An Empirical Mode Decomposition Fuzzy Forecast Model for COVID-19

Bo-Lun Chen[1,2] · Yi-Yun Shen[1] · Guo-Chang Zhu[1] · Yong-Tao Yu[1] · Min Ji[1]

Abstract

At present, the Corona Virus Disease 2019 (COVID-19) is ravaging the world, bringing great impact on people's life safety and health as well as the healthy development of economy and society, so the research on the prediction of the development trend of the epidemic is crucial. In this paper, we focus on the prevention and control of epidemic using the relevant technologies in the field of artificial intelligence and signal analysis. With the unknown principle of epidemic transmission, we first smooth out the complex and variable epidemic data through the empirical mode decomposition model to obtain the change trends of epidemic data at different time scales. On this basis, the change trends under different time scales are trained using an extreme learning machine to obtain the corresponding prediction values, and finally the epidemic prediction results are obtained by fitting through Adaptive Network-based Fuzzy Inference System. The experimental results show that the algorithm has good learning ability, especially in the prediction of time-series sequences can guarantee the accuracy rate while having low time complexity. Therefore, this paper not only plays a theoretical support for epidemic prevention and control, but also plays an important role in the construction of public emergency health system in the long run.

Yi-Yun Shen, Guo-Chang Zhu, Yong-Tao Yu and Min Ji have contributed equally to this work

---

✉ Yi-Yun Shen
   yiyun_shen@163.com

   Bo-Lun Chen
   chenbolun1986@163.com

   Guo-Chang Zhu
   zhuguochang1996@163.com

   Yong-Tao Yu
   yuyongtao@hyit.edu.cn

   Min Ji
   jimin@hyit.edu.cn

[1] Department of Computer Science, Huaiyin Institute of Technology, Huaiyin 223003, Jiangsu, China

[2] Institute of Informatics, University of Zurich, 8050 Zurich, Switzerland

1 Introduction

At present, the COVID-19 is ravaging the world, bringing great impact to the people's life safety and health as well as the healthy development of the economy and society of each country. At the same time, the prevention and control of the epidemic is serious, complex and arduous, which also puts the social governance system and governance capacity of each country to a great test. In order to win this battle as early as possible, all relevant departments need to prevent the disease before it happens and minimize the damage. With the continuous development of new information technologies such as big data, cloud computing, artificial intelligence, and the Internet of Things, we should use better means to prevent and combat epidemics.

With the popularity of modern network communication media and the wide application of new media technologies, the Internet has gradually become an important position for the generation, fermentation and dissemination of social opinion. Especially in major public health events, network public opinion not only runs through the outbreak, development and extinction of major public health events, but also plays a significant influence in the development of major public health events. The prevention and control of the new crown epidemic is an unprecedented battle in world history, and its network public opinion is far hotter, more intense, more sensitive and more diffuse than ever before. This major public health event is not only a test for the public health emergency management system, but also a new challenge to the epidemic management capacity of each country. With the rise and practicality of artificial intelligence technology, artificial intelligence and big data technology have been widely used in the fields of epidemic prevention and control [1, 2], social network analysis [3, 4], computer vision [5, 6], and biomedicine [7, 8]. Estimating the basic regeneration number through mathematical model is helpful to determine the possibility and severity of epidemic outbreak, and provides key information for determining the type and intensity of disease intervention.

Therefore, it is important to develop realistic mathematical models to predict short-term and long-term epidemics. For example, the model can effectively predict the number of population in the peak and final infection or incubation periods of an epidemic, which provides important reference values for the epidemic prevention department to solve the problems of how many wards, how many medical personnel, and how to allocate medical supplies in the future. Similarly, by using the prediction model, we can get the number of infected people in different cities at a certain period of the outbreak. When the number of infections spread by an epidemic is greater than or equal to a certain threshold, the disease may spread and grow explosively in the city. Then, the epidemic prevention department can take some specific epidemic prevention measures according to the actual situation of the city. For example, it is important for users in key communities to reduce the risk of infection by reducing travel, wearing masks at all times, and purchasing disinfectants and hand sanitizers, among other behaviors. Another example is that the state can adopt a closed city epidemic prevention strategy to prevent the spread of the epidemic while the epidemic is still growing explosively outside the country. Tang et al. used mathematical knowledge to help verify the important role of the city closure strategy in slowing down the national epidemic [9]. A team from Oxford University, Harvard Medical School, and Boston Children's Hospital published a paper in Science in which they analyzed the correlation between the spread of online information and the epidemic based on data from China's Baidu Map migration big data platform, providing strong data support for studying the spread of the epidemic [10].

Therefore, it is of great theoretical importance to analyze, understand and predict the topology, function and dynamic behavior of the network through artificial intelligence and big data technology in epidemic prevention and control, which provides solutions for government departments to carry out epidemic prevention and control, information traceability, etc., and provides strong theoretical support and realistic guarantee for the security and stability of society and economic development.

In summary, our work has the following contributions:

(1) The complex and changeable epidemic data are stabilized by empirical mode decomposition in signal analysis, and the change trend of the epidemic data under different time scales is obtained.
(2) According to the change trend under different time scales, the extreme learning machine is used for training. The weight of hidden layer nodes is given randomly and does not need to be updated. It has advantages in learning rate and generalization ability.
(3) Using adaptive Network-based fuzzy inference system for data fitting and self-adjusting fuzzy inference control rules through off-line training and on-line learning algorithms not only has certain advantages in prediction accuracy, but also has low time complexity.

2 Relate Work

2.1 Research Status of Relevant Models

Yu et al. used Susceptible regeneration number, same-day infection rate and latent infection rate to obtain the evolutionary dynamics equation of virus, and modified the SIR (Susceptible-Infected-Recovered) model [11]. By analyzing the data of the development of COVID-19, Fan et al. proposed an analysis method based on stage division and model fitting for each stage. Based on the traditional SIR model, the concept of "undiagnosed infected person", "cured person" and "dead person" was introduced to obtain the phased SIR-F model [12]. Based on the SIR model, Sedaghat et al. added the infected population into the SIR model by using Weibull distribution function and proposed a Weibull-SIRD method [13]. Sedaghat et al. proposed the SEIR-PAD model in view of the SIR model being insufficient to fully adapt to the multiple clinical data of the new coronavirus, an infectious disease. The model is composed of 7 sets of ordinary differential equations with 8 unknown coefficients, which were solved numerically using an optimization algorithm to fit the clinical data of 4 sets of COVID-19 patients consisting of cumulative infected, dead, recovered, and susceptible populations [14].

Chen et al. proposed a SEIR (Susceptible Exposed Infectious Recovered) epidemic model based on the study of the spread of COVID-19 in China. The model divided the confirmed population into two groups: those who had been admitted and those who had not been admitted. Firstly, analyzing the key problems of the model such as disease-free equilibrium point and its stability, basic regeneration number. Then, the development of the epidemic was analyzed and predicted based on the actual data [15]. Lin et al. proposed a new SEIR model based on the traditional SEIR model by redefining latent and infected persons and adding the concept of invisible spreader. After testing, this model has obvious advantages over traditional models in COVID-19 epidemic fitting and prediction [16]. Cao et al. proposed the SEIR infectious disease dynamics model for COVID-19. The key of this method is that the dynamic parameters of SEIR model are obtained by fitting actual data, and then euler numerical method is used to solve the modified SEIR infectious disease dynamic model

[17]. Li et al. proposed the SEIR model of dynamic infection rate, which first used the SEIR dynamics model to fit the R_0 curve of COVID-19 regeneration coefficient, and then completed effective prediction of epidemic data based on the dynamic R_0 [18]. Mahmoud et al. revised SEIR and Logistic mathematical models to To predict the development trend of the epidemic. In order to study the influence of different parameters on model prediction, sensitivity analysis was carried out [19]. Li et al. considered isolation group and multiple time-varying parameters, introduced abc-smc algorithm to solve the problem of parameter reasoning, and proposed SEIQR model [20].

Bai et al. constructed a non-autonomous dynamic model, taking into account the transmission mechanism of COVID-19 and public interventions such as tracing and isolating close contacts of infected people and treating them. In this method, the least square method and MCMC (Markov-Chain-Monte-Carlo) were used to calculate the parameters to be estimated in the dynamics model, and then the estimated parameters were used to predict the epidemic trend [21]. Friji et al. proposed a generalized mechanism model with eight states to describe the evolution of COVID-19 from an susceptible state to a discharged state as it passed through isolation and hospitalization. The question has three observational inputs: number of infections, number of deaths, and number of reported cases. The objective function of the model was weighted within the training days to guide the fitting algorithm to approach the latest pandemic period, making trend prediction more accurate and forecasting ability stronger [22].

Mustafa et al. analyzed the time series by box Jenkins method and proposed the ARIMA (autorepressant integrated moving average) model to predict the number of coronavirus infections in Iraq in 2019. The results show that the ARIMA(2,1,5) model was a better prediction model [23]. Using available data on COVID-19 patients in India, Singh et al. used the SIRD (Susceptible-Infected-Recovered-Dead) model to predict the correlation between outbreak intensity and time by generating the best parameter values and taking into account the number of deaths caused by the pandemic [24].

2.2 Traditional Prediction Algorithm

Cihan et al. used a fuzzy rules-based system to predict the daily number of COVID-19 cases in response to the global spread of COVID-19 [25]. Kurniawan et al. applied K-means clustering and correlation calculation to the COVID-19 data set. Among them, K-means was used to cluster unknown clusters of COVID-19 infected countries; Correlation calculations were used to calculate the correlation between the number of deaths and the attributes of critically ill patients to predict the number of deaths in different ethnic groups [26]. In order to prove that the machine learning model can effectively predict the number of infected people, Rustam et al. Compared and verified four standard prediction models such as linear expression [27].

Shaikh et al. used linear regression and polynomial regression, using R2 and error values to evaluate, to determine the optimal regression model and to predict the Novel Coronavirus in India [28]. Sharma et al. Decompose the time series data into different subsets through eigenvalue decomposition of Hankel matrix technology, take them as the input of ARIMA prediction model, and finally take the superimposed prediction value as the final result [29]. KKumar et al. modeled the evolution of the new crown pneumonia epidemic, using ARIMA and Prophet time series models to predict, and used average absolute error, root mean square error, relative square error and average absolute percentage error to evaluate the effectiveness

of the model. The investigation shows that the ARIMA model has a good effect in predicting the prevalence of COVID-19 [30].

Gaglione et al. applied the algorithms to the tracking and target prediction of missiles and ships to epidemiology COVID-19, that is, Bayesian sequence and adaptive dynamic estimation, which can reliably estimate and predict the evolution of infection [31]. The EN-COF (elastic net covid-19 predictor) proposed by Johnsen et al. aimed to provide an intuitive, universal and easy approach to apply predictor. The predictor novel coronavirus pneumonia was a multilinear regression trained by temporal data and can be used to predict the number of new crown pneumonia cases per day. Compared with BiLSTM model(bidirectional long short term memory), the EN-COF maintained high accuracy and had the advantages of transparency and generalization [32].

2.3 Prediction Algorithm Based on Deep Learning

In view of the problem that the virus infection rate in SEIR model is constant and cannot be accurately modeled for epidemic trend prediction, Gan et al. proposed a method combining the prediction method of virus infection rate of the LSTM (Long short-term Memory Network) with the SEIR model then constructed the COVID-19 prediction model LS-NET (LSTM-SEIR Network) [33]. On the basis of RNN (Recurrent Neural Network), Iqbal et al. proposed the LSTM model. They used RNN and LSTM respectively to train the data of patients in Pakistan, then calculated the mean absolute percentage error to determine the prediction effect of the model on different LSTM units, batch sizes and time points [34]. After comparing four different models: LSTM network, box Jenkins method, prophet method, and Holt winters addition method with damping trend, helli et al. concluded that adding death data to the input would lead to lower prediction error than using only the total number of cases as the input. The Holt winters addition method with damping trend was better than the LSTM network in predicting the total number of COVID-19 cases [35].

Xie et al. replaced the basic infectious number R_0 with the combined dynamic transmission rate and proposed a nonlinear time-varying transmission rate model based on support vector regression. This method effectively solved the difficulties in accurately estimating the basic infectious number R_0 and the low prediction accuracy of single model in traditional epidemic epidemiology [36]. Gupta et al. compared the classification methods of random forest, linear model, support vector machine, decision tree and neural network, and pointed out that random forest model was superior to other methods. Then they used random forest for prediction and analysis, and used K-fold cross-validation to measure the consistency of the model [37]. Shahid et al. used predictive models such as ARIMA, SVR (Support Vector Regression), LSTM, and BI-LSTM to predict the time series of confirmed cases, deaths, and recovery in 10 countries primarily affected by COVID-19. The performance of the model was measured by mean absolute error, root mean square error, and r^2 index. The best-performing to lowest-performing models were BI-LSTM, LSTM, GRU, SVR, and ARIMA [38]. When masum et al. predicted the daily cumulative confirmed cases, aiming at the randomness factors of neural network optimization and random initialization of weights, and considering the poor repeatability of the results produced by the LSTM model, they proposed the r-LSTM framework, and used the Z-score outliers to detect the repeatability and robustness of the framework [39]. Based on the ARIMA model and LSTM-based recursive neural network, Assimakis et al. studied the applicability of Kalman filter as decision support in COVID-19 and other infectious disease early warning and emergency response systems, and predicted cumulative daily confirmed cases [40].

Beche et al. attempted to solve the problem of predicting novel coronavirus transmission by exploiting the ability of cyclic autoencoders on time series and semi-supervised training processes. And the cumulative number of confirmed cases in any country was estimated by introducing the concept of neighboring countries. The results showed that this method can reliably predict the period of 30 days data [41]. Mousavi et al. proposed a new COVID-19 confirmed case prediction system platform based on transmission rate, temperature, humidity, and other factors. The platform can systematically derive a set of features suitable for training RNN, and improve the prediction of the trend of confirmed cases by using the stationary and non-stationary features [42].

As discussed, although much effort has been dedicated to predict the development trend of the epidemic, these algorithms are subject to the following limitations and challenges.

(1) The epidemic data are continuous and time-series data, and the change law is nonlinear and non-stationary. Some traditional methods need complex modeling and analysis, and optimize the model, which has high time complexity.
(2) The traditional algorithm has achieved good results in the dataset with uniform rules. However, due to the rapid outbreak of the outbreak and the variation of the epidemic situation, the corresponding data change rule is different. The traditional method lacks the generalization ability of the fresh data outside the learning set with the same rule, it is difficult to learn the law behind the data, and the prediction effect is not very ideal.

3 Problem Modeling

3.1 EMD (Empirical Mode Decomposition)

EMD is a method to find out the internal characteristics of signals at different time scales by analyzing the change trend of signals, so as to obtain sub signals with different frequencies [43]. Compared with other signal analysis algorithms, this method does not need to set the basis function. Its essence is to smooth the signal first, find out the change trend of the signal under different characteristic time scales, and form the IMF (Intrinic Mode Function).

The decomposition process of the EMD algorithm is as follows: Firstly, finding all the maximum points of the original signal $x(t)$ and fitting into the upper envelope; second finding all the minimum points and fitting them into the lower envelope; then taking the mean of the two to fit the mean envelope and calling it $m(t)$;finally removing $m(t)$ from $x(t)$ and getting a new signal $h(t)$:

$$h(t) = x(t) - m(t) \tag{1}$$

The conditions of the IMF are no negative local maximum and positive local minimum. If not, $x(t)$ is replaced by $h(t)$, and the above steps are repeated until $h(t)$ meets the conditions of the IMF. If so, the signal $h(t)$ is the first IMF. Repeat the above steps for this IMF to obtain more IMF until the N-order IMF component or residual component is less than the preset value or monotone function or constant, the EMD decomposition ends. Figure 1 is a schematic diagram of the EMD.

3.2 ELM (Extreme Learning Machine)

Limit learning machine is a single hidden layer feedforward neural network in which the connection weight of input layer and hidden layer and the threshold of hidden layer can
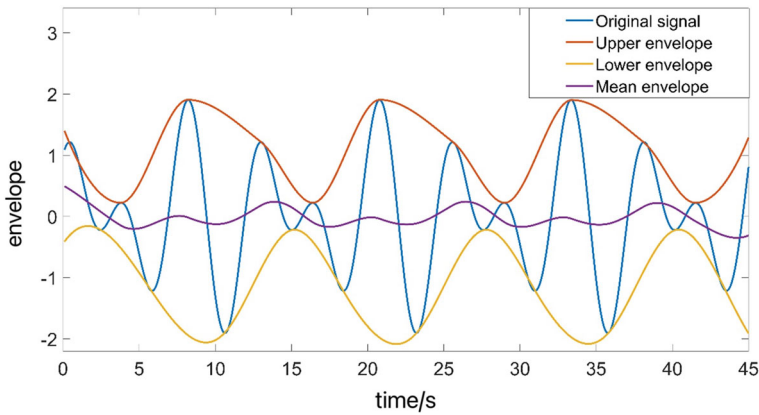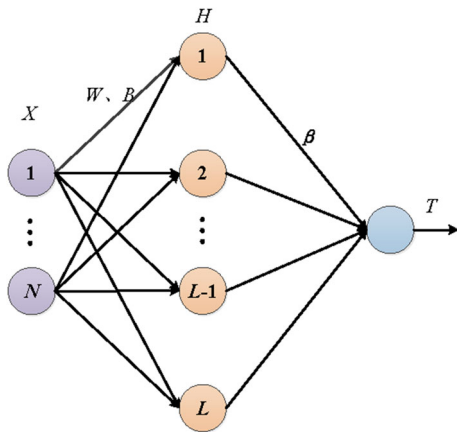
Fig. 1 Schematic diagram of EMD process

Fig. 2 Schematic diagram of ELM

be set randomly and do not need to be adjusted [44]. The connection weights between the hidden layer and the output layer can be solved by solving the equations, and the optimal solution can be obtained at one time, so as to reduce the amount of computation and improve the running speed. Compared with the traditional neural network, this method simplifies the training process, improves the speed and generalization ability of the algorithm, and solves the problems of over-fitting and local minimization. Figure 2 is the schematic diagram of the ELM.

3.3 ANFIS (Adaptive Network-Based Fuzzy Inference System)

ANFIS based network is a neural fuzzy system proposed by Jang [45]. The ANFIS structure consists of five layers. Figure 3 shows the ANFIS system structure with two inputs and one output. Assuming ANFIS has two inputs , and one output y, the rule library has the following rules:

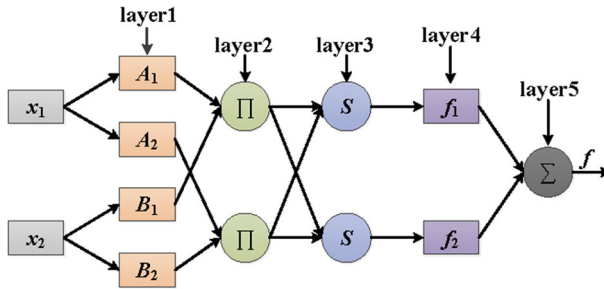(1) If x_1 is A_1 and x_2 is B_1, then $y = p_1x_1 + q_1x_2 + r_1$;

Fig. 3 Schematic diagram of the network structure of ANFIS

(2) If x_1 is A_2 and x_2 is B_2, then $y = p_2x_1 + q_2x_2 + r_2$;

Among them, A_1 and A_2 are fuzzy sets, p_i, q_i and r_i are result parameters, also known as subsequent parameters.

The first layer is the fuzzy layer. The membership function is used to fuzzize the input variables and get the corresponding membership degree. Each node represents a membership function, which usually includes Gaussian membership function, triangular membership function, bell function, trapezoidal function, etc. Gaussian membership function is mainly used in this paper, and the formula is as follows:

$$\mu_{F_i^j}(x_i) = exp\left[-\frac{(x_i - c_{ij})^2}{\sigma_j^2}\right], \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m \qquad (2)$$

where $\mu_{F_i^j}$ is the jth membership function of x_i, c_{ij} is the center of the jth Gaussian membership function of x_i, σ_j is the width of the jth Gaussian function of x_i, n is the number of input variables, m is the number of membership functions, and also represents the total number of rules of the system.

The second layer is the regular intensity release layer. The algebraic product of all input signals is used as the output to obtain the excitation intensity of each rule. The third layer is the regular strength normalization layer. The trigger intensity of each rule obtained from the second layer is normalized to calculate the trigger proportion of the rule in the whole rule base. The fourth layer is the fuzzy rule output layer. In this layer, all nodes are adaptive nodes with node functions, and the output of each rule can be obtained. The fifth layer is the summation output layer, used to calculate the sum of all input signals to get the total output.

4 The Detailed Process of The Algorithm

4.1 Overall Architecture

We use the EMD-FFM (Empirical Mode Decomposition Fuzzy Forecast Model) to predict the daily number of new COVID-19 diagnoses. First, the EMD was used to decompose the original data to obtain the changing trend of the number of newly confirmed cases at different time scales. Then, the ELM was used to make training prediction for the IMF. Finally, the ANFIS was used to fit the training results to obtain the final prediction value. The process
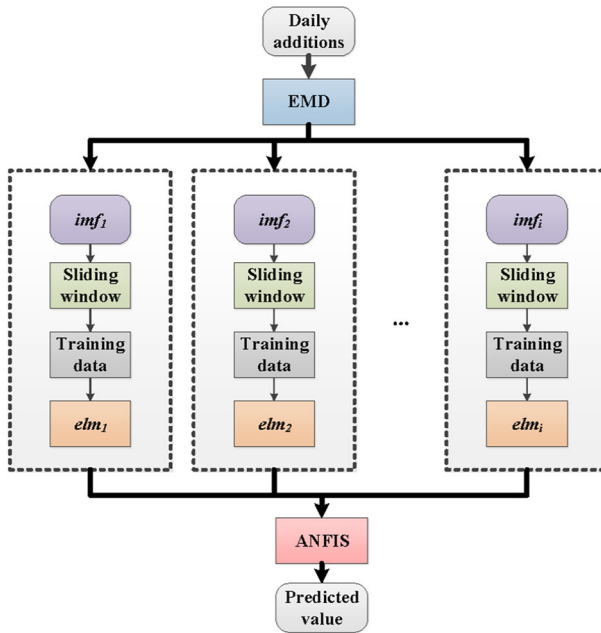
Fig. 4 EMD-FFM structure diagram

Table 1 Empirical mode decomposition fuzzy forecast model

Algorithm 1 Empirical Mode Decomposition Fuzzy Forecast Model

Input: COVID-19 data;

Output: Predicted value y of data from day $t + 1$ to day $t + c$

1: New=pretreat(COVID-19 data);

2: $IMF = (imf_1, imf_2, \ldots, imf_i, \ldots, imf_n)$=EMD($New$);

3: **for** i=1:n

4:    x_i=win(imf_i);

5:    elm_i=elm(x_i);

6: **End**

7: $ELM = (elm_1, elm_2, \ldots, elm_i, \ldots, elm_n)$;

8: y=ANFIS(ELM)

structure diagram is shown in Fig. 4 and the forecast flow chart of the Empirical Mode Decomposition based Fuzzy Prediction Model (EMD-FPM) is shown in the Table 1.

4.2 Detailed Steps of Algorithm

In summary, the process of signal decomposition fuzzy prediction algorithm for COVID-19 proposed in this paper is described as follows, taking the daily newly confirmed cases in the United States as an example:

The number of newly confirmed cases in the United States was normalized to obtain $New(t)$, where t is the number of days; EMD is used to decompose data $New(t)$, and the steps are as follows:

Firstly, all the maximum points in $New(t)$ are found, and the upper envelope $max(t)$ is obtained by function fitting. All minimum points were found, and the lower envelope $min(t)$ was obtained by fitting with the same method. The mean value of the upper and lower envelope is denoted as $m_i(t)$:

$$m_1(t) = \frac{max(t) + min(t)}{2} \tag{3}$$

Then, Subtract $m_1(t)$ from the original signal to get a new signal $h_1^1(t)$ with the low frequencies removed:

$$h_1^1(t) = New(t) - m_1(t) \tag{4}$$

As the original data $New(t)$ is disorganized and irregular, $h_1^1(t)$ obtained by the first calculation generally cannot meet the IMF conditions, and the above steps need to be repeated k times until $h_1^1(t)$ meets the IMF conditions, that is, the first-order IMF component of $New(t)$, denoted by $imf_1(t)$:

$$imf_1(t) = h_1^1(t) \tag{5}$$

Subtract $imf_1(t)$ from $New(t)$ and get a new signal $r_1(t)$ without high frequency component:

$$r_1(t) = New(t) - imf_1 \tag{6}$$

The above process is repeated for $r_1(t)$, and the second order IMF component is obtained. And so on, until the n-order IMF component $imf_n(t)$ residual component is less than the preset value or monotone function or constant, and the decomposition stops. Getting IMF after n decomposition of $New(t)$, $IMF = (imf_1, imf_2, \dots, imf_n)$:

For imf_i, a window with a size of $1*(c+1)$ is used to slide with a step size of one. The first c value of each sliding window is included in ELM training set X, and the $c+1$ value is composed of test set T. The first 70% of training set X is input and the last 30% is output.

The input and output obtained in Step 3 were trained using the ELM, and elm_i was obtained, then the corresponding $ELM = (elm_1, elm_2, \dots, elm_n)$:;

Assume that the training set X has n samples and the network has L hidden layer nodes.

For a single hidden layer neural network, input x_i, corresponding expected output is t_i, activation function is $g(x)$, input weight w_i and bias b_i from input layer to hidden layer, and output weight β_i from hidden layer can be obtained:

$$\sum_{i=1}^{L} g(x_i w_i + b_i)\beta_i = t_i \tag{7}$$

The matrix can be expressed as:

$$H\beta = T \tag{8}$$

where H is the hidden layer output matrix, β is the output weight, and T is the desired output:

$$H = \begin{bmatrix} g(x_1 w_1 + b_1), \dots, g(x_1 w_L + b_L) \\ \cdots \\ g(x_N w_1 + b_1), \dots, g(x_N w_L + b_L) \end{bmatrix}_{N \times L} \tag{9}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \cdots \\ \beta_L^T \end{bmatrix}_{L \times m} \tag{10}$$

$$T = \begin{bmatrix} T_1^T \\ \cdots \\ T_L^T \end{bmatrix}_{N \times m} \tag{11}$$

Since input weight w_i and bias b_i are randomly determined and will not change in the ELM, which means that the output matrix H of the hidden layer is uniquely determined, the training process of the ELM can be transformed into solving the output weight β, that is:

$$\beta = H^- T \tag{12}$$

The resulting weight β is multiplied by the first 70% of the test set T to obtain $ELM = (elm_1, elm_2, \ldots, elm_n)$.

Use the ANFIS to fit the elm_i obtained in Step 4, then the output of the first layer of the ANFIS is:

$$O_{ij}^1 = \mu_{F_j^i}(e_i) \tag{13}$$

where μ_F is is the membership function determined by some parameters, known as the antecedent parameter.

The output of the second layer is as follows:

$$O_j^2 = \omega_j = \prod \mu_{F^i}(e_i) \tag{14}$$

The output of the third layer is as follows:

$$O_j^3 = \overline{\omega_j} = \frac{\omega_j}{\sum \omega_j} \tag{15}$$

where $\overline{\omega_j}$ is the normalized excitation intensity transmitted from the upper layer.

The output of the fourth layer is as follows:

$$O_j^4 = \overline{\omega_j} f_j = \overline{\omega_j} \left(\sum p_j elm_j + q_j \right) \tag{16}$$

p_j and q_j are parameter sets of this node, which are called aftereffects.

The output of the fifth layer is as follows:

$$O^5 = \sum \overline{\omega_j} f_j = \frac{\sum (\omega_j f_j)}{\sum \omega_j} \tag{17}$$

In each iterative training, the root mean square error between the actual output and the expected output will be reduced. When the termination condition of training times is met or the error range is within a certain threshold, the training will be stopped. Through learning, the optimal parameters of membership function can be calculated effectively, so that the fuzzy reasoning system can best simulate the actual relationship between the input and output, and finally obtain the prediction result y.

5 Experimental Results and Analysis

5.1 Dataset

In order to verify the performance of the algorithm, the experimental data in this paper comes from the public data set provided by the center for systems science and Engineering (CSSE) of Johns Hopkins University. The data set takes days as the time unit and is classified by countries (https://systems.jhu.edu/). The data set has strong accuracy and scientificity, which is of great significance to objectively evaluate the accuracy of prediction algorithm. In this paper, the number of newly confirmed coronavirus infections per day in the US, UK, Italy, France, Germany and Iran was predicted. The selected countries are distributed in the Americas, Europe, the Middle East and other parts of the world, which are very representative.

5.2 Evaluation Criterion

In order to compare the algorithms, our evaluation indexes are as follows:

(1) MSE (Mean Square Error)
To calculate the error between the predicted value and the real value, the formula is as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{18}$$

(2) MAE (Mean Absolute Error)
To reflect the actual error between the predicted value and true value, the formula is as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{19}$$

(3) SD (Standard Deviation)
To reflect the degree of dispersion between the predicted value or true value and its mean. Taking the true value as an example, the formula is as follows:

$$SD = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} \tag{20}$$

where $\bar{y}$ is the predicted value, y is the true value, $\bar{y}$ is the average value of the true value, and n is the number of data.
In order to intuitively show the prediction ability of the model and the correlation between the predicted value and the real value, the following indicators are selected for evaluation in this paper:

(4) Cosine similarity
The cosine of the Angle between two vectors in the vector space is used to measure the difference between two individuals. The range of cosine value is between $[-1,1]$, and the closer it is to 1, the closer the included Angle is to 0 degree, that is, the more similar the two vectors are. The closer you get to -1, the more opposite the two vectors are; It's close to zero, which means the two vectors are nearly orthogonal. The formula is as follows:

The cosine of the angle between two vectors in the vector space is used to measure the difference between two individuals. Therefore, the prediction effect can be measured by calculating the similarity between the predicted value vector $X = (x_1, x_2, \ldots, x_n)$ and the real value vector $Y = (y_1, y_2, \ldots, y_n)$. The smaller the included angle, the more similar it represents. The range of the cosine value is between $[-1, 1]$. The closer it is to 1, the closer the included angle is to 0 degrees, that is, the more similar the two vectors are; The closer to -1, the more opposite the direction of the two vectors. The cosine value is close to 0 means that the two vectors are almost orthogonal. The calculation formula is as follows:

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{21}$$

(5) Pearson's correlation coefficient

Similarly, Pearson correlation coefficient is also an index used to measure the correlation between two vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, and its value is between -1 and 1. When the value is 1, it indicates that the two variables are completely positively correlated; When it is -1, it indicates that the two variables are completely negatively correlated. A value of 0 indicates that there is no linear relationship between the two variables. The formula is as follows:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\left(\sum_{i=1}^{n}(x_i - \overline{X})^2\right)\left(\sum_{i=1}^{n}(y_i - \overline{Y})^2\right)} \tag{22}$$

(6) Coefficient of determination

It is generally used in regression models to evaluate the degree of agreement between the predicted value and actual value. The closer it is to 1, the better the independent variable explains the dependent variable. The formula is as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}} \tag{23}$$

where SS_{tot} is the sum of total squares, SS_{reg} is the sum of regression squares, and SS_{res} is the sum of residual squares.

5.3 Results and Analysis

In this experiment, three groups of comparative experiments were selected, including ELM, BP, and NAR. BP neural network is a hierarchical neural network with multiple layers [46], including two processes of signal forward propagation and error back propagation. Forward propagation process, data from the input layer to the hidden layer to the output layer. If the actual output is inconsistent with the expected output, the error backpropagation process is carried out, and the error is backpropagated from the output layer to the hidden layer and then to the input layer, and the error is apportioned to all units of each layer to correct the weight of the previous layer. After repeated learning and training, the weights and thresholds corresponding to the minimum error are determined. NAR neural network [47] is a neural network with timing properties, which uses itself as a regression variable, that is, a linear combination of random variables at some time to describe the nonlinear target time.
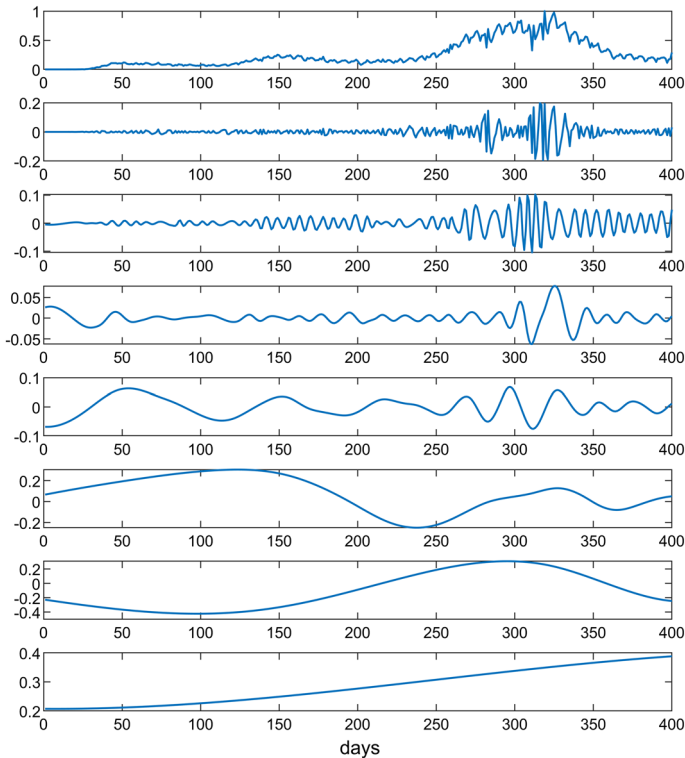
**Fig. 5** Schematic diagram of EMD:take US as an example

5.3.1 Signal Decomposition Result

Taking the United States as an example, Fig. 5 shows the EMD decomposition process of the daily newly confirmed COVID-19 cases, in which the first one is the original normalized data. The rest are 7 IMFs decomposed by the EMD.

In Fig. 5, the abscissa represents the time series, and the ordinate represents the value of the normalized daily newly diagnosed patients after EMD decomposition. It is obvious from the figure that the original data is irregular and non-linear. Through multiple decomposition, the original uneven and smooth data gradually become a smooth curve. As can be seen from the previous three decomposition, the data are relatively stable at the beginning, and the fluctuation range will be more obvious in the later period. The last three times of decomposition can be seen as the overall fluctuation, the last time showed an upward trend, which is consistent with the overall trend of the original data. This also reflects the high efficiency of the EMD decomposition in dealing with nonlinear and non-stationary data at different time scales.

5.3.2 Short-Term Forecast

For daily additions, we use the data of the previous five days to predict the value of the sixth day, which is regarded as a short-term prediction. The comparison between the predicted value and the actual value is shown in Fig. 6.
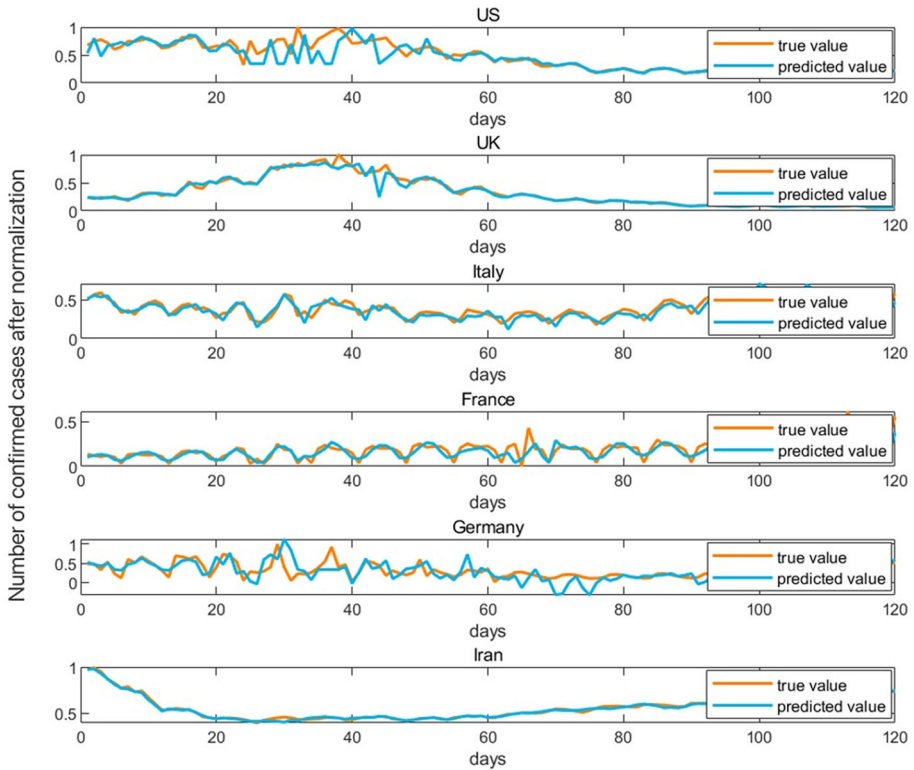
Fig. 6 Short-term forecast comparison of real and predicted values

Table 2 The results of short-term prediction under different similarity evaluation indicators

	US	UK	Italy	France	Germany	Iran
$\cos(\theta)$	0.9929	0.9995	0.9990	0.9979	0.9921	0.9999
ρ	0.7787	0.9830	0.8978	0.7411	0.7394	0.9927
R^2	0.5661	0.9655	0.7914	0.5132	0.4406	0.9845

From Fig. 6, we can see that the real value is highly consistent with the predicted value. In the UK, for example, the true value peaked at about 40 days, followed by a downward trend, and the predicted value was the same as the trend and also reached its maximum at about 40 days. We calculated the correlation between the predicted value and the real value, and the results are shown in Table 2.

From Table 2, we can see that the index values of Britain and Italy are close to 1, which indicates that the predicted value is very accurate. The low correlation value between the United States and Germany indicates that the government and society may have taken intervention measures to suppress the spread of the epidemic, so the accuracy rate is low. However, the positive values indicate that the predicted value has the same trend as the real value and the prediction effect is good. On the whole, the short-term prediction effect is more accurate.

Next, we compare the EMD-FFM with BP, ELM, and NAR. The experimental results are as follows:
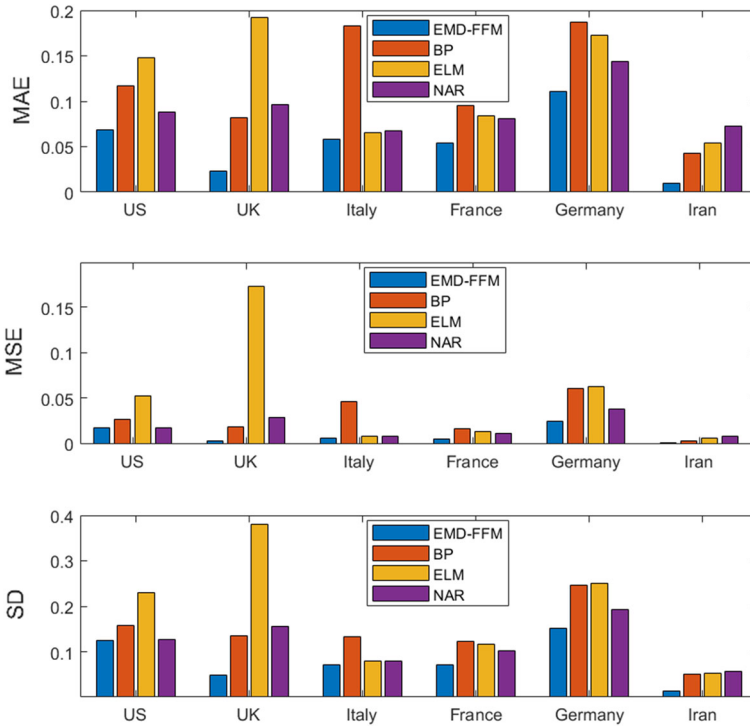
Fig. 7 Comparison of short-term prediction results of different algorithms

As can be seen from Fig. 7, the EMD-FFM algorithm has good accuracy and generalization performance in predicting the daily number of newly confirmed COVID-19 cases, and all three evaluation indexes are superior to the other algorithms. This is because the ANFIS model combines the advantages of adaptive network and fuzzy inference system, which can change system parameters according to the prior knowledge when fitting data and get output values closer to the real. Experiments show that the EMD-FFM algorithm not only has high accuracy, but also has strong applicability and wide application range.
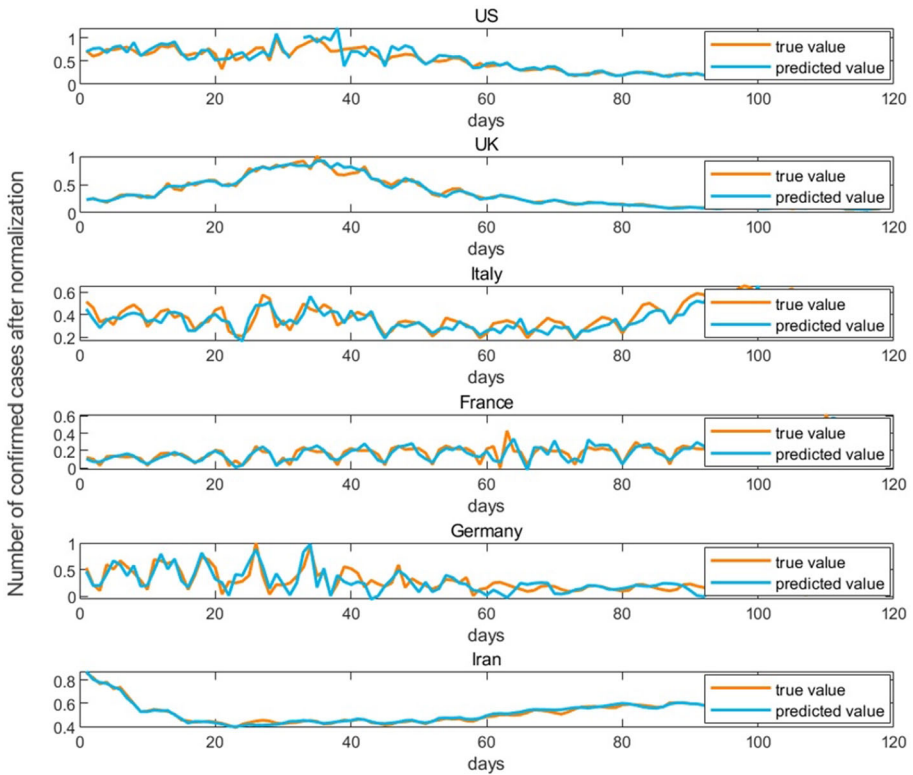
5.3.3 Long-Term Forecast

For the daily increase, we use the data of the previous fifteen days to predict the value of the sixteenth day, so as to simulate the long-term forecast. The comparison between the predicted value and the actual value is shown in Fig. 8.

From Fig. 8, we can see that the real value is highly consistent with the predicted value. For example, In France, the data fluctuation trend was consistent within 20-50 days and peaked around 63 days, followed by fluctuations. We calculated the correlation between the predicted value and the real value, and the results are shown in Table 3.

From Table 3, we can see that the evaluation indexes of Britain and Iran are still close to 1. The average data of the other countries are also around 0.8, which indicates that the predicted value not only has the same trend with the real value, but also has a high accuracy compared with the real value. Vaccinations, national policies and people's attention are still

Fig. 8 Long-term forecast comparison of real and predicted values

Table 3 The results of long-term prediction under different similarity evaluation indicators

	US	UK	Italy	France	Germany	Iran
$\cos(\theta)$	0.9958	0.9997	0.9985	0.9978	0.9919	0.9999
ρ	0.8608	0.9910	0.8402	0.7926	0.7398	0.9896
R^2	0.6881	0.9820	0.6767	0.5687	0.4702	0.9719

contributing to the error. But on the whole, the long-term prediction effect of this method is still good.

Next, we compare the EMD-FFM with BP, ELM and NAR, and the experimental results are as follows:

As shown in Fig. 9, consistent with the short-term prediction results, the three evaluation indexes of the EMD-FFM algorithm are still far smaller than other algorithms, which is due to the superiority of the ANFIS model in data fitting. Among the other three algorithms, the ELM performs better. Therefore, the EMD-FFM algorithm also has good performance in long-term prediction.
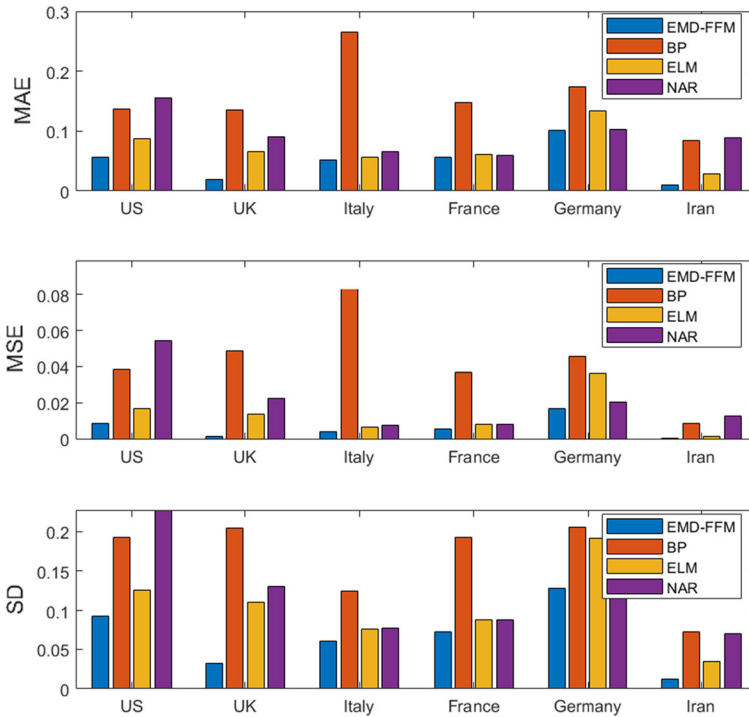
Fig. 9 Comparison of short-term prediction results of different algorithms

5.3.4 Time Complexity Analysis of Algorithm

Time complexity is an important index for designing prediction algorithm. In the experiment, we compare the calculation time of the EMD-FFM algorithm with BP, ELM and NAR. Figure 10 shows the comparison results of running time of different algorithms in different data sets.

It can be seen from the figure that the running time predicted by the EMD-FFM algorithm is higher than that predicted by the ELM directly, but lower than that of the BP and NAR algorithm. For example, in the short-term prediction, when predicting the epidemic data of the United States, the running speed of our algorithm is about 0.14 s, while that of the BP algorithm is about 0.24 s, which reduces the time by 42%. Similarly, in the long-term prediction, when predicting the epidemic data of Iran, the running speed of our algorithm is about 0.1 s, while the NAR algorithm is about 0.28 s, which only needs 35% of its time. Because the EMD-FFM algorithm trains the model through limit learning machine, the connection weights of input layer and hidden layer and the threshold of hidden layer of the model can be set randomly and do not need to be adjusted. The connection weight between the hidden layer and the output layer can obtain the optimal solution at one time by solving the equations, so as to reduce the amount of calculation and improve the running speed.
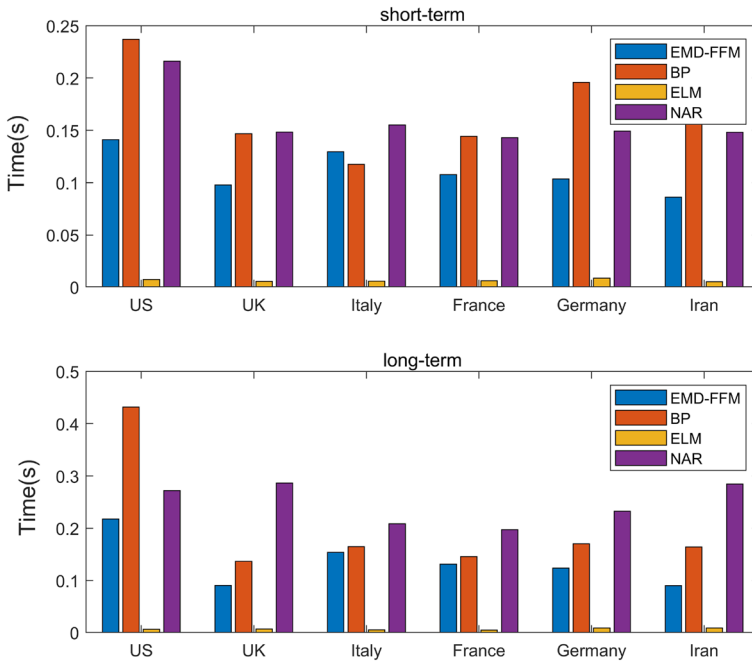
Fig. 10 Comparison of running time of different algorithms

6 Discussion

In this paper, we combine EMD and ELM to predict the epidemic data based on ANFIS. The advantage of this method is that the time-series data of the epidemic is considered as a signal sequence, which can be decomposed directly without any prior analysis and research, and it has obvious advantages in dealing with non-stationary and non-linear data, which is suitable for analyzing non-linear and non-stationary signal sequences with high signal-to-noise ratio and overcoming the problem of non-adaptive basis functions. In addition, the data obtained from the EMD is used to train the model by the ELM, which has the advantage that the hidden layer does not require iteration and has a fast learning rate and good generalization performance. Finally, the fuzzy inference part of the adaptive network realizes the combination of fuzzy logic inference and neural network. This structure form has the advantages of both easy expression of human knowledge in fuzzy logic and distributed information storage and learning capability of neural network, which can perform parallel data processing and improve the prediction efficiency of the algorithm, and has very good effect in solving nonlinear problems.

However, the method also has certain drawbacks, for example, when the epidemic outbreak grows, the different prevention strategies that each department will adopt to stop the spread of the epidemic will have an impact on the real-time data change pattern of the epidemic, and some outlier points and other situations will appear. When there is intermittency phenomenon caused by abnormal events in the signal, such as intermittent signal, impulse interference and noise, the decomposition result of EMD will have a modal mixing situation. In addition, the ELM model only considers the empirical risk and not the structured risk, so it may lead to the problem of overfitting, and by calculating the least squares solution in the model, the user

cannot fine-tune it according to the characteristics of the data set, and the controllability is poor, and the performance of the model will be greatly affected if there are outliers in the data set.

7 Conclusion

The COVID-19 is related to people's own health and the economic development of society, so prediction of the epidemic situation is very important. As an important algorithm for signal analysis, EMD can well analyze the change trend of epidemic data, stabilize the complex and changeable epidemic data, and obtain the change trend of epidemic data at different time scales. Using ELM training to obtain the corresponding predicted values for the changing trends at different time scales, and fitting with the ANFIS to obtain the final epidemic prediction results. The experimental results show that the signal decomposition fuzzy prediction model has good learning ability, and has good accuracy for various data predictions in the epidemic. In the future, we will analyze the abnormal conditions in the input signal in detail, reduce the impact of signal data instability caused by abnormal events on the accuracy and efficiency of the model, and design a robust algorithm model.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Ding W, Abdel-Basset M, Hawash H, Elkomy OM (2021) Mt-ncov-net: a multitask deep-learning framework for efficient diagnosis of covid-19 using tomography scans. IEEE Trans Cybern
2. Cao J, Jiang X, Zhao B et al (2020) Mathematical modeling and epidemic prediction of covid-19 and its significance to epidemic prevention and control measures. J Biomed Res Innov 1(1):1–19
3. Guo J, Li Y, Wu W (2019) Targeted protection maximization in social networks. IEEE Trans Netw Sci Eng 7(3):1645–1655
4. Ding W, Pedrycz W, Triguero I, Cao Z, Lin C-T (2020) Multigranulation supertrust model for attribute reduction. IEEE Trans Fuzzy Syst 29(6):1395–1408
5. Zhang G, Ge Y, Dong Z, Wang H, Zheng Y, Chen S (2021) Deep high-resolution representation learning for cross-resolution person re-identification. IEEE Trans Image Process 30:8913–8925
6. Chen B-L, Wan J-J, Chen T-Y, Yu Y-T, Ji M (2021) A self-attention based faster r-cnn for polyp detection from colonoscopy images. Biomed Signal Process Control 70:103019
7. Ding W, Lin C-T, Cao Z (2018) Shared nearest-neighbor quantum game-based attribute reduction with hierarchical coevolutionary spark and its application in consistent segmentation of neonatal cerebral cortical surfaces. IEEE Trans Neural Netw Learn Syst 30(7):2013–2027
8. Ding W, Abdel-Basset M, Hawash H, Pedrycz W (2021) Multimodal infant brain segmentation by fuzzy-informed deep learning. IEEE Trans Fuzzy Syst
9. Tang B, Wang X, Li Q, Bragazzi NL, Tang S, Xiao Y, Wu J (2020) Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions. J Clin Med 9(2):462

10. Hale T, Angrist N, Goldszmidt R, Kira B, Petherick A, Phillips T, Webster S, Cameron-Blake E, Hallas L, Majumdar S et al (2021) A global panel database of pandemic policies (oxford covid-19 government response tracker). Nat Hum Behav 5(4):529–538

11. Yanzhe L, Bingxiang L (2020) Evaluation and prediction of covid-19 based on time-varying sir model. In: International conference on web information systems and applications. Springer, pp 176–183

12. Yourong Fan, Yang Tao KH (2020) Assessment and prediction of covid-19 based on staged sir-f model. Comput Appl Softw 37(11):51–56

13. Mosavi A, Sedaghat A (2020) Predicting covid-19 (coronavirus disease) outbreak dynamics using sir-based models: comparative analysis of SIRD and Weibull-SIRD. Available at SSRN 3739532

14. Sedaghat A, Band S, Mosavi A, Nadai L (2020) Covid-19 (coronavirus disease) outbreak prediction using a susceptible-exposed-symptomatic infected-recovered-super spreaders-asymptomatic infected-deceased-critical (SEIR-PADC) dynamic model. In: 2020 IEEE 3rd international conference and workshop in Óbuda on electrical and power engineering (CANDO-EPE). IEEE, pp 000275–000282

15. Xingzhi C (2021) Evaluation and prediction of prevention and control effects of the covid-19 epidemic based on the SEIR model. Appl Math Mech 199–211

16. Lin J-f (2020) Assessment and prediction of covid-19 based on SEIR model with undiscovered people. J Univ Elec Sci Tech China 49:375–382

17. Cao S, Feng P, Shi P (2020) Study on the epidemic development of covid-19 in Hubei province by a modified SEIR model. J Zhejiang Univ (Med Sci) 49(2):178–184

18. Li W, Rong D, Shudong C (2020) Analysis of transmission characteristics of covid-19 and prediction of the development trend of epidemic situation. J Xiamen Univ (Nat Sci) 59(6):1025–1033

19. Mahmoud TM, Abu-Tafesh MS, ElOcla NM, Mohamed AS (2020) Forecasting of covid-19 in Egypt and Oman using modified SEIR and logistic growth models. In: 2020 2nd novel intelligent and leading emerging sciences conference (NILES). IEEE, pp 606–611

20. Li B, Liu Z, Zhao C, Sun Y (2020) A susceptible-exposed-infected-quarantined-recovered (SEIQR) model for predicting the trajectory of the covid-19 epidemic. In: 2020 Chinese automation congress (CAC). IEEE, pp 58–63

21. Ning B, Song C, Xu R (2021) Predicting the trend and evaluating control strategies of covid-19 based on a dynamical model. Acta Math Appl Sin 43(3):483–493

22. Friji H, Hamadi R, Ghazzai H, Besbes H, Massoud Y (2021) A generalized mechanistic model for assessing and forecasting the spread of the covid-19 pandemic. IEEE Access 9:13266–13285

23. Mustafa HI, Fareed NY (2020) Covid-19 cases in Iraq; forecasting incidents using box-Jenkins Arima model. In: 2020 2nd Al-Noor international conference for science and technology (NICST). IEEE, pp 22–26

24. Singh S, Raj P, Kumar R, Chaujar R (2020) Prediction and forecast for covid-19 outbreak in India based on enhanced epidemiological models. In: 2020 second international conference on inventive research in computing applications (ICIRCA). IEEE, pp 93–97

25. Cihan P (2020) Fuzzy rule-based system for predicting daily case in covid-19 outbreak. In: 2020 4th international symposium on multidisciplinary studies and innovative technologies (ISMSIT). IEEE, pp 1–4

26. Kurniawan R, Abdullah SNHS, Lestari F, Nazri MZA, Mujahidin A, Adnan N (2020) Clustering and correlation methods for predicting coronavirus covid-19 risk analysis in pandemic countries. In: 2020 8th international conference on cyber and IT service management (CITSM). IEEE, pp 1–5

27. Rustam F, Reshi AA, Mehmood A, Ullah S, On B-W, Aslam W, Choi GS (2020) Covid-19 future forecasting using supervised machine learning models. IEEE Access 8:101489–101499

28. Shaikh S, Gala J, Jain A, Advani S, Jaidhara S, Edinburgh MR (2021) Analysis and prediction of covid-19 using regression models and time series forecasting. In: 2021 11th international conference on cloud computing, data science & engineering (confluence). IEEE, pp 989–995

29. Sharma RR, Kumar M, Maheshwari S, Ray KP (2020) Evdhm-Arima-based time series forecasting model and its application for covid-19 cases. IEEE Trans Instrum Meas 70:1–10

30. Kumar N, Susan S (2020) Covid-19 pandemic prediction using time series forecasting models. In: 2020 11th international conference on computing, communication and networking technologies (ICCCNT). IEEE, pp 1–7

31. Gaglione D, Braca P, Millefiori LM, Soldi G, Forti N, Marano S, Willett PK, Pattipati KR (2020) Adaptive Bayesian learning and forecasting of epidemic evolution-data analysis of the covid-19 outbreak. IEEE Access 8:175244–175264

32. Johnsen TK, Gao JZ (2020) Elastic net to forecast covid-19 cases. In: 2020 international conference on innovation and intelligence for informatics, computing and technologies (3ICT). IEEE, pp. 1–6

33. Gan Y, Wu Y, Wang J (2021) Epidemics trend prediction model of covid-19. CAAI Trans Intell Syst 16(3):528–536

34. Iqbal M, Al-Obeidat F, Maqbool F, Razzaq S, Anwar S, Tubaishat A, Khan MS, Shah B (2021) Covid-19 patient count prediction using LSTM. IEEE Trans Comput Soc Syst

35. Helli SS, Demirci Ç, Çoban O, Hamamci A (2020) Short-term forecasting covid-19 cases in Turkey using long short-term memory network. In: 2020 medical technologies congress (TIPTEKNO). IEEE, pp 1–4

36. Xie X, Luo K, Yin Z, Wang G (2021) Nonlinear combinational dynamic transmission rate model and its application in global covid-19 epidemic prediction and analysis. Mathematics 9(18):2307

37. Gupta VK, Gupta A, Kumar D, Sardana A (2021) Prediction of covid-19 confirmed, death, and cured cases in India using random forest model. Big Data Min Anal 4(2):116–123

38. Shahid F, Zameer A, Muneeb M (2020) Predictions for covid-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos Solitons Fractals 140:110212 110212

39. Masum M, Shahriar H, Haddad HM, Alam MS (2020) r-LSTM: time series forecasting for covid-19 confirmed cases with LSTMbased framework. In: 2020 IEEE international conference on big data (big data). IEEE, pp 1374–1379

40. Assimakis N, Ktena A, Manasis C, Mele E, Kunicina N, Zabasta A, Juhna T (2020) Using the time varying Kalman filter for prediction of covid-19 cases in Latvia and Greece. In: 2020 IEEE 61th international scientific conference on power and electrical engineering of Riga Technical University (RTUCON). IEEE, pp 1–7

41. Beche R, Baila R, Marginean A (2020) Covid-19 spread forecast using recurrent auto-encoders. In: 2020 IEEE 16th international conference on intelligent computer communication and processing (ICCP). IEEE, pp 117–122

42. Mousavi M, Salgotra R, Holloway D, Gandomi AH (2020) Covid-19 time series forecast using transmission rate and meteorological parameters as features. IEEE Comput Intell Mag 15(4):34–50

43. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen N-C, Tung CC, Liu HH (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond Ser A Math Phys Eng Sci 454(1971):903–995

44. Huang G-B, Zhu Q-Y, Siew C-K (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541), vol 2. IEEE, pp 985–990

45. Jang J-S (1993) Anfis: adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23(3):665–685

46. Jin W, Li ZJ, Wei LS, Zhen H (2000) The improvements of BP neural network learning algorithm. In: WCC 2000-ICSP 2000. 2000 5th international conference on signal processing proceedings. 16th world computer congress 2000, vol 3. IEEE, pp 1647–1649

47. Chow T, Leung C (1996) Neural network based short-term load forecasting using weather compensation. IEEE Trans Power Syst 11(4):1736–1742