COVID-19 training data. Then, 51 groups of COVID-19 test data were substituted into the trained wavelet neural network to predict the cumulative confirmed cases. The prediction results of wavelet neural network are shown in Fig. 9 (a).

B. Support Vector Machine prediction

The regression model established by SVM is used to fit the cumulative confirmed cases. The data set is a 553×8 double matrix, which is preprocessed. It is necessary to adjust the penalty parameter (c) and the kernel function parameter (g) in SVM for regression prediction. Then select the best parameters through cross-validation in this paper. Compared with randomly selected parameters, the parameters obtained in this way are more effective in predicting. The best parameters obtained above are used to train and test SVM.

In this paper, Cross-validation is used to optimize SVM parameters. The parameter selection results are shown in Fig. 7 (a) and Fig. 7 (b), where best $c=0.57435$, $g=1$, $CVmse=0.0023782$.

Fig. 7. SVM parameter selection process. (a) Parameter selection 2D process. (b) Parameter selection 3D process The X-axis is the value of $\log_2 c$, and the Y-axis is the value of $\log_2 g$. The contour line is the accuracy rate of the corresponding cross-validation method after taking the corresponding parameters c and g .

The prediction results of SVM are shown in Fig. 9 (b), the SVM prediction curve fits the true value curve and maintains the same growth trend.

C. Recurrent Neural Network prediction

The number of input and output nodes of RNN is determined by the input and output dimensions of samples. So the number of input nodes of the network is 7 and the number of output nodes is 1.

The RNN regression fitting diagram is shown in Fig. 8, which can more intuitively display the prediction accuracy of the RNN model. The prediction result of RNN is shown in Fig. 9 (b). It can be seen from that the RNN neural network can better predict the trend of cumulative diagnoses, but the network prediction results still have large errors. In the next section, we will

study the LSTM neural network optimized on the basis of RNN in order to get better prediction results.

Fig. 8. The prediction effect of RNN. The closer the correlation coefficient is to 1, the better the prediction effect of the regression model. It can be seen that the correlation coefficients between the predicted value and the real value in the training set, the verification set, the test set and all set are 0.97212, 0.96374, 0.96841 and 0.9703 respectively, indicating that the RNN regression prediction effect is good.

D. Long Short-Term Memory Neural Network prediction

The working process of LSTM is mainly forward propagation, back propagation and weight update. The first step is to calculate the output value of each cell state forward, and then we use the output value and the actual value to get an error term. The next step is to reverse calculate the gradient of each weight parameter, and the obtained gradient is the adjustment direction and amplitude of the parameter. Finally, the weight of each network layer is updated. Stop updating until the error meets the expected value given.

The LSTM prediction result is shown in Fig. 9 (d). The abscissa is the number of test set groups, and the ordinate is the cumulative number of confirmed cases worldwide. It can be seen that compared with the RNN prediction curve, the LSTM prediction curve has a higher degree of fit with the actual curve.

5. Discussion

In this paper, we proposed five COVID-19 trend prediction models. Fig. 9 shows the comparison of five prediction results, where the abscissa are the days of the test set, 51 sets of data were selected as the test set in this study, and the longitudinal coordinates are the cumulative number of confirmed cases worldwide. The red curve is the development trend of real value, and the green, black, purple and yellow curves are the prediction results of WNN, SVM, RNN and LSTM respectively. It can be seen from the figure that the yellow (LSTM) and black (SVM) curves have a higher degree of fit with the red curve (actual value), and the purple

(RNN) curve has the lowest degree of fit. It can be initially judged that the LSTM and SVM models have higher prediction accuracy compared with the other three models, while the RNN model has a poor prediction effect.

Fig. 9. Prediction results. (a) WNN prediction results. (b) SVM prediction results. (c) SVM prediction results. (d) LSTM prediction results.

In order to compare the prediction performance of the model more objectively, we calculated the evaluation indicators MAE, RMSE and MAPE of the five models. Table 1 shows the evaluation index values of the five models.

Table 1

Calculation results of evaluation indexes of five prediction models.

Models	Evaluation index		
	MAE (%)	MAPE(%)	RMSE(%)
ARIMA	1.0609	1.0227	1.4779
WNN	3.4587	3.3869	4.0031
SVM	0.7168	0.7248	0.7954
RNN	5.3639	5.4243	5.5362
LSTM	0.0936	0.0962	0.1123

By comparing the evaluation indexes in Table 1, we can rank the prediction performance of the five prediction models: LSTM > SVM > ARIMA > WNN > RNN. Here, the MAE (5.3639%), MAPE (5.4243%) and RMSE (5.5362%) of RNN model are the highest. So its prediction performance is the worst compared with the other four prediction models. The structure of LSTM is optimized based on RNN, and its MAE, MAPE and RMSE are reduced to 0.0936%, 0.0962% and 0.1123%. The prediction performance of LSTM has been greatly

improved compared with RNN, it shows that the cumulative confirmed case data sequence has the characteristics of long-term dependence. And compared to the other four methods, the LSTM network structure is the most complicated. Therefore, LSTM can extract complex features of data, and its prediction performance is also optimal.

Finally, the LSTM model is used to predict the global cumulative confirmed cases from August 15, 2021 to August 20, 2021. The final prediction results are shown in Table 2. It can be seen from Table 2 that the global cumulative confirmed cases will continue to rise, and the number of cumulative confirmed cases in the future may exceed 205 million. Therefore, governments in various countries should continue to strengthen prevention and control measures, and remind people to take protective measures when going out.

Table 2

Prediction results of cumulative confirmed cases in the next five days using LSTM.

Dates	2021.08.15	2021.08.16	2021.08.17	2021.08.18	2021.08.19
Predictive value	204075975	204486628	204893902	2053070755	2057312836

6. Conclusion

Focusing on the global cumulative confirmed cases, we explored new ideas for the development trend time series prediction of COVID-19. The historical and future cumulative confirmed cases were analyzed from two aspects: the linear relationship in the time series and the nonlinear relationship in the time series. ARIMA was selected as the linear prediction model, and WNN, SVM, RNN and LSTM were selected as nonlinear prediction models. A variety of methods are used to establish the prediction model, which can more accurately and comprehensively extract the characteristic laws in the data. Then, according to the deviation and the evaluation indicators between the fitted values of the five different models and the historical true values, the model with the best prediction performance was obtained. With the continuous improvement of nonlinear system prediction and identification system, the prediction accuracy has also been improved. Finally, the optimal model LSTM was used to predict the global cumulative confirmed cases in the next 5 days. The results show that the predicted trend is in line with the actual global cumulative confirmed case change range. Due to the limited time, we only analyzed and predicted the development trend of COVID-19 from the perspective of time series, and lacked the application of spatial information. In the next research, we will explore the epidemic prediction model from the perspective of region and environment, so as to provide more scientific and effective reference for relevant epidemic prevention and control departments.

Acknowledgements

We would like to thank our colleagues in the Modern Optimization Algorithm Laboratory of North University of China for their help in data collection and analysis.

Author contributions

Data curation: T. X, R. C;

Analysis and interpretation of data: Y. C, H. H;

Methodology: Y. C;

Software: Y. C, Y. B;

Supervision: T. X, H. H;
Writing – original draft: Y. C;
Writing – review & editing: Y. C, Y. B.

Funding

Research project supported by National Nature Science Foundation of China (Grant No. 61774137), Shanxi Natural Science Foundation (Grant No. 201801D121026 and 201701D221121), Shanxi Scholarship Council of China (Grant No.2020-104 and 2016-088), the Key Research and Development Projects of Shanxi Province (Grant No.201903D121156) and Graduate Innovation Research Project of School of Science of North University of China (No.20210813).

Competing interests

The authors have declared that no competing interests exist

Data availability

The data sets used in this study are freely available to public on the webpage: https://sa.sogou.com/new-weball/page/sgs/epidemic?type_page=VR , and are provided in the Supplementary Dataset.

References

- [1] Khan, M. A. & Atangana, A. Modeling the dynamics of novel coronavirus (2019-nCov) with fractional derivative. *Alex. Eng. J.* **4**, 2379-2389 (2020).
- [2] Youssef, H. M., Alghamdi, N. A., Ezzat, M. A., El-Bary, A. A. & Shawky, A. M. A new dynamical modeling SEIR with global analysis applied to the real data of spreading COVID-19 in Saudi Arabia. *Math. Biosci. Eng.* **17**, 7018-7044(2020).
- [3] He, S. B., Peng, Y. X., Ezzat, M. A. & Sun, K. H. SEIR modeling of the COVID-19 and its dynamics. *Nonlinear. Dynam.* **101**, 1667-1680(2020).
- [4] Kissler, S. M., Tedijanto, C., Goldstein, E. Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*. **368**, 860-868 (2020).
- [5] Li, R. Y. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*. **368**,489-493 (2020).
- [6] Kwekha-Rashid, A. S., Abduljabbar, H. N. & Alhayani, B. Coronavirus disease (COVID-19) cases analysis using machine-learning applications. *Appl. Nanosci.* <https://dx.doi.org/10.1007%2Fs13204-021-01868-7>
- [7] Zimmerman, A. & Kalra, D. Usefulness of machine learning in COVID-19 for the detection and prognosis of cardiovascular complications. *Rev. Cardiovasc. Med.* **21**, 345-352(2020).
- [8] Amin, S., Uddin, M. I., Al-Baity, H. H., Zeb, M. A. & Khan, M. A. Machine Learning Approach for COVID-19 Detection on Twitter. *Cmc-Comput. Mater. Con.* **68**, 2231-2247(2021).
- [9] Neslihanoglu, S. Nonlinear models: a case of the COVID-19 confirmed rates in top 8 worst affected countries. *Nonlinear Dynam.* <https://dx.doi.org/10.1007%2Fs13204-021-01868-7>
- [10] Huang, H. N. et al. The impact of air pollution on COVID-19 pandemic varied within

- different cities in South America using different models. *Environ. Sci. Pollut. R.* <https://dx.doi.org/10.1007%2Fs13204-021-01868-7>
- [11] Pei, L. J. & Zhang, M. Y. Long-term predictions of current confirmed and dead cases of COVID-19 in China by the non-autonomous delayed epidemic models. *Cogn. Neurodynamics*. <https://dx.doi.org/10.1007%2Fs11571-021-09701-1>
- [12] Musa, S. S. et al. Estimation of COVID-19 under-ascertainment in Kano, Nigeria during the early phase of the epidemics. *Alex. Eng. J.* **60**, 4547-4554 (2021).
- [13] Stehlik, M., Kiselak, J., Dinamarca, M. A., Li, Y. & Ying, Y. On COVID-19 outbreaks predictions: Issues on stability, parameter sensitivity, and precision. *Stoch. Anal. Appl.* **39**, 383-404 (2021).
- [14] Oliveira, R. F. A. P., Bastos-Filho, C. J. A., Medeiros, A. C. A. M. V. F., Buarque, P. & Freire, D. L. Machine learning applied in SARS-CoV-2 COVID 19 screening using clinical analysis parameters. *IEEE. Lat. Am. T.* **19**, 978-985 (2021).
- [15] Abdulkareem, A. B. Predicting COVID-19 Based on Environmental Factors with Machine Learning. *Intell. Autom. Soft. Co.* **28**, 305-320 (2021).
- [16] El-dosuky, M. A., Soliman, M. & Hassanien, A. E. COVID-19 vs influenza viruses: A cockroach optimized deep neural network classification approach. *Int. J. Imag. Syst. Techo.* **31**, 472-482 (2021).
- [17] Tiwari, D., Bhati, B. S., Al-Turjman, F. & Nagpal, B. Pandemic coronavirus disease (Covid-19): World effects analysis and prediction using machine-learning techniques. *Expert. Syst.* <https://dx.doi.org/10.1111%2Fexsy.12714>
- [18] Rustam, F. et al. COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access.* **8**, 101489-101499 (2020).
- [19] Wang, S. R., McCormick, T. H. & Leek, J. T. Methods for correcting inference based on outcomes predicted by machine learning. *P. Natl. Acad. Sci. USA.* **117**, 30266-30275 (2020).
- [20] Fernandez-Ruiz, I. Machine learning predicts risk in ACS. *Nat. Rev. Cardiol.* **18**, 230-230 (2021).
- [21] Joodaki, H., Gepner, B. & Kerrigan, J. Leveraging machine learning for predicting human body model response in restraint design simulations. *Comput. Method. Biomec.* **24**, 597-611 (2021).
- [22] Feldman, J., Thomas-Bachli, A., Forsyth, J., Patel, Z. H. & Khan, K. Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise. *J. Am. Med. Inform. Assn.* **26**, 1355-1359 (2019).
- [23] Agany, D. D. M., Pietri, J. E. & Gnimpieba, E. Z. Assessment of vector-host-pathogen relationships using data mining and machine learning. *Comput. Struct. Biotec.* **18**, 1704-1721 (2020).
- [24] Liu, Z. T. & Hauskrecht, M. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artif. Intell. Med.* **65**, 5-18 (2015).
- [25] Lee, J. M. & Hauskrecht, M. Modeling multivariate clinical event time-series with recurrent temporal mechanisms. *Artif. Intell. Med.* **112**, 102021 (2021).
- [26] Lu, W. J., Ma, L., Chen, H., Jiang, X. J. & Gong, M. A Clinical Prediction Model in Health Time Series Data Based on Long Short-Term Memory Network Optimized by Fruit Fly Optimization Algorithm. *IEEE. Access.* **8**, 136014-136023 (2020).
- [26] Lu, W. J., Ma, L., Chen, H., Jiang, X. J. & Gong, M. A. Clinical Prediction Model in Health

- Time Series Data Based on Long Short-Term Memory Network Optimized by Fruit Fly Optimization Algorithm. *IEEE. Access.* **8**, 136014-136023 (2020).
- [27] He, M. CausalBG: Causal Recurrent Neural Network for the Blood Glucose Inference with IoT Platform. *IEEE. Internet Things.* **7**, 598-610 (2020).
- [28] Zecchin, C., Facchinetti, A., Sparacino, G., De Nicolao, G. & Cobelli, C. Neural Network Incorporating Meal Information Improves Accuracy of Short-Time Prediction of Glucose Concentration. *IEEE. T. Bio-Med. Eng.* **59**, 1550-1560 (2012).
- [29] Wang, W. B., Tong, M., Yu, M. & Lee, J. M. Blood Glucose Prediction with VMD and LSTM Optimized by Improved Particle Swarm Optimization. *IEEE. Access.* **8**, 217908-217916 (2020).
- [30] Li, Y. T., Bao, T. F., Gong, J., Shu, X. S. & Zhang, K. The Prediction of Dam Displacement Time Series Using STL, Extra-Trees, and Stacked LSTM Neural Network. *IEEE. Access.* **8**, 94440-94452 (2020).
- [31] Cheng, L. F. et al. Sparse multi-output Gaussian processes for online medical time series prediction. *BMC. Med. Inform. Decis.* **20**, 152-152(2020).
- [32] Xu, X. J., Tang, Q., Xia, H. Y., Zhang, Y. L., Li, W. Q. & Huo, X. Chaotic time series prediction for prenatal exposure to polychlorinated biphenyls in umbilical cord blood using the least squares SEATR model. *Sci. Rep-UK.* **6**, 287-288 (2016).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarydataset.xlsx](#)
- [SupplementaryInformation.docx](#)