

RESEARCH ARTICLE

Prediction and analysis of Corona Virus Disease 2019

Yan Hao¹, Ting Xu², Hongping Hu², Peng Wang², Yanping Bai^{2*}¹ School of Information and Communication Engineering, North University of China, Taiyuan, China,² Department of Mathematics, School of Science, North University of China, Taiyuan, China

* baiyp666@163.com



Abstract

The outbreak of Corona Virus Disease 2019 (COVID-19) in Wuhan has significantly impacted the economy and society globally. Countries are in a strict state of prevention and control of this pandemic. In this study, the development trend analysis of the cumulative confirmed cases, cumulative deaths, and cumulative cured cases was conducted based on data from Wuhan, Hubei Province, China from January 23, 2020 to April 6, 2020 using an Elman neural network, long short-term memory (LSTM), and support vector machine (SVM). A SVM with fuzzy granulation was used to predict the growth range of confirmed new cases, new deaths, and new cured cases. The experimental results showed that the Elman neural network and SVM used in this study can predict the development trend of cumulative confirmed cases, deaths, and cured cases, whereas LSTM is more suitable for the prediction of the cumulative confirmed cases. The SVM with fuzzy granulation can successfully predict the growth range of confirmed new cases and new cured cases, although the average predicted values are slightly large. Currently, the United States is the epicenter of the COVID-19 pandemic. We also used data modeling from the United States to further verify the validity of the proposed models.

OPEN ACCESS

Citation: Hao Y, Xu T, Hu H, Wang P, Bai Y (2020) Prediction and analysis of Corona Virus Disease 2019. PLoS ONE 15(10): e0239960. <https://doi.org/10.1371/journal.pone.0239960>

Editor: Jerry Chun-Wei Lin, Western Norway University of Applied Sciences, NORWAY

Received: April 26, 2020

Accepted: September 16, 2020

Published: October 5, 2020

Copyright: © 2020 Hao et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data underlying the results presented in the study are available from <http://wjw.wh.gov.cn/> and <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>, and are provided in the Supporting Information files.

Funding: YP B National Nature Science Foundation of China (Grant No. 61774137) <http://www.nsf.gov.cn/> Key Research and Development Projects of Shanxi Province (Grant No.201903D121156) <http://kjt.shanxi.gov.cn/> HP H Shanxi Natural Science Foundation (Grant No.201801D121026) <http://kjt.shanxi.gov.cn/> P W Shanxi Scholarship

Introduction

Infectious diseases are caused by various pathogens that can be transmitted from person to person, animal to animal, or person to animal. They can be transmitted in various ways, and the speed of transmission is fast. Early diagnosis of infectious diseases is crucial, and prevention and control are paramount.

In December 2019, an unexplained viral pneumonia was reported in Wuhan, China. The virus, named the 2019 Novel Coronavirus (2019-nCoV) by the World Health Organization (WHO) on January 12, 2020, causes Corona Virus Disease 2019 (COVID-19), and is currently the seventh known species of coronavirus that can infect human beings. The remaining six species are human coronavirus (HCoV)-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1, severe acute respiratory syndrome (SARS)-CoV, and Middle East respiratory syndrome (MERS)-CoV. COVID-19 and severe acute respiratory syndrome (SARS), which broke out in 2003, are both caused by coronaviruses. Their clinical symptoms are similar but not identical.

Council of China (Grant No.2016-088) <http://kjt.shanxi.gov.cn/> Z J Development Projects of Shanxi Province (202003D31011/GZ) <http://kjt.shanxi.gov.cn/> We have added a new fund based on the original funding information (Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (No. 2020L0283)). This work was supported by these funders, but the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

People infected with the 2019-nCoV will experience varying symptoms, including fever, mild cough, or pneumonia, occasionally leading to death. The mortality rate of COVID-19 is approximately 2% to 4%, although this is an extremely early percentage and may change as more information becomes available. Meanwhile, this does not mean that the virus is not serious, it simply means that not everyone infected with it will face the worst outcome. We are currently in a tense state of prevention and control, and are concerned about a global pandemic.

2019-nCoV is an intractable and unexpected virus with the following characteristics. First, it has a strong ability to camouflage itself. The virus can be present asymptotically, making an infected person appear healthy, or can manifest with various other symptoms in which an infected person seems to be suffering from a common respiratory disease. It also has a long latency period. According to a report from *Lancet*, the median latency time of the new coronavirus is approximately 20 days, reaching up to 37 days for some patients. In addition, it is transmitted in various ways. The transmission method of the 2019-nCoV is similar to that of various other infectious diseases. In addition to a traditional droplet transmission, it also includes contact, air, fecal–oral, and blood transmissions. A high relapse rate has also been seen. Since the initial outbreak of COVID-19, there have been many cases of recovered patients testing positive again after a reexamination, even leading to death after a relapse. The virus has also shown variability. As a foreboding aspect of the future development of novel coronaviruses, their rapid genetic recombination results in a mutation into new strains. Each recombination can increase the toxicity and infectivity, invalidating original treatment methods and drugs. Forty variants of novel coronaviruses have been discovered by scientists in Iceland. Finally, it is highly infectious. Scientific data have shown that the affinity between the S-protein of 2019-nCoV and angiotensin converting enzyme 2 (ACE2) is 10 to 20 times that of SARS, which means that the infectivity of COVID-19 is significantly higher than that of SARS.

There are currently no specific treatments for COVID-19. However, many symptoms can be treated, and treatment must be given according to the clinical status of the patient. In addition, supplementary care for those infected may be effective. Self-protection includes maintaining basic hand and respiratory hygiene, adhering to safe eating habits, and avoiding close contact with anyone who shows symptoms of a respiratory disease (such as coughing or sneezing). At present, owing to the widespread nature of COVID-19, factories are being shut down, schools are being suspended, and people are isolating in their own homes, significantly disrupting daily life. It is therefore extremely important to reasonably predict and analyze the development trend of this pandemic.

In previous studies, the prediction methods for the occurrence, spread, and change in infectious diseases mainly included regression prediction models [1, 2], Markov chain models [3], Bayesian networks [4–6], and other machine learning methods [7–9]. Most of the previous studies have been based on research related to influenza, with a few studies conducted on human immunodeficiency virus (HIV) infection [10, 11] and SARS [12, 13].

For example, Ray [14] used ensemble methods with three component models to predict the seasonal timing of influenza and severity in the United States. Xue et al. [15] established five models for predicting and assessing influenza outbreaks across 10 regions of the United States. Experiments on Google Flu Trends (GFT) and Centers for Disease Control (CDC) data have shown that GFT and historical influenza data are partially complementary and that GFT + CDC regression is a preferable model. In addition, Alkouz et al. [16] proposed a system called Tweetluenza, which can predict the spread of influenza in real time.

Many scholars have published reports on COVID-19. For example, in [17–20], its clinical features are described, and in [19] and [20], the authors demonstrated that COVID-19 has a high comorbidity with hypertension, diabetes, and cardiovascular diseases. Anastassopoulou et al. [21] estimated the case fatality and case recovery ratios based on a susceptible infectious-

recovered-dead (SIDR) model, along with their 90% confidence intervals as the outbreak evolves. In [22–27], the authors use machine learning methods to diagnose COVID-19 based on X-ray and CT images. Randhawa et al. [28] also used machine learning-based methods to achieve ultra-fast, scalable, and highly accurate classification of whole COVID-19 virus genomes based on the COVID-19 intrinsic virus genomic signature. In [29–31], other characteristics related to COVID-19 were studied.

To date, few studies related to the use of quantitative prediction methods to predict the development trend and growth range of COVID-19 have been conducted. In the present study, three methods, namely, an Elman neural network, LSTM, and SVM are applied to predict and analyze COVID-19 data from January 23, 2020 to April 6, 2020 in Wuhan, Hubei Province, China, including cumulative confirmed cases, confirmed new cases, cumulative deaths, new deaths, and cumulative cured cases and new cured cases. First, the Elman neural network, LSTM, and SVM were used to predict the development trend of cumulative confirmed cases, cumulative deaths, and cumulative cured cases. Next, the SVM with fuzzy granulation was used to predict the growth range of confirmed new cases, new deaths, and new cured cases. Experimental results showed that the Elman neural network and SVM adopted in this study can accurately predict the development trend of COVID-19, whereas LSTM is more suitable for the prediction of cumulative confirmed cases. The SVM with fuzzy granulation is effective for the growth range prediction of confirmed new cases, new deaths, and new cured cases, despite the larger averages. The same models were also used on data in the United States to verify the robustness of the models.

The rest of this paper is organized as follows. Section 2 introduces the methods applied. Section 3 describes the experiments and analysis results. Finally, some concluding remarks and areas of future research are provided in Section 4.

Methods

There are two main research topics in this paper. The first is the development trend prediction and analysis of the cumulative confirmed cases, deaths, and cured cases. The second is the growth range prediction of the confirmed new cases, new deaths, and new cured cases. Because the present state of cumulative confirmed cases, deaths, and cured cases is related to the state of the previous day, and shows a non-linear growth, recurrent neural networks (RNNs) are applicable. In this study, we mainly apply the Elman neural network and LSTM. In addition, the SVM has a better prediction effect on non-linear data, therefore, the SVM model is also introduced for development trend prediction. Because there are many factors affecting the growth of new cases, the growth trend fluctuates significantly. Here, only the growth range of the new cases is predicted. To predict the change space of the three datasets more accurately, this paper introduces a SVM model with fuzzy granulation.

A. Elman neural network

The Elman neural network is generally divided into four layers: input, hidden, receiving, and output layers. The receiving layer is also called the context layer or state layer, and is used to memorize the output value of the hidden layer unit at the previous moment and return it to the network input. It can be considered a one-step delay operator. As the characteristic of the Elman neural network, the output of the hidden layer is self-connected to the input of the hidden layer through the delay and storage of the receiving layer. This self-connected method makes it sensitive to historical data. The addition of an internal feedback network enhances the ability of the network to process dynamic information, thereby achieving the purpose of dynamic modeling. The network structure of the Elman neural network is shown in Fig 1.

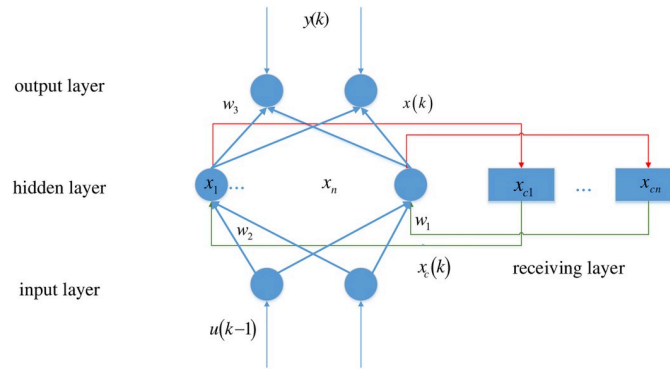


Fig 1. Network structure of the Elman neural network.

<https://doi.org/10.1371/journal.pone.0239960.g001>

The connections of the input layer, hidden layer, and output layer are similar to those of a feedforward network. Unlike a feedforward network, the Elman neural network contains a receiving layer, which takes the hidden layer state of the previous moment together with the network input of the current moment as the input of the hidden layer.

Here, $x_c(k) = x(k - 1)$ indicates that the feedback state vector is equal to the hidden layer state at the previous moment. The hidden layer state is expressed as follows:

$$x(k) = f(w_1x_c(k) + w_2(u(k - 1))).$$

The output layer state is expressed as

$$y(k) = g(w_3x(k)),$$

where u is the input vector, x_c is the feedback state vector, w_1 is the weight of the receiving layer to the hidden layer, w_2 is the weight of the input layer to the hidden layer, w_3 is the weight of the hidden layer to the output layer, f^* is the activation function of the hidden layer, and g^* is the activation function of the output layer.

B. Long short-term memory

LSTM networks are variants of RNNs. General RNNs only have short-term memory owing to a disappearance of the gradients. LSTM networks combine short-term and long-term memory through a subtle gate control and solve the problem of disappearing gradients to a certain extent. LSTM was proposed by Hochreiter and Schmidhuber [32] in 1997. It has three special structures: a forget gate, an input gate, and an output gate. LSTM networks can delete or add information to a cell state through a structure called a gate. Fig 2 shows a diagram of the LSTM network.

The diagram shows the details of the forget gate, input gate, and output gate. The first step is to decide what information to discard from the cell state, which is done through a layer called the forget gate. The forget gate reads the hidden state of the previous moment h_{t-1} and the current input data x_t , and then outputs a vector between zero and 1. The value between zero and 1 in this vector indicates how much information is retained or discarded in the cell state c_t . A value of zero means all information is discarded, and 1 means all information is retained.

$$z_f = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f).$$

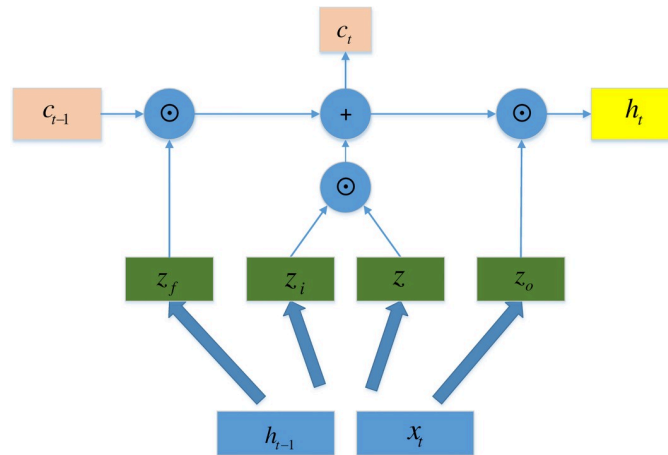


Fig 2. Diagram of the LSTM.

<https://doi.org/10.1371/journal.pone.0239960.g002>

The next step is to decide how much new information is added to the cell state. Two steps are needed to achieve this. First, we use h_{t-1} and x_t to decide which information to update through an input gate operation. We then use h_{t-1} and x_t to obtain a new candidate cell state z through a tanh layer.

$$z_i = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i),$$

$$z = \tanh(W \cdot [h_{t-1}, x_t] + b).$$

Next, the cell state is updated as follows:

$$c_t = z_f \odot c_{t-1} + z_i \odot z.$$

Finally, we need to determine the output value. After updating the cell state, we need to determine which part of the cell state will be output according to the input h_{t-1} and x_t . Here, we need to pass the input through a sigmoid layer called the output gate to obtain the judgment conditions. We then need to pass the cell state through the tanh layer to obtain a vector between -1 and 1, which is multiplied by the judgment conditions obtained by the output gate to obtain the output.

$$z_o = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),$$

$$h_t = z_o \odot \tanh(c_t),$$

where z_i , z_f and z_o are the gate control state of the forget gate, input gate, and output gate, respectively. In addition, z is the input through a tanh layer, which is a so-called candidate cell state. Finally, \odot represents a multiply operation of the corresponding elements of the matrix.

C. Support vector machine

SVM is a classic model that can be used not only for classification but also for regression. We do not delve into its theoretical derivation here, however. Unlike with a classification problem, the output of the regression problem is no longer a discrete value, but a continuous value. In reality, it is often impossible to accurately predict the value of the COVID-19, and to this end,

it is particularly important to predict the development trend and change space for the important parameters of this disease.

Without considering other factors, time is clearly an important independent variable affecting COVID-19. To more accurately predict the change space of each group of data on this disease, in this paper, a fuzzy granulation model is introduced. First, we divide the original data into multiple granulation windows and then apply fuzziness to each window.

The so-called fuzzy granulation means that the information granulation adopted is based on the model of fuzzy set theory. The key to fuzzy granulation is fuzzification, here, we use the Pedrycz's granulation method [33]. For a given time series, we consider the entire time series X as a window and then apply fuzziness to it. Fuzzification is a task used to establish a fuzzy particle P on X , that is, a fuzzy concept G that can reasonably describe X . Fuzzy particles can reasonably represent the original data and have certain specialties. A fuzzy particle P can be described simply as $P = A(x)$, where A is a membership function of the fuzzy concept G . Commonly used fuzzy particles include triangular, ladder, Gaussian, and parabolic particles. In this study, we choose triangular fuzzy particles, the membership function of which is as follows:

$$A(x, a, m, b) = \begin{cases} 0, & x < a \\ \frac{x-a}{m-a}, & a \leq x \leq m \\ \frac{b-x}{b-m}, & m < x \leq b \\ 0, & x > b \end{cases}$$

Here, m is the kernel of triangular fuzzy particles, and a and b are the lower and upper bounds of the support, respectively.

Results

We studied COVID-19 data from January 23, 2020 to April 6, 2020, which were published by the Wuhan Municipal Health Commission (<http://wjw.wh.gov.cn/>) [34]. We used the Elman neural network, LSTM, and SVM to predict and analyze the development trend of cumulative confirmed cases, cumulative deaths, and cumulative cured cases. As an additional study, we also used the SVM with fuzzy granulation to predict and analyze the growth range of confirmed new cases, new deaths, and new cured cases. To further demonstrate the robustness of the models, we used the same models on data from the United States from March 23, 2020 to June 5, 2020. The data were obtained from the World Health Organization (WHO) [35]. The data from the United States contains confirmed cases, deaths while without cured cases. All experiments were conducted in MATLAB (R2014a). Fig 3(a) shows a line chart of the confirmed new cases. From Fig 3(a), it can be seen that, from January 23, 2020 to February 11, 2020, the number of confirmed new cases showed a fluctuating upward trend. confirmed new cases fluctuated sharply on February 12, showing a peak, the reason for which was the announcement of 12,364 clinically confirmed cases for the first time on that day, with a sharp increase in confirmed new cases. Fig 3(b) shows a line chart of the new deaths and cured cases. The number of new cured cases showed an increasing trend before February 27, after which the growth trend of the cured cases slowed. Moreover, the number of new deaths showed a significant downward trend after February 24, demonstrating a promising future recovery.

A. Development trend prediction for data of Wuhan

In this section, the Elman neural network, LSTM, and SVM were used to predict and analyze the development trends of the cumulative confirmed cases, deaths, and cured cases. In the

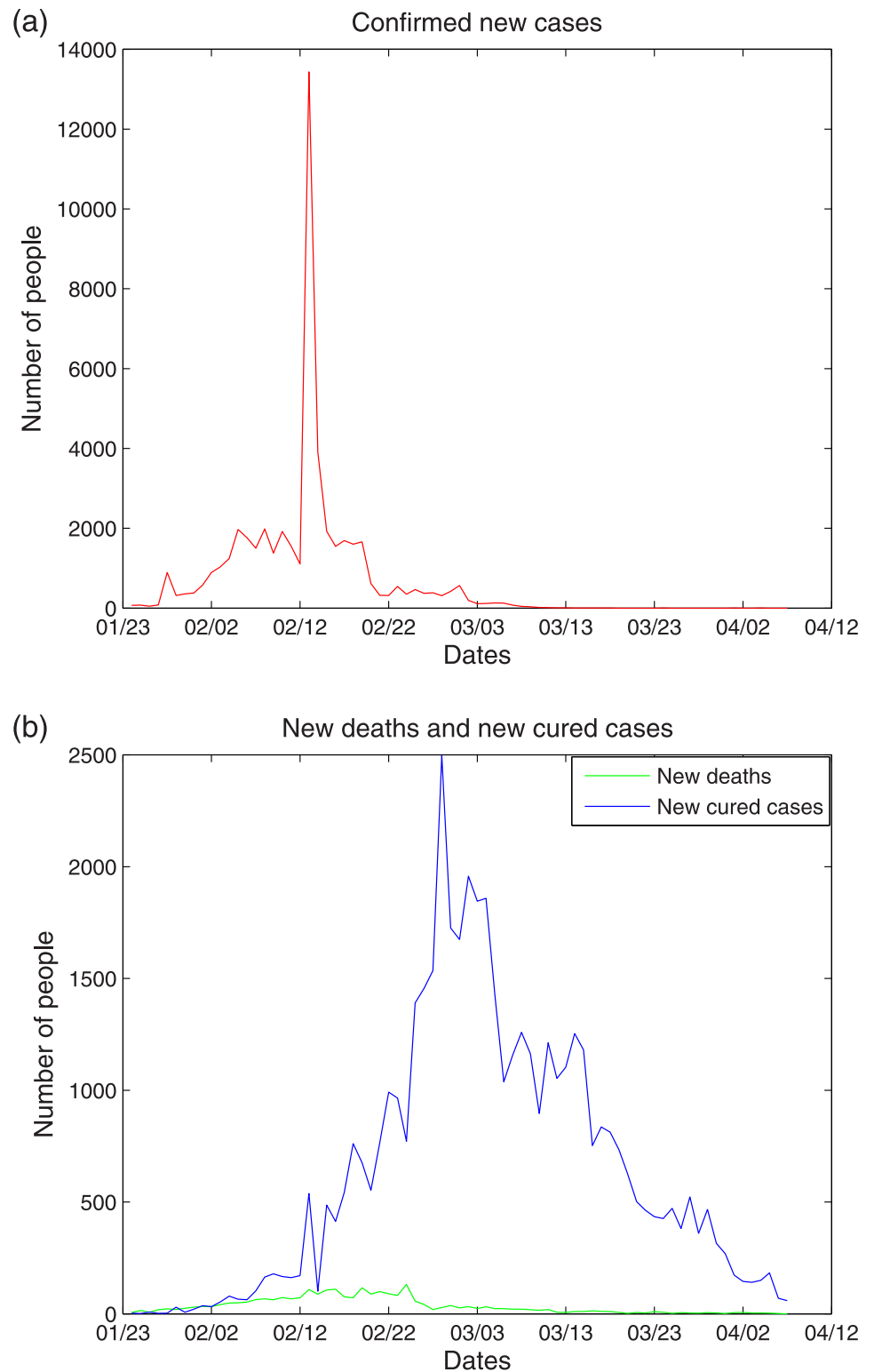


Fig 3. (a) The line chart of confirmed new cases. (b) The line chart of new deaths and new cured cases.

<https://doi.org/10.1371/journal.pone.0239960.g003>

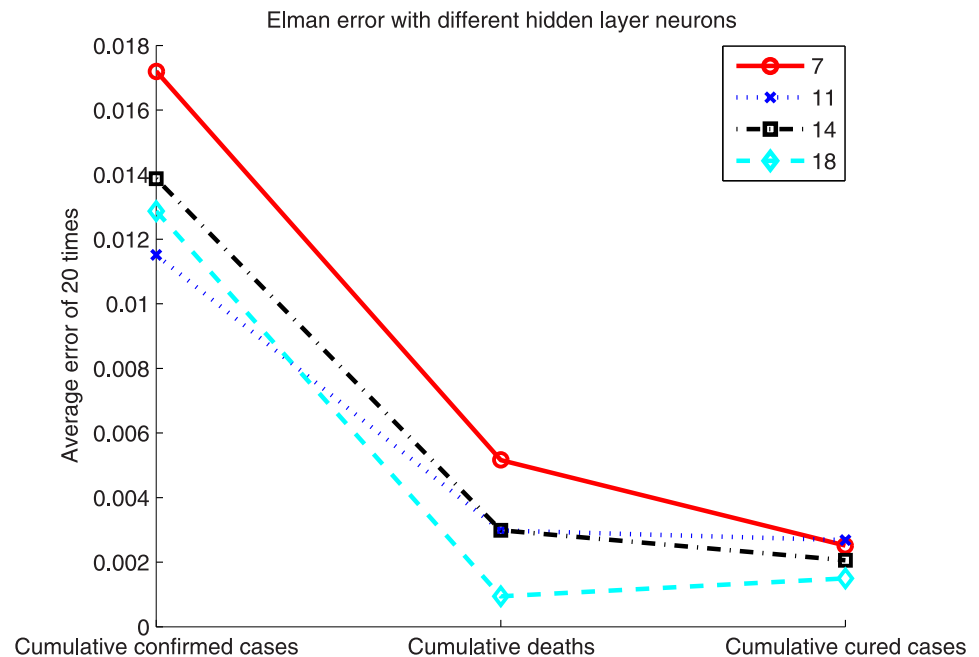


Fig 4. Error of the Elman neural network with different number of hidden layer neurons.

<https://doi.org/10.1371/journal.pone.0239960.g004>

Elman neural network and LSTM, the experimental data are in the form of the first three days to the fourth (i.e., using data from the first three days to predict the data on day four). Thus, the original 75 groups of data are converted into 72 groups, and the last group of data are used as the test sample. According to the data, both the Elman neural network and the LSTM in this paper contain only one hidden layer.

Because the initial weights of the Elman neural network and LSTM were randomly initialized, the average of 20 experimental results is given here as the final outcome. In addition, the number of hidden layer neurons is an important factor. We chose 7, 11, 14, and 18 hidden layer neurons according to experience. Fig 4 shows the error in the Elman neural network based on these different numbers of hidden layer neurons.

Fig 4 shows that the effect of the number of hidden neurons on the error is not too obvious. By comparison, with 18 hidden neurons, the prediction result is slightly better. In addition, the prediction error of the cumulative confirmed cases is larger than that of the cumulative deaths and cumulative cured cases because the number of confirmed cases increased dramatically owing to the clinically confirmed cases announced on February 12, which has had a significant impact on the prediction results. Fig 5 shows the error rates of the Elman neural network and LSTM when the number of hidden layer neurons is 18. The prediction errors of the two methods are listed in Table 1.

As can be seen from Fig 5 and Table 1, the prediction results of the LSTM are worse than those of the Elman neural network, particularly for cumulative cured cases. As is well-known, LSTM is a powerful model for time-series prediction. However, the novel coronavirus pneumonia data have no obvious periodicity, and LSTM has a relatively poor prediction effect on data with a weak periodicity. Regarding the poor result of the cumulative cured cases, no effective treatment has been found, and fewer people were cured during the initial stage of the pandemic. In addition, it takes time between treatment and recovery. The treatment period varies from person to person. With continuous research, the number of cured cases continues to increase nonlinearly. For irregular growth, neural networks tend to fall into the local optimum,

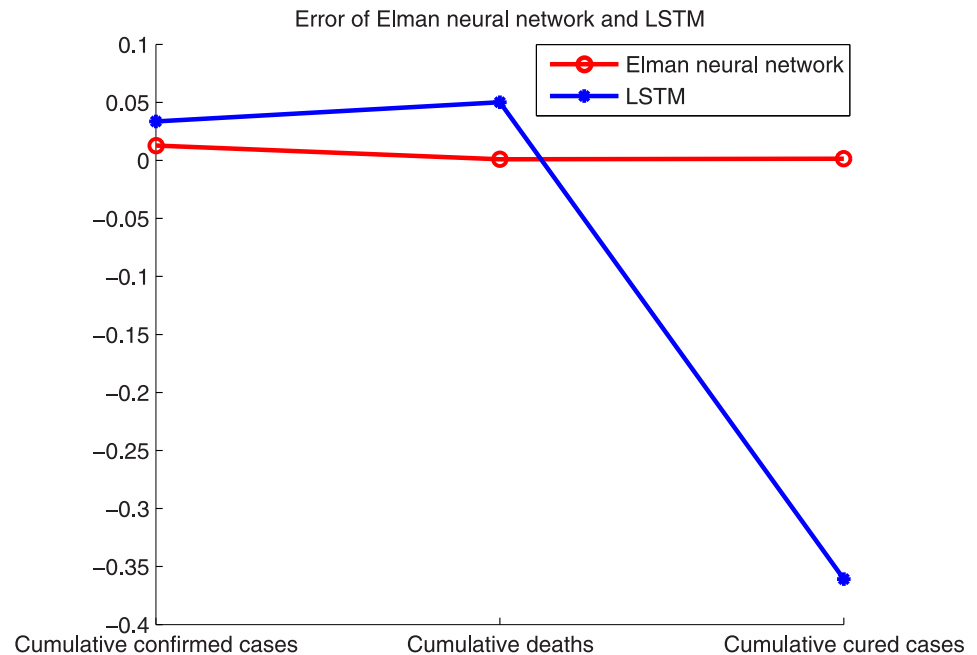


Fig 5. Error of the Elman neural network and LSTM.

<https://doi.org/10.1371/journal.pone.0239960.g005>

making it difficult to achieve an accurate prediction, which is also a major drawback of a neural network.

In the SVM experiment, we used a one-day-to-the-next approach, that is, we used the data from the previous day to predict data for the next day. In this study, data of 61 days from January 23 to March 23 were used as training samples, and data from March 24 to April 6 were used as test samples. The radial basis function was used in the experiments. Fig 6 shows the prediction results of the cumulative confirmed cases, deaths, and cured cases.

The prediction mean square error and square correlation coefficient of the cumulative confirmed cases, deaths, and cured cases are shown in Table 2.

It can be seen from the results that the SVM can predict the three sets of data well, and the square correlation coefficients of the cumulative deaths and cumulative cured cases between the predicted output and the actual output are all above 99%. Although the mean square error of the cumulative confirmed cases is relatively large and the square correlation coefficient is relatively small, it can be seen from Fig 6(a) that the fluctuation range of the cumulative confirmed cases from March 24 to April 6 is extremely small, and only two of the 14-day prediction results are incorrect.

B. Development trend prediction for data of the United States

Similar to the aforescribed experiments, herein we predict and analyze the development trend of cumulative confirmed cases and deaths from the data in the United States. Here, 12

Table 1. Error of the Elman neural network and LSTM for data of Wuhan.

Methods	Cumulative confirmed cases	Cumulative deaths	Cumulative cured cases
Elman neural network	0.0129	0.0009	0.0015
LSTM	0.0336	0.0501	-0.3609

<https://doi.org/10.1371/journal.pone.0239960.t001>

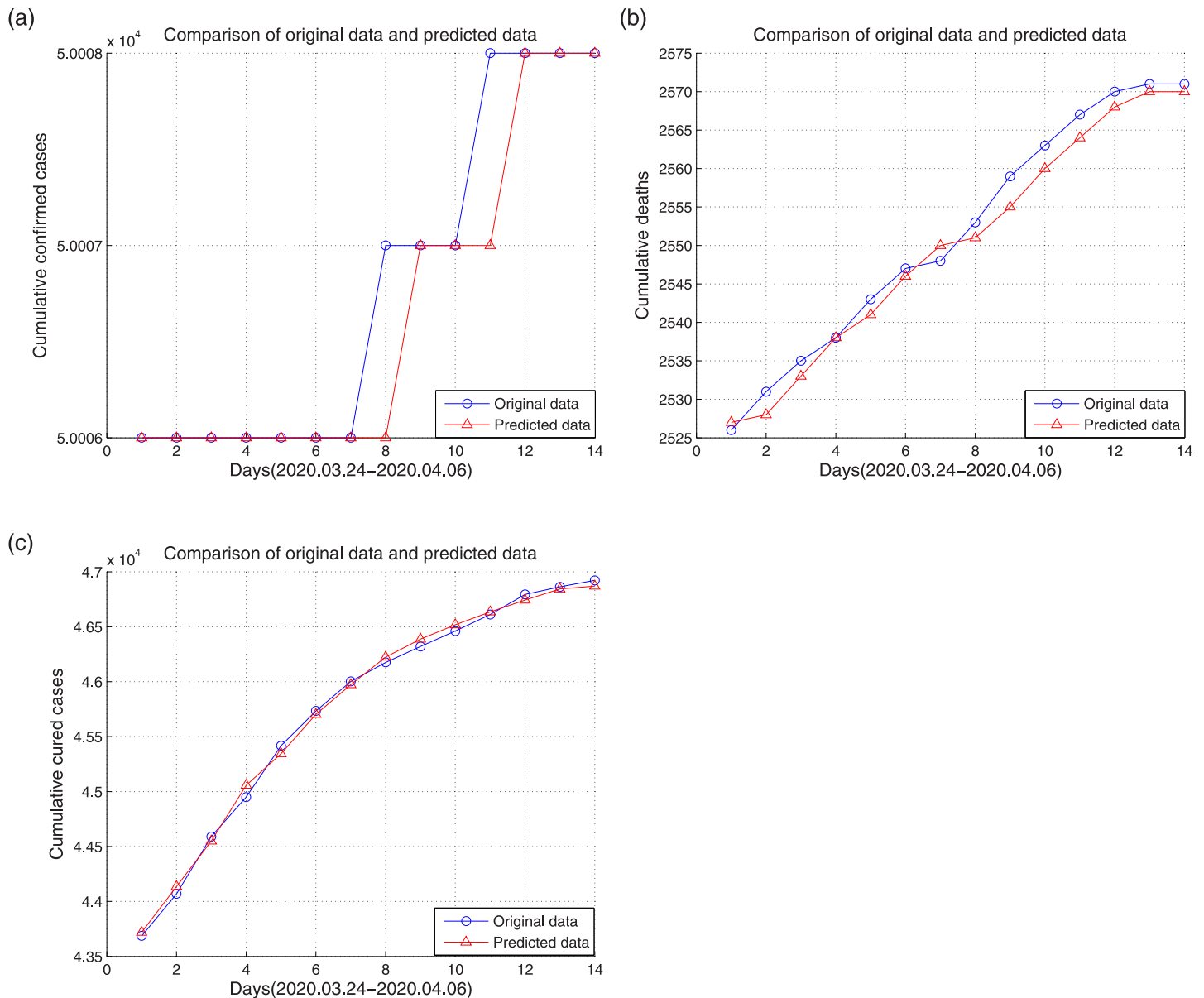


Fig 6. Prediction results of the cumulative confirmed cases, deaths and cured cases. (a) cumulative confirmed cases, (b) cumulative deaths, (c) cumulative cured cases.

<https://doi.org/10.1371/journal.pone.0239960.g006>

hidden layer neurons of the Elman neural network and LSTM are applied. Table 3 shows the prediction error of the two methods on data from the United States.

In the SVM experiment, data of 61 days from March 23 to May 22 were used as training samples, and data from May 23 to June 5 were used as test samples. The prediction results of the cumulative confirmed cases and deaths are shown in Fig 7. The mean square error and

Table 2. Mean square error and square correlation coefficient for data of Wuhan.

Evaluation criterion	Cumulative confirmed cases	Cumulative deaths	Cumulative cured cases
Mean square error	0.0320	0.0023	0.0003
Square correlation coefficient	84.51%	99.06%	99.74%

<https://doi.org/10.1371/journal.pone.0239960.t002>

Table 3. Error of the Elman neural network and LSTM for data of the United States.

Methods	Cumulative confirmed cases	Cumulative deaths
Elman neural network (18)	-0.0049	-0.0011
LSTM (18)	-0.1277	0.1118

<https://doi.org/10.1371/journal.pone.0239960.t003>

square correlation coefficient of the cumulative confirmed cases and deaths are shown in [Table 4](#).

Comparing the experimental results of Wuhan and the United States, the Elman neural network and SVM can well predict the development trend of Wuhan and the United States, whereas the prediction result of LSTM for the United States is clearly worse than that of Wuhan. This occurs because the data from the United States used in this study are national data, not data from a certain city, and the data growth is more irregular and without any periodicity.

C. Growth range prediction for data of Wuhan

In addition to predicting the development trend of COVID-19, we also used a SVM with fuzzy granulation to predict the growth range of the confirmed new cases, new deaths, and new cured cases. Taking 5 days as a granulation window, the data from January 23, 2020 to April 1, 2020 were used as the training samples, and the growth range of the next 5 days was predicted, that is, the growth range of the confirmed new cases, new deaths, and new cured cases from April 2 to April 6. To reduce the impact of the peak value on the experimental results, we first smoothed the data of confirmed new cases and new cured cases. [Table 5](#) shows the predicted growth range of confirmed new cases, new deaths, and new cured cases.

[Table 5](#) shows that the growth ranges predicted in this paper are basically reasonable, although the growth range of new deaths is slightly large. However, the average predicted values are larger than actual average values. This is because the algorithm depends on the data, regardless of the actual situation. Owing to continuous medical research and strict control, there are increasingly fewer confirmed new cases and existing confirmed cases in Wuhan, resulting in increasingly fewer new deaths and cured cases. In the middle of the epidemic, the numbers of confirmed new cases and new cured cases increased rapidly, leading to a higher predicted average value. In addition, the number of confirmed cases had steadily increased from January 23 to February 11. However, on February 12, 12,364 clinically confirmed cases were announced for the first time and included in the confirmed new cases. The subsequent growth trends were completely inconsistent, which severely affected the prediction results.

D. Growth range prediction for data of the United States

The data from March 23, 2020 to May 31, 2020 were used as the training samples, and the growth range of the confirmed new cases and new deaths from June 1 to June 5 were predicted. [Table 6](#) shows the results.

The predicted growth range of confirmed new cases and new deaths is smaller than the actual growth. The predicted average for confirmed new cases is larger than the actual average of 19,648.6, whereas the predicted average for new deaths is smaller than the actual average of 879. The negative value in the public data has a significant influence on the prediction results (the number of confirmed new cases on May 10 is -99, and the number of new deaths on May 4 is -1696). There is no obvious downward trend in confirmed new cases and new deaths, and strict control is still required. Individuals should therefore strengthen their awareness of prevention and control and be responsible for themselves and others.

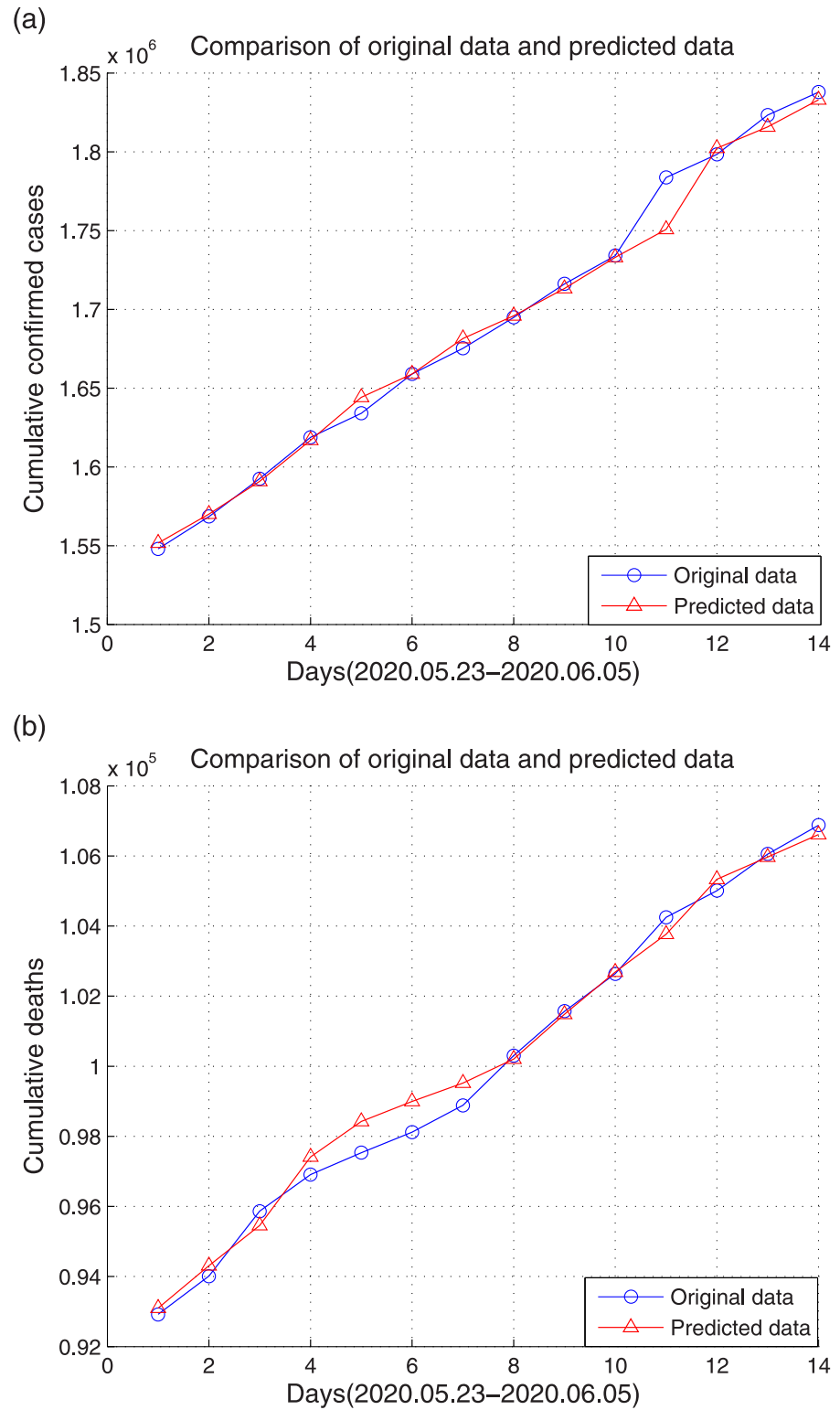


Fig 7. Prediction results of the cumulative confirmed cases and deaths. (a) cumulative confirmed cases, (b) cumulative deaths.

<https://doi.org/10.1371/journal.pone.0239960.g007>

Table 4. Mean square error and square correlation coefficient for data of the United States.

Evaluation criterion	Cumulative confirmed cases	Cumulative deaths
Mean square error	0.0011	0.0011
Square correlation coefficient	99.03%	99.12%

<https://doi.org/10.1371/journal.pone.0239960.t004>

Table 5. Predicted growth range of the confirmed new cases, new deaths and new cured cases for data of Wuhan.

Dates	Actual growth					Predicted growth range	Predicted average
	April 2	April 3	April 4	April 5	April 6		
Confirmed new cases	0	1	0	0	0	[0,260]	160
New deaths	4	4	3	1	0	[4,10]	8
New cured cases	141	150	183	69	59	[24,498]	222

<https://doi.org/10.1371/journal.pone.0239960.t005>

Table 6. Predicted growth range of confirmed new cases and new deaths for data of the United States.

Dates	Actual growth					Predicted growth range	Predicted average
	June 1	June 2	June 3	June 4	June 5		
Confirmed new cases	17962	26116	14692	24890	14583	[18761,21027]	20045
New dead cases	1073	693	761	1043	825	[752,926]	813

<https://doi.org/10.1371/journal.pone.0239960.t006>

Conclusions and future perspectives

The emergence of COVID-19 has been a heavy burden. During the past 2 months, however, significant progress has been made in the diagnosis and treatment of this disease. In this study, machine learning was used to predict and analyze the development trend and growth range of COVID-19. The results using data from Wuhan showed that the Elman neural network and SVM can better predict the development trend of cumulative deaths and cumulative cured cases, whereas the prediction error of cumulative confirmed cases is relatively large. In comparison, the LSTM model has a worse predictive effect on cumulative confirmed cases, deaths, and cured cases owing to the aperiodic data. Moreover, the prediction results of the two recurrent neural network models used in this study are unstable because neural networks tend to fall into a local optimum for data with irregular growth. This is a major drawback of neural networks, and needs to be improved in the future. For the prediction of the growth range of confirmed new cases, new deaths, and new cured cases, the SVM with fuzzy granulation introduced in this paper was shown to be effective. However, the predicted average values are larger than the actual average values, which indicates that the model is less robust to data with large fluctuations and still needs to be improved. The experimental results on data from the United States verified the robustness of the models, although the prediction result of LSTM was worse than that of Wuhan. In future research, we will improve the models based on the aforementioned problems and continue to improve the generalizability of the models.

Supporting information

S1 Data. Data of the United States.

(XLSX)

S2 Data. Data of Wuhan.

(XLSX)

S1 File. Data description.

(DOCX)

S2 File. Manuscript with marks.

(DOCX)

Author Contributions**Data curation:** Ting Xu, Hongping Hu.**Methodology:** Yan Hao.**Software:** Yan Hao, Peng Wang.**Supervision:** Ting Xu, Hongping Hu, Peng Wang, Yanping Bai.**Writing – original draft:** Yan Hao.**Writing – review & editing:** Yan Hao, Yanping Bai.**References**

1. Hirose H, Wang L. Prediction of Infectious Disease Spread Using Twitter: A Case of Influenza. *International Symposium on Parallel Architectures*. IEEE, 2013.
2. Chen Y, Chu CW, Chen MIC, Cook AR. The utility of LASSO-based models for real time forecasts of endemic infectious diseases: a cross country comparison. *Journal of Biomedical Informatics*, 2018, 81.
3. Jewell CP, Kyraios T, Christley RM, Roberts GO. A novel approach to real-time risk prediction for emerging infectious diseases: A case study in Avian Influenza H5N1. *Preventive Veterinary Medicine*, 2009, 91(1):19–28. <https://doi.org/10.1016/j.prevetmed.2009.05.019> PMID: 19535161
4. Vinarti RA, Hederman LM. A personalized infectious disease risk prediction system. *Expert Systems with Application*, 2019, 131:266–274.
5. Aulia R, Hederman L. Personalization of Infectious Disease Risk Prediction: Towards Automatic Generation of a Bayesian Network. *IEEE International Symposium on Computer-Based Medical Systems 2017*. IEEE, 2017.
6. Stojanović O, Leugering J, Pipa G, Ghozzi S, Ullrich A. A Bayesian Monte Carlo approach for predicting the spread of infectious diseases. *PLoS ONE* 14(12): e0225838. <https://doi.org/10.1371/journal.pone.0225838> PMID: 31851680
7. Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics*, 2017, 66:82–94. <https://doi.org/10.1016/j.jbi.2016.12.007> PMID: 28034788
8. Ma Q, Xiao L. Prediction model of BP neural network based on improved genetic algorithm optimization for infectious diseases. *2017 Chinese Automation Congress (CAC)*, 2017, pp. 4225–4229.
9. Zhu S, Emrich SJ, Chen DZ. Predicting Local Inversions Using Rectangle Clustering and Representative Rectangle Prediction. *IEEE Transactions on NanoBioscience*, 2019, 18(3): 316–323. <https://doi.org/10.1109/TNB.2019.2915060> PMID: 31180865
10. Ying H, Lin F, Macarthur RD, Cohn JA, Barth-Jones DC, Ye H, et al. A Self-Learning Fuzzy Discrete Event System for HIV/AIDS Treatment Regimen Selection. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics*, 2007, 37(4):966–979.
11. Andonie R, Fabry-Asztalos L, Abdul-Wahid CB, Abdul-Wahid S, Barker GI, Magill LC. Fuzzy ARTMAP Prediction of Biological Activities for Potential HIV-1 Protease Inhibitors Using a Small Molecular Data Set. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, 2011, 8(1):80–93.
12. Mitchell HD, Eisfeld AJ, Sims AC, McDermott JE, Matzke MM, Webb-Robertson BJ, et al. A Network Integration Approach to Predict Conserved Regulators Related to Pathogenicity of Influenza and SARS-CoV Respiratory Viruses. *PLoS ONE* 2013, 8(7): e69374. <https://doi.org/10.1371/journal.pone.0069374> PMID: 23935999
13. Bai Y, Jin Z. Prediction of SARS epidemic by BP neural networks with online prediction strategy. *Chaos Solitons & Fractals*, 2005, 26(2):559–569.

14. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. *Plos Computational Biology*, 2018, 14(2):e1005910. <https://doi.org/10.1371/journal.pcbi.1005910> PMID: 29462167
15. Xue H, Bai Y, Hu H, Liang H. Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network. *IEEE Access*, 2018, 6: 563–575.
16. Alkouz B, Aghbari ZA, Abawajy JH. Tweetluenza: Predicting Flu Trends from Twitter Data. *Big Data Mining and Analytics*, 2019(4).
17. Woelfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Mueller MA, et al. Clinical presentation and virological assessment of hospitalized cases of coronavirus disease 2019 in a travel-associated transmission cluster. *MedRxiv*, 2020. <https://www.medrxiv.org/content/10.1101/2020.03.05.20030502v1>
18. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 2020, 395: 497–506.
19. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA*, 2020, 323 (11): 1061–1069. <https://doi.org/10.1001/jama.2020.1585> PMID: 32031570
20. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical characteristics of 2019 novel coronavirus infection in China. *MedRxiv*, 2020. <https://www.medrxiv.org/content/10.1101/2020.02.06.20020974v1>
21. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS ONE*, 2020, 15(3): e0230405. <https://doi.org/10.1371/journal.pone.0230405> PMID: 32231374
22. Rajaraman S, Siegelman J, Alderson PO, Folio LS, Folio LR, Antani SK. Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays. *IEEE Access*, 2020, 8.
23. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, 2020, 43, 635–640. <https://doi.org/10.1007/s13246-020-00865-4> PMID: 32524445
24. Butt C, Gill J, Chun D, Babu BA. Deep learning system to screen coronavirus disease 2019 pneumonia. *Applied Intelligence*, 2020(5).
25. Hemdan ED, Shouman MA, Karar ME. COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images. *arXiv:2003.11055*, 2020. <https://arxiv.org/abs/2003.11055>
26. Ucar F, Korkmaz D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Medical Hypotheses*, 2020, 140.
27. Barstugan M, Ozkaya U, Ozturk S. Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods. *arXiv:2003.09424*, 2020. <https://arxiv.org/abs/2003.09424>.
28. Randhawa GS, Soltysiak MPM, El Roz H, de Souza CPE, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS ONE* 15(4): e0232391. <https://doi.org/10.1371/journal.pone.0232391> PMID: 32330208
29. Chan JF, Yuan SF, Kok KH, To KK, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 2020, 395: 514–523.
30. Li JY, You Z, Wang Q, Zhou ZJ, Qiu Y, Luo R, et al. The epidemic of 2019-novel-coronavirus (2019-nCoV) pneumonia and insights for emerging infectious diseases in the future. *Microbes and Infection*, 2020; 22(2):80–85. <https://doi.org/10.1016/j.micinf.2020.02.002> PMID: 32087334
31. Xu YH, Dong JH, An WM, Lv XY, Yin XP, Zhang ZJ, et al. Clinical and computed tomographic imaging features of Novel Coronavirus Pneumonia caused by SARS-CoV-2. *Journal of Infection*, 2020; 80 (4):394–400. <https://doi.org/10.1016/j.jinf.2020.02.017> PMID: 32109443
32. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural computation*, 1997, 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735> PMID: 9377276
33. Pedrycz W, Song M. A granulation of linguistic information in AHP decision-making problems. *Information Fusion*, 2014, 17:93–101.
34. Wuhan Municipal Health Commission. <http://wjw.wh.gov.cn>
35. Coronavirus disease (COVID-2019) situation reports from the World Health Organization(WHO). <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>.