SCIENTIFIC REPORTS
nature research

OPEN

A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China

Jiangyan Gu[1,2], Lizhong Liang[3], Hongquan Song[1,2,4*], Yunfeng Kong[1,2*], Rui Ma[1], Yane Hou[1], Jinyu Zhao[1], Junjie Liu[1], Nan He[1] & Yang Zhang[5]

Hand-foot-mouth disease (HFMD) is a common infectious disease in children and is particularly severe in Guangxi, China. Meteorological conditions are known to play a pivotal role in the HFMD. Previous studies have reported numerous models to predict the incidence of HFMD. In this study, we proposed a new method for the HFMD prediction using GeoDetector and a Long Short-Term Memory neural network (LSTM). The daily meteorological factors and HFMD records in Guangxi during 2014–2015 were adopted. First, potential risk factors for the occurrence of HFMD were identified based on the GeoDetector. Then, region-specific prediction models were developed in 14 administrative regions of Guangxi, China using an optimized three-layer LSTM model. Prediction results (the R-square ranges from 0.39 to 0.71) showed that the model proposed in this study had a good performance in HFMD predictions. This model could provide support for the prevention and control of HFMD. Moreover, this model could also be extended to the time series prediction of other infectious diseases.

Hand-foot-mouth disease (HFMD) is a common viral infectious disease in children under 5 years old, which is commonly caused by the enteric pathogen coxsackievirus A16 (CoxA16) and enterovirus 71 (EV 71)[1,2]. Severe HFMD could be associated with serious complications, such as poliomyelitis and brainstem encephalitis, which may be life-threatening[3]. HFMD has resulted in several outbreaks throughout the world and become a public health issue in Asia[4–7]. HFMD ranked first among the notifiable infectious diseases in China in 2017[8]. The incidence and mortality rate of HFMD are particularly severe in Guangxi Zhuang Autonomous Region of China[9]. There are no specific drugs or vaccines to prevent HFMD[10,11] and therefore it is essential to establish a reliable prediction model for the prevention of HFMD.

HFMD has obvious periodic variations, such as its peak period usually occurs during summer months in the northern hemisphere[12]. Previous studies have revealed that HFMD is closely related to meteorological conditions[13–16], such as the mean temperature, relative humidity, wind speed, and sunshine hours. It should be possible to establish models to predict the occurrence of HFMD based on these associations. The prediction model would enable individuals as well as hospitals and clinics formulate precautions and minimize health risks.

Numerous studies have been carried out to develop HFMD prediction models. There are three categories of prediction models, including linear regression, time series, and machine learning. The linear regression model was established by analyzing the correlations between the incidence of HFMD and the influential factors[17]. However, it is difficult to capture the non-linear association between HFMD and impacting factors and maintain the spatial stationary assumption over a large area[18]. The time series model uses the relationship in the sequential lag time series to predict the incidence of HFMD, such as the seasonal auto-regressive integrated moving average model (ARIMA)[19,20]. These models did not consider the relationship between HFMD and potential impacting factors. With the development of artificial intelligence (AI), machine learning algorithms have shown their advantages in

[1]Laboratory of Geospatial Technology for the Middle and Lower Yellow River Regions, Ministry of Education, Henan University, Kaifeng, Henan, 475004, China. [2]Institute of Urban Big Data, College of Environment and Planning, Henan University, Kaifeng, Henan, 475004, China. [3]The Affiliated Hospital of Guangdong Medical University, Zhanjiang, 524001, China. [4]Henan Key Laboratory of Integrated Air Pollution Control and Ecological Security, Henan University, Kaifeng, Henan, 475004, China. [5]Institute for Global Innovation and Development, East China Normal University, Shanghai, 200062, China. *email: hqsong@henu.edu.cn; yfkong@henu.edu.cn

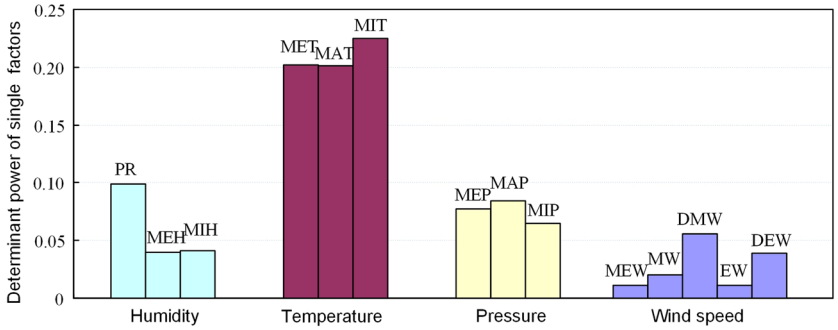

Figure 1. Determinant power of the potential impacting factors of HFMD.

predictions and recognitions[21–23]. Gradient boosting tree (GBT) and random forest (RF) were found to be capable of identifying both mild and severe HFMD, which is helpful for early surveillance and control in HFMD[24,25]. Deep learning methods such as Back Propagation Neural Networks (BPNN) were also adopted to predict the incidence of HFMD[26]. However, conventional machine learning methods such as BPNN cannot effectively deal with the trend prediction of HFMD since the temporal pattern must be taken into account when predicting infectious diseases.

To overcome the limitations mentioned above, this study proposed a HFMD prediction method using the GeoDetector (http://geodetector.org) and the Long Short-Term Memory Neural Network (LSTM). GeoDetector measures the association between input factors and dependent factors according to their temporal-spatial distributions by the indicator q-statistic (q) value, the value $\in [0,1]$ increases as the association between the input factor and HFMD increase[27]. LSTM is an advanced kind of Recurrent Neural Network (RNN), which has the ability to learn temporal pattern and store the useful memory for a longer time. In this study, the GeoDetector was employed to analyze the impact of every meteorological factor and the interactive effects between different factors on HFMD. And then the dominant impacting factors were input into LSTM model to predict the weekly cases of HFMD in 14 subregions of Guangxi, China.

Results

Identification of potential impacting factors. Figure 1 shows the contributions of 14 meteorological factors to the occurrence of HFMD in 14 subregions of Guangxi from 2014 to 2015. The 14 meteorological factors were divided into four categories, including (1) Humidity: minimum relative humidity (MIH), mean relative humidity (MEH), and precipitation (PR); (2) temperature: mean temperature (MET), maximum temperature (MAT), and minimum temperature (MIT); (3) pressure: mean pressure (MEP), maximum pressure (MAP), and minimum pressure (MIP); (4) wind speed: mean wind speed (MEW), maximum wind speed(MW), the direction of maximum wind speed (DMW), extreme wind speed (EW), and the direction of extreme wind speed (DEW). It can be seen that the primary impacting factor is the temperature of MIT ($q = 0.23$) and MET ($q = 0.20$), followed by PR ($q = 0.10$) and wind speed (DEW, $q = 0.04$; MW, $q = 0.02$; EW, $q = 0.01$; MEW, $q = 0.01$). This indicated that the q value was similar for the category of potential impacting factors, which means that they may have the similar contribution to the occurrence of HFMD, but does not mean that they influence the HFMD in the same way.

The interactive effect is not simply equal to the sum of the q values of the two influencing factors' ($X1$ and $X2$) effects on HFMD, which is represented as $q(X1 \cap X2)$[28]. Figure 2 shows the interactive results between potential influencing factors (only show a subset ($q(X1 \cap X2) > 0.20$) of all the interactions due to the space limitation, entire results could be found in the Supplementary Fig. S1). This indicated that any two combined factors could play a more important role than their single effects on HFMD. The combinations of the temperatures and other factors had more dominated influences among all the combinations. The most four primary interactions MIT and EW ($q = 0.30$), MIT and MEW ($q = 0.30$), MIT and MIP ($q = 0.28$), and MET and MEW ($q = 0.28$). Taking the effects of both single and combined effects into consideration, MW and DEW were eliminated from the risk factors.

LSTM prediction model. To minimize the spatial difference of the effects of meteorological conditions on HFMD and improve the prediction accuracy of LSTM, we developed the region-specific models for 14 subregions of Guangxi. The HFMD cases of the previous 80 weeks were taken as the training set to train the LSTM model, and the HFMD cases of the next 24 weeks were taken as the testing set to evaluate the prediction model. The model was saved after 5000 iterations, which was applied to predict the HFMD cases in the subsequent 24 weeks. We run 20 times for each model and the mean value of the runs was considered as the prediction value.

Figure 3 shows predictions of the region-specific LSTM models compared with observations in subregions (Beihai and Wuzhou not shown here due to the lower cases of HFMD, entire results could be found in the Supplementary Fig. S2). The predictions and observations had good consistence in all subregions, which indicated that the region-specific LSTM models had good performance in the prediction of HFMD. To quantify the performance of the region-specific models, the metrics of R Square (R^2) and Mean Absolute Percent Error ($MAPE$) were adopted to evaluate the performance of these models. Table 1 shows the performance of the 14 region-specific models in Guangxi. The mean R^2 and $MAPE$ of these models was 0.60 and 0.73, respectively. The R^2 ranges from 0.39 to 0.71 and the $MAPE$ ranges from 23% to 131%. The region-specific model showed best and worst performance in Baise ($R^2 = 0.71$, $MAPE = 30.46\%$) and Chongzuo ($R^2 = 0.39$, $MAPE = 55.29\%$), respectively.

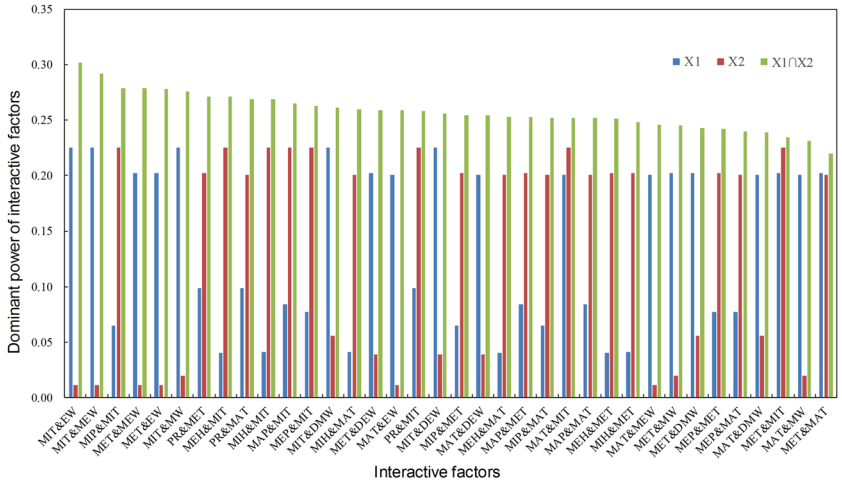

Figure 2. Interactive effects between the potential influencing factors on HFMD. The x-axis label *X*1 & X2 denotes q values of *X*1 (blue), *X*2 (red), and the interaction between *X*1 and *X*2 (green).

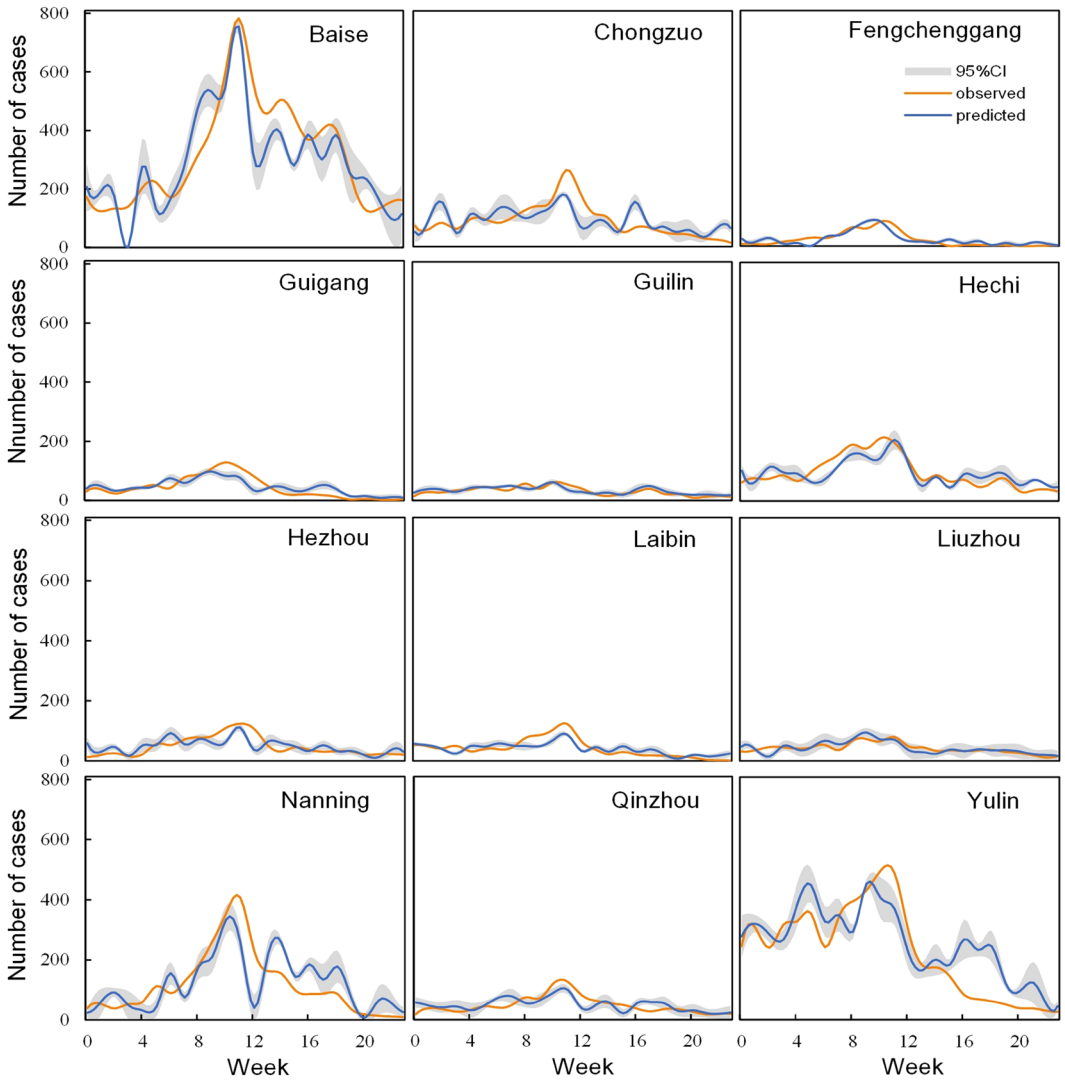

Figure 3. Region-specific model predictions of HFMD compared with observations in subregions. The grey shaded areas demote the 95% confidence interval (CI) of the predictions.

Region	Chongzuo	Hezhou	Qinzhou	Liuzhou	Nanning	Beihai	Guilin
R^2	0.39	0.40	0.49	0.54	0.56	0.60	0.61
MAPE	0.55	0.52	0.43	0.23	0.92	1.32	0.35
Region	**Laibin**	**Wuzhou**	**Guigang**	**Yulin**	**Fangchenggang**	**Hechi**	**Baise**
R^2	0.64	0.65	0.68	0.68	0.70	0.70	0.71
MAPE	1.05	1.38	0.96	0.75	1.07	0.36	0.30

Table 1. The performance of the region-specific models in subregions of Guangxi.

Discussion

It has been reported that the incidence of HFMD significantly increased in recent years in China[29], particularly in the southeast areas. This study proposed a method for predictions of the HFMD occurrence using GeoDetector and LSTM model.

Primary potential impacting factors for HFMD were identified by the GeoDetector. Compared with the conventional regression models, the GeoDetector can not only identify non-linear associations but also detect interactive effects from multiple variables. According to the detection results of the GeoDetector, the *q* values of the temperatures and the precipitation are at high levels, indicating that they could be crucial influential factors of the HFMD. Particularly, the temperatures rank higher associations than other variables. In addition, interactions of temperatures and wind speeds rank highest, which means that these combinations play an important role in the occurrence of HFDM. This finding is consistent with previous studies[13–16]. Existing researches explained the association between climate and disease, such as wind can promote circulation and distribution of air pollutants like particulate matter carrying enterovirus to accelerate the transportation of HFMD[30]. Temperature change could not only influence the children's immune capacity[31], but can also influence human's direct contact to increase the opportunity of HFMD transmissions, as physical activities among individuals are increasing in warm months[32]. A study in Hefei, China, has found that more than half of HFMD cases occurred on rainy days, because the wet weather is suitable for the HFMD virus to survive and multiply[33].

Predictions of the region-specific models in 14 subregions of Guangxi proved that LSTM has the ability to predict the occurrence of HFMD. Numerous studies have contributed to the construction of HFMD prediction model[18–20]. However, this may be the first time to adopt the LSTM, to the best of our knowledge, to predict the occurrence of HFMD based on meteorological conditions. The mean R^2 and MAPE of the 14 region-specific models was 0.60 and 0.73, respectively. Compared with previous works predicted by other methods[34,35] (mean $R^2 = 0.54$, mean MAPE = 1.02), indicating that the model proposed in this study had higher accuracy in the HFMD predictions. However, the performance of the LSTM model had spatial variations, which may be due to the meteorological conditions may not be the primary driving factors in some regions. For example, the socio-economic factors may be the primary driving forces for the incidence of HFMD[36].

It should also be noted that there were some limitations in the region-specific LSTM prediction model proposed in this study. The training dataset may be partial and insufficient due to the uneven distribution of hospitals, which may result in uncertainties for the prediction of HFMD. However, we believe that this can capture the temporal trends of HFMD occurrence for each subregion, because the dataset adopted in this study was collected continuously from most hospitals in each region. In addition, we adopted normalized and optimized method to minimize the prediction errors in the model construction. Moreover, the HFMD is also affected by socio-economic factors such as population density, rural population, and proportion of student population[17,36]. However, we only considered meteorological factors in this study, which may possibly lead to inaccurate predictions in regions where meteorological factors are not strong determinants of the HFMD. Taking other potential impacting factors into the LSTM model development would improve the prediction accuracy of HFMD occurrence.

In conclusion, this study proposed a method for predicting HFMD using GeoDetector and LSTM. The method was proved to be accurate and effective. Although this model cannot be applied directly in other studies due to the parameters in deep learning algorithm models vary with training data, the framework proposed in this study can be extended to predict other infectious diseases in other study areas. The capability of LSTM in dealing with time series issues could be applied more extensively in further researches.

Materials and Methods

Data sources. We collected the daily historical first page data of medical records and meteorological data from January 2014 to December 2015 in Guangxi, China. Guangxi is located in southeast China, adjacent to the South China Sea. Most of it is in the subtropical zone with a monsoon humid and rainy climate (Fig. 4).

The first page data of medical records for hospitals were collected from 14 administrative regions of Guangxi, including patients age, source, admission time, hospital stay, diagnosis, operation, payment, and the form of payment. Considering the scale of the first page data, the big data technologies such as data cleaning and denoising were adopted during the first page data preprocessing. The HFMD cases were defined according to the International Classification of Diseases B08.4. Finally, there were 170920 records of HFMD were adopted in this study, which were divided by 14 administrative regions and 104 weeks.

Meteorological data were obtained from the China Meteorological Data Sharing Service System. The original data were collected at 99 meteorological stations in and around Guangxi, including 14 meteorological factors, MIH, MEH, PR, MET, MAT, MIT, MEP, MAP, MIP, MEW, MW, DMW, EW, and DEW. In order to get the meteorological data for each week and each region, a model equipped with iterator was built by using the Model Builder function of ArcMap 10 (https://desktop.arcgis.com/en/arcmap/). The main function of this model was to turn the

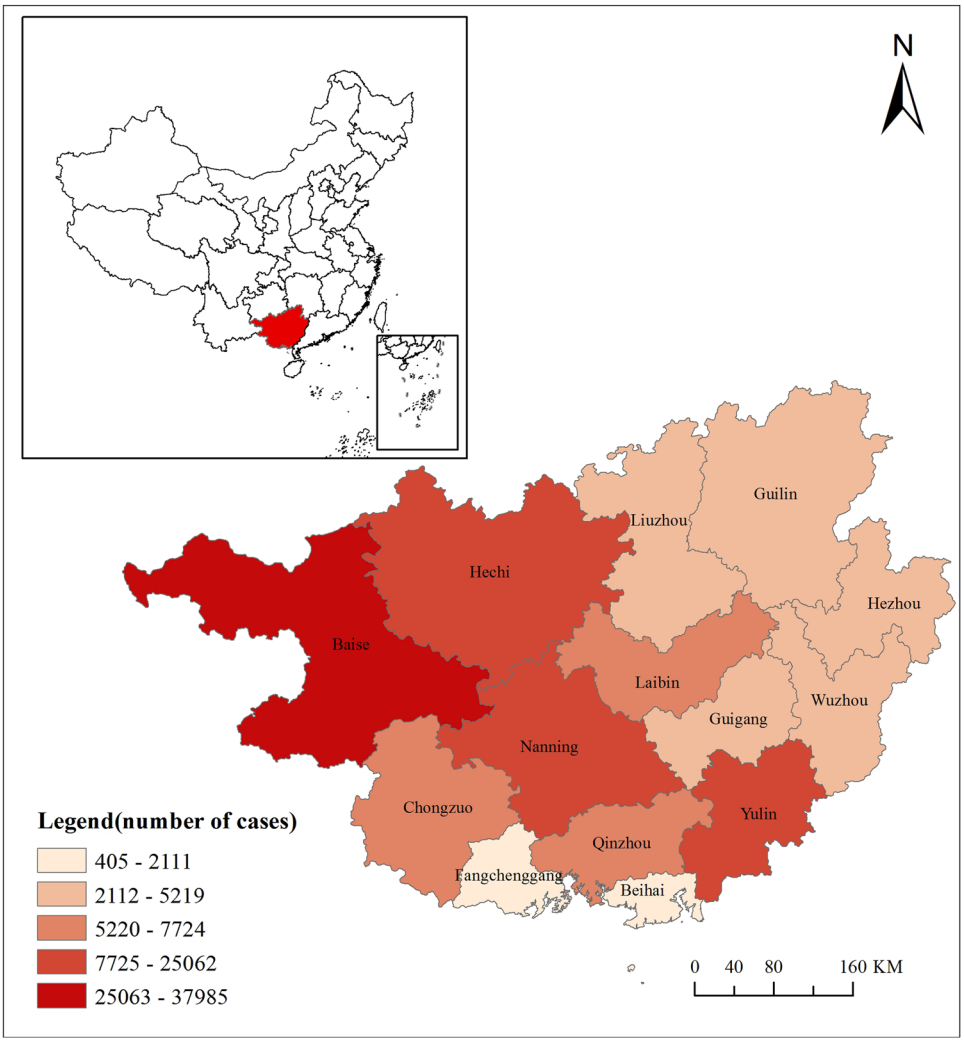

Figure 4. Location of Guangxi Zhuang Autonomous Region in China and the total number of HFMD cases during 2014–2015.

daily station records to weekly region records by spatiotemporal Kriging interpolation method. Spatiotemporal interpolation method was an extended interpolation way based on spatial temporary relativity, considering the variables both in space and time, the estimated value of a time-space spot is calculated by the weighted sum of its surrounding observations. Main methods include spatiotemporal Kriging, BME (Bayesian Maximum Entropy), and synthesis method[37]. Among all the spatial interpolation ways, the spatiotemporal Kriging was a simple way and commonly used for the interpolation of climatic data[38,39].

GeoDetector. GeoDetector is a statistical method to detect the temporal-spatial heterogeneity. This tool has been widely used in many areas, such as heavy metal differentiation, land use, and disease risk factor detection[40–42]. The assumption is that, if a potential factor leads to a disease, this factor would show a temporal-spatial distribution similar to the disease. In this study, GeoDetector was adopted to identify the risk factors from the 14 candidate meteorological factors that caused the temporal spatial stratified heterogeneity of HFMD in Guangxi from 2014 to 2015. Then, the impacting factors were identified according to the ranking of q values. The calculation of q is as follows:

$$q = 1 - \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(Y_{hi} - \overline{Y}_h)^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} \quad (1)$$

This equation assumes that the study area is composed of N units and is stratified into h ∈ [1, 2, 3 ..... L] strata, Y_i is the value of sample i, i is the whole sample population, Y_{hi} means the value of sample i in stratum h, $\overline{Y}_h$ is the mean value of stratum h, $\overline{Y}$ is the mean value of population. A higher value of q indicates a stronger spatially stratified heterogeneity of Y; it means that factor X can explain 100 * q % of the temporal-spatial pattern of Y. Moreover, according to the rules and principle of GeoDetector, the X should be a categorized variable instead of numerical variables[28,43], therefore, the continuous meteorological factors were categorized into six levels using k-means cluster algorithm.

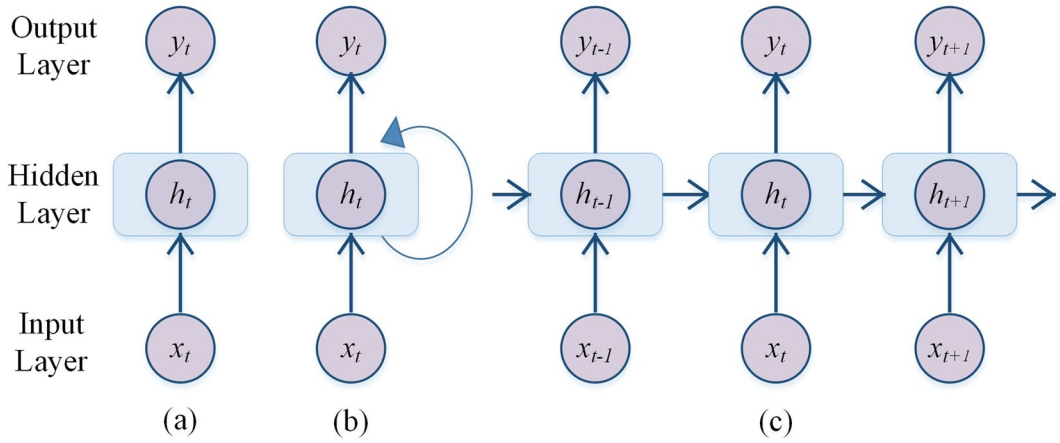

Figure 5. Architecture of artificial neural networks. (**a**) Architecture of feed-forward neural network. (**b**) Architecture of RNN. (**c**) Architecture of RNN unfolded in time.

LSTM neural networks. LSTM is a special kind of RNN (Recurrent Neural Network). RNN has the ability to learn patterns and extract features from data that contains time series due to the special structure in context layer, a hidden layer with repeated connections in neurons. LSTM has been widely used in recognitions of image and speech due to its high accuracy[44,45]. However, it remains a long-term dependency issue in RNN due to the exploding gradient problem resulting from gradient propagation over many layers. LSTM was designed to overcome this issue through cell-and-gate structure, which enables the LSTM to learn when they forget and update memory[46]. LSTM has a better performance than tradition statistical models and has been applied in predicting such as emotional state, traffic flow, and disease, especially when combined with convolutional neural networks[47–49].

In order to learn the pattern in time series, the conventional deep feed-forward neural networks must be improved (Fig. 5a). The lack of connections among the nodes within the hidden layers may result in failure in dealing with time series problems. Therefore, RNN, a kind of neural network equipped with recurrent connections in the neurons of hidden layers, has been developed and improved[50]. Figure 5b,c show a basic RNN architecture and unfolded architecture in time. The connections or loops can transport feedback from the previous state to the current state, allowing information to be passed between the consecutive temporal steps. Hence, it can be seen that the output not only depends on the input information, but also depends on the output of the previous hidden layer.

The model in Fig. 5a can be expressed mathematically as follows:

$$h_t = f_1(w_1 x_t + b_1) \tag{2}$$

$$y_t = f_2(w_2 h_t + b_2) \tag{3}$$

where x_t is the input variable, h_t is the temporary variable in hidden layer or hidden state; w_1 and w_2 are weight metrics from input layer to hidden layer and from hidden layer to output layer, respectively; b_1 and b_2 are bias vectors; f_1, and f_2 are hidden and output activation functions, respectively. The activation functions are nonlinear functions, making the neural networks approximate any continual nonlinear functions with any precision[51].

The mathematical expressions of the feedback loop in a hidden layer are as follow equations (Fig. 5b):

$$h_t = f_1(w_{t1} x_t + w_{t2} h_{t-1} + b_1) \tag{4}$$

$$y_t = f_2(w_2 h_t + b_2) \tag{5}$$

where the w_{t1}, w_{t2} and w_2 are weight matrices; b_1, b_2, f_1 and f_2 have the same meaning as described for conventional deep feed-forward neural networks. The same weights are used at each time step to calculate the output y_t. The loop makes the h_t at time t calculated not only by the input x_t but also by the previous output h_{t-1}, which is consistent with the unfolded RNN architecture shown in Fig. 5b. It can be seen that the temporal information is continuously reflected over time. The RNN is actually a very deep neural network trained using back propagation algorithm in time direction. However, due to the vanishing gradient or exploding gradient problems that always occur in very deep neural networks[52], the accuracy of the RNN deteriorates quickly over a long period of time, which means the RNN can only store short-term memory. This is called the long-term dependency of RNNs.

The LSTM was proposed by Sepp Hochreiter and Jürgen Schmidhuber to overcome the long-term dependency of RNNs in 1997[53]. For LSTM, the hidden neurons of the RNN are replaced by LSTM memory units (Fig. 6a). The memory units are mainly composed of three gates and one cell (Fig. 6b), which aim to control information flow and storage, including input gate, forget gate, output gate and memory cell. The system is determined by the state of these structures at each time step whether the information will be retained. Thus, the LSTM could hold important short-term memory for a longer period by the filtration of the memory units.

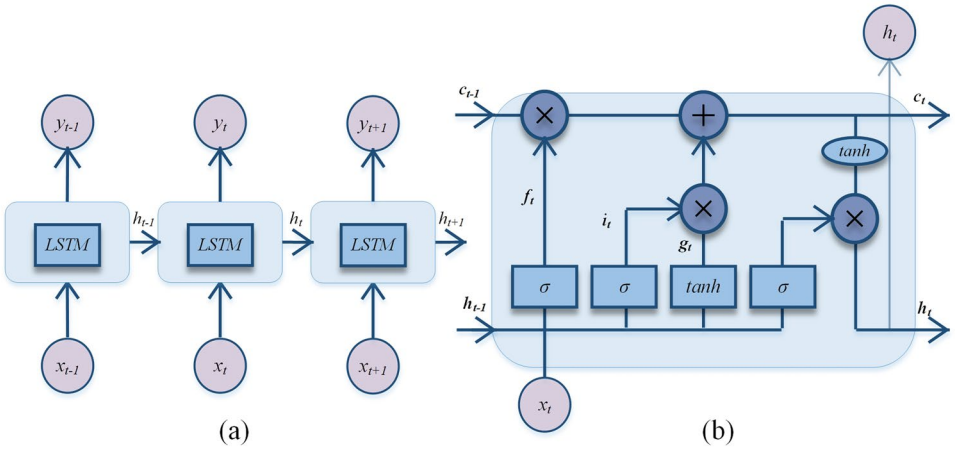

Figure 6. Architecture of LSTM. (**a**) Architecture of LSTM. (**b**) Architecture of LSTM memory unit.

The formulas for calculations in the memory unit are performed as follows:

$$i_t = \sigma(w_{i1}x_t + w_{i2}h_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(w_{f1}x_t + w_{f2}h_{t-1} + b_f) \tag{7}$$

$$o_t = \sigma(w_{o1}x_t + w_{o2}h_{t-1} + b_0) \tag{8}$$

$$g_t = tanh(w_{g1}x_t + w_{g2}h_{t-1} + b_g) \tag{9}$$

$$c_t = f_t * c_{t-1} + i_t * g_t \tag{10}$$

$$h_t = o_t * \tanh(c_t) \tag{11}$$

here, the input gate i_t, forget gate f_t, and output gate o_t take variable x_t and previous hidden state h_{t-1} as inputs at time t, multiplied with the weight matrices w_{i1} w_{i2} w_{f1} w_{f2} w_{01} x_t w_{02}, then plus bias vectors b_i, b_f, b_o, σ means the sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$, generating the outputs of these gates range from 0 to 1. The larger output value is, the more information allowed, i.e., if the output value equals 1, it means the information is fully entered. The c_{t-1} means previous state of memory cell; the * indicates element-wise multiplication. *tanh* is a kind of activation function such that, $tanh(z) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. The calculations are as follows: after inputting the candidate information g_t to be stored, the actual amount stored is determined by the input gate i_t. Then in the Eq.(10), the final cell sta$_t$e c_t depends on the sum of the amount of newly input informat*i*on $i_t * g_t$ and the amount of forgotten information $f_t * c_{t-1}$, h_t means the final output of the memory unit, controlled by the output gate o_t and the final cell state c_t. The LSTM model was constructed in Python 3.5 and was supported by modules including Tensorflow, Numpy, and Pandas.

Model design. In order to identify the risk meteorological factors of the temporal-spatial distribution of HFMD in Guangxi, the preprocessed data were sent to GeoDetector, including weekly normalized cases and categorized meteorological factors in each region. In addition, the weekly cases need to be normalized in data preprocessing due to the collected data cannot cover all the hospitals, so the case number instead of incidence was adopted as dependent variable. However, the number of cases between regions had large spatial differences because of the uneven distribution of hospitals and population. To solve this issue, we standardizes the number of cases to eliminate dimension. Z-score was adopted to normalize the data. The Z-score calculation method is:

$$x^* = \frac{x - \overline{x}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}} \tag{12}$$

The environmental factors are different in different regions of Guangxi. To minimize the effect of spatial scale and improve the accuracy of LSTM, the region-specific model was built for each of the 14 regions of Guangxi. After the identification by using GeoDetector, the identified impacting factors and weekly cases in any given region were used as inputs and outputs, respectively. To develop the LSTM prediction model, the hospital data of the first 80 weeks work as the training set and the last 24 weeks work as the testing set. The number of input layers was equal to the number of identified factors and the neuron in hidden layer was set as 10 after experiments and adjustments. Because the output is a continuous variable, the output layer was set as 1 without activation functions. To study the model with continuous variables, the root-mean-square error was set as loss function.

Backward propagation with gradient descent was used as a training algorithm to minimize the result of loss function. The optimization of the LSTM prediction model had three parts: L2 regularization, moving average model and exponential decay learning rate. First, in order to avoid the overfitting problem caused by the limited number of data, L2 regularization was adopted in loss function as shown in equations[54], where c_0 denotes the original loss function; the ε denotes the regularization rate, usually set as a very small number, such as 0.0001 in this paper; and $\sum_w w^2$ indicates the sum of squares of all the weights.

$$c_0 = \frac{1}{n}\sum_{i=1}^{n}(y_{pi} - y_i)^2 \tag{13}$$

$$c = c_0 + \frac{\varepsilon}{2n}\sum_{w} w^2 \tag{14}$$

Second, the moving average model is always accompanied by a gradient decay algorithm to improve the capacity and extensiveness of the final output model. Shadow variables following the variables are held and updated with iterations by the model to control the update rate[55], after experiments, the decay rate was determined as 0.99.

$$shadow_i = \text{decay\_rate} * shadow_{i-1} + (1 - \text{decay\_rate}) * variable_i \tag{15}$$

The other important parameter in neural networks is learning rate. It has been proved that dynamic rate has a better effect than fixed rate in training[56]. Therefore, in this study, an exponential decay learning rate was adopted, and the mathematical mechanism is expressed in Eq. (16). A better solution is using a higher learning rate at first, called learning rate base, then it will decrease gradually as the increase in global step over time. The learning rate base was determined as 0.1, and the decay rate was 0.99 after experiments.

$$\text{learning\_rate} = \text{base\_rate} * (decay\_rate)^{\frac{global\_step}{decay\_steps}} \tag{16}$$

Ethics approval. The present study was approved by the medical and research ethics committee of Henan University (see the ethical approval in the related manuscript file), all methods were carried out in accordance with relevant guidelines and regulations. Since there were no individual information involved and the records are anonymized in this study, the approving IRB has waived the need to collect informed consent, so the informed consent was not required in the present study.

Data availability

The meteorological dataset used during the current study are available from China Meteorological Data Sharing Service System (http://data.cma.cn/). The hospital datasets can be extracted as presented in the website of data collection, or contact the corresponding author on reasonable request.

Received: 3 June 2019; Accepted: 14 November 2019;
Published online: 29 November 2019

References

1. Wang, J. F. *et al*. Hand, foot and mouth disease: spatiotemporal transmission and climate. *International Journal of Health Geographics* **10**, https://doi.org/10.1186/1476-072x-10-25 (2011).
2. Xu, C. D. & Xiao, G. X. Spatiotemporal risk mapping of hand, foot and mouth disease and its association with meteorological variables in children under 5 years. *Epidemiology and Infection* **145**, 2912–2920, https://doi.org/10.1017/s0950268817001984 (2017).
3. Ventarola, D., Bordone, L. & Silverberg, N. Update on hand-foot-and-mouth disease. *Clinics in Dermatology* **33**, 340–346, https://doi.org/10.1016/j.clindermatol.2014.12.011 (2015).
4. Schmidt, N. J., Lennette, E. H. & Ho, H. H. An apparently new enterovirus isolated from patients with disease of the central nervous system. *The Journal of infectious diseases* **129**, 304–309, https://doi.org/10.1093/infdis/129.3.304 (1974).
5. Chen, K.-T., Chang, H.-L., Wang, S.-T., Cheng, Y.-T. & Yang, J.-Y. Epidemiologic features of hand-foot-mouth disease and herpangina caused by enterovirus 71 in Taiwan, 1998-2005. *Pediatrics* **120**, E244–E252, https://doi.org/10.1542/peds.2006-3331 (2007).
6. Ang, L. W. *et al*. Epidemiology and Control of Hand, Foot and Mouth Disease in Singapore, 2001-2007. *Annals Academy of Medicine Singapore* **38**, 106–112 (2009).
7. Lu, Q.-B. *et al*. Circulation of Coxsackievirus A10 and A6 in Hand-Foot-Mouth Disease in China, 2009–2011. *Plos One* **7**, https://doi.org/10.1371/journal.pone.0052073 (2012).
8. National Survey of Epidemic Situation of Notifiable Communicable Diseases in 2017 (accessed on 1 January *2019)*, http://www.nhc.gov.cn/jkj/s3578/201802/de926bdb046749abb7b0a8e23d929104.shtml (1990).
9. Jiang, L. *et al*. Epidemiological characteristics and temporal-spatial clustering of hand,foot and mouth disease in Guangxi from 2008 to 2015. *Chinese. Journal of Disease Control & Prevention* **21**, 340–344 (2017).
10. Wu, K. X., Ng, M. M. L. & Chu, J. J. H. Developments towards antiviral therapies against enterovirus 71. *Drug Discovery Today* **15**, 1041–1051, https://doi.org/10.1016/j.drudis.2010.10.008 (2010).
11. Lee, B. Y. *et al*. Forecasting the economic value of an Enterovirus 71 (EV71) vaccine. *Vaccine* **28**, 7731–7736, https://doi.org/10.1016/j.vaccine.2010.09.065 (2010).
12. Druyts-Voets, E. Epidemiological features of entero non-poliovirus isolations in Belgium 1980–94. *Epidemiology and infection* **119**, 71–77, https://doi.org/10.1017/s0950268897007656 (1997).
13. Onozuka, D. & Hashizume, M. The influence of temperature and humidity on the incidence of hand, foot, and mouth disease in Japan. *Science of the Total Environment* **410**, 119–125, https://doi.org/10.1016/j.scitotenv.2011.09.055 (2011).
14. Xu, C. Spatio-Temporal Pattern and Risk Factor Analysis of Hand, Foot and Mouth Disease Associated with Under-Five Morbidity in the Beijing-Tianjin-Hebei Region of China. *International Journal of Environmental Research and Public Health* **14**, https://doi.org/10.3390/ijerph14040416 (2017).

15. Zhang, X., Xu, C. & Xiao, G. Space-time heterogeneity of hand, foot and mouth disease in children and its potential driving factors in Henan, China. *Bmc Infectious Diseases* **18**, https://doi.org/10.1186/s12879-018-3546-2 (2018).
16. Liao, J., Qin, Z., Zuo, Z., Yu, S. & Zhang, J. Spatial-temporal mapping of hand foot and mouth disease and the long-term effects associated with climate and socio-economic variables in Sichuan Province, China from 2009 to 2013. *Science of the Total Environment* **563**, 152–159, https://doi.org/10.1016/j.scitotenv.2016.03.159 (2016).
17. Bo, Y.-C., Song, C., Wang, J.-F. & Li, X.-W. Using an autologistic regression model to identify spatial risk factors and spatial risk patterns of hand, foot and mouth disease (HFMD) in Mainland China. *Bmc Public Health* **14**, https://doi.org/10.1186/1471-2458-14-358 (2014).
18. Xu, M. *et al*. Non-Linear Association between Exposure to Ambient Temperature and Children's Hand-Foot-and-Mouth Disease in Beijing, China. *Plos One* **10**, https://doi.org/10.1371/journal.pone.0126171 (2015).
19. Kim, B. I., Ki, H., Park, S., Cho, E. & Chun, B. C. Effect of Climatic Factors on Hand, Foot, and Mouth Disease in South Korea, 2010-2013. *Plos One* **11**, https://doi.org/10.1371/journal.pone.0157500 (2016).
20. Peng, Y. *et al*. Application of seasonal auto-regressive integrated moving average model in forecasting the incidence of hand-foot-mouth disease in Wuhan, China. *Journal of Huazhong University of Science and Technology-Medical Sciences* **37**, 842–848, https://doi.org/10.1007/s11596-017-1815-8 (2017).
21. Dong, X., Si, W. & Huang, W. ECG-based identity recognition via deterministic learning. *Biotechnology & Biotechnological Equipment* **32**, 769–777, https://doi.org/10.1080/13102818.2018.1428500 (2018).
22. Yao, Y. *et al*. A paired neural network model for tourist arrival forecasting. *Expert Systems with Applications* **114**, 588–614, https://doi.org/10.1016/j.eswa.2018.08.025 (2018).
23. Kucukoglu, I., Simsek, B. & Simsek, Y. Multidimensional Bernstein polynomials and Bezier curves: Analysis of machine learning algorithm for facial expression recognition based on curvature. *Applied Mathematics and Computation* **344**, 150–162, https://doi.org/10.1016/j.amc.2018.10.012 (2019).
24. Zhang, B. *et al*. Machine Learning Algorithms for Risk Prediction of Severe Hand-Foot-Mouth Disease in Children. *Scientific Reports* **7**, https://doi.org/10.1038/s41598-017-05505-8 (2017).
25. Liu, G. *et al*. Developing a Machine Learning System for Identification of Severe Hand, Foot, and Mouth Disease from Electronic Medical Record Data. *Scientific Reports* **7**, https://doi.org/10.1038/s41598-017-16521-z (2017).
26. Xiang, L., Yuan, G., Yang, X. & Zhu, M. The model of back-propagation neural network about meteorological factors and hand-foot-mouth disease in Baoshan District,Shanghai City. *Chinese*. *Journal of Disease Control & Prevention* **19**, 138–141 (2015).
27. Wang, J.-F., Zhang, T.-L. & Fu, B.-J. A measure of spatial stratified heterogeneity. *Ecological Indicators* **67**, 250–256, https://doi.org/10.1016/j.ecolind.2016.02.052 (2016).
28. Wang, J. & Xu, C. Geodetector: Principle and prospective. *Acta Geographica Sinica* **72**, 116–134 (2017).
29. Li, P. *et al*. Temporal-spatial variation of hand-foot-mouth disease in 2008 to 2014,China. *Journal of Nanjing Medical University*. *Natural Sciences Edition* **38**, 380–385 (2018).
30. Liao, Y., Ouyang, R., Wang, J. & Xu, B. A study of spatiotemporal delay in hand, foot and mouth disease in response to weather variations based on SVD: a case study in Shandong Province, China. *Bmc Public Health* **15**, https://doi.org/10.1186/s12889-015-1446-6 (2015).
31. Cheng, J. *et al*. Impact of temperature variation between adjacent days on childhood hand, foot and mouth disease during April and July in urban and rural Hefei. *China*. **60**, 883–890, https://doi.org/10.1007/s00484-015-1082-y (2016).
32. Suminski, R. R., Poston, W. C., Market, P., Hyder, M. & Sara, P. A. Meteorological conditions are associated with physical activities performed in open-air settings. *International Journal of Biometeorology* **52**, 189–197, https://doi.org/10.1007/s00484-007-0110-y (2008).
33. Cheng, J. *et al*. Associations between extreme precipitation and childhood hand, foot and mouth disease in urban and rural areas in Hefei, China. *Science of the Total Environment* **497**, 484–490, https://doi.org/10.1016/j.scitotenv.2014.08.006 (2014).
34. Wei, J. *et al*. The Effect of Meteorological Variables on the Transmission of Hand, Foot and Mouth Disease in Four Major Cities of Shanxi Province, China: A Time Series Data Analysis (2009-2013). *Plos Neglected Tropical Diseases* **9**, https://doi.org/10.1371/journal.pntd.0003572 (2015).
35. Du, Z. *et al*. Predicting the hand, foot, and mouth disease incidence using search engine query data and climate variables: an ecological study in Guangdong, China. *Bmj Open* **7**, https://doi.org/10.1136/bmjopen-2017-016263 (2017).
36. Xu, C., Zhang, X. & Xiao, G. Spatiotemporal decomposition and risk determinants of hand, foot and mouth disease in Henan, China. *Science of the Total Environment* **657**, 509–516, https://doi.org/10.1016/j.scitotenv.2018.12.039 (2019).
37. Wang, J. *et al*. Spatiotemporal data analysis in geography. *Acta Geographica Sinica* **69**, 1326–1345 (2014).
38. Sha, L. I., Hong, S. H. U. & Lin, D. Research and realization of Kriging interpolation based on spatial-temporal variogram. *Computer Engineering and Application* **47**, 25–26,38 (2011).
39. Sha, L. I., Hong, S. H. U. & Zhengquan, X. U. Study on Spatial-temporal Kriging Interpolation of Monthly Precipitation in Three Provinces of Northeast China. *Hydrology* **31**, 31–35 (2011).
40. Zuo, S., Dai, S., Li, Y., Tang, J. & Ren, Y. Analysis of Heavy Metal Sources in the Soil of Riverbanks Across an Urbanization Gradient. *International journal of environmental research and public health* **15**, https://doi.org/10.3390/ijerph15102175 (2018).
41. Shi, T. *et al*. Geo-detection of factors controlling spatial patterns of heavy metals in urban topsoil using multi-source data. *Science of the Total Environment* **643**, 451–459, https://doi.org/10.1016/j.scitotenv.2018.06.224 (2018).
42. Fei, X. *et al*. The association between heavy metal soil pollution and stomach cancer: a case study in Hangzhou City, China. *Environmental geochemistry and health*, https://doi.org/10.1007/s10653-018-0113-0 (2018).
43. Wang, J.-F. *et al*. Geographical Detectors-Based Health Risk Assessment and its Application in the Neural Tube Defects Study of the Heshun Region, China. *International Journal of Geographical Information Science* **24**, 107–127, https://doi.org/10.1080/13658810802443457 (2010).
44. He, K., Zhang, X., Ren, S., Sun, J. & Ieee. In *2016 Ieee Conference on Computer Vision and Pattern Recognition IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
45. Graves, A., Mohamed, A.-r., Hinton, G. & Ieee. In *2013 Ieee International Conference on Acoustics, Speech and Signal Processing International Conference on Acoustics Speech and Signal Processing ICASSP* 6645–6649 (2013).
46. Baek, Y. & Kim, H. Y. ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module. *Expert Systems with Applications* **113**, 457–480, https://doi.org/10.1016/j.eswa.2018.07.019 (2018).
47. Woellmer, M., Schuller, B., Eyben, F. & Rigoll, G. Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening. *Ieee Journal of Selected Topics in Signal Processing* **4**, 867–881, https://doi.org/10.1109/jstsp.2010.2057200 (2010).
48. Chen, Z., Liu, Y. & Liu, S. In *Proceedings of the 36th Chinese Control Conference* Chinese *Control Conference* (eds T. Liu & Q. Zhao) 3876–3881 (2017).
49. Kim, Y., Roh, J.-H. & Kim, H. Y. Early Forecasting of Rice Blast Disease Using Long Short-Term Memory Recurrent Neural Networks. *Sustainability* **10**, https://doi.org/10.3390/su10010034 (2018).
50. Fausett, L. V. *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*. (Prentice-Hall, 1994).
51. Chen, T. & Chen, H. J. I. T. N. N. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. **6**, 911–917 (1995).

52. Bengio, Y., Simard, P. & Frasconi, P. %J IEEE Transactions on Neural Networks. *Learning long-term dependencies with gradient descent is difficult*. **5**, 157–166 (1994).
53. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735 (1997).
54. Xu, Z., Chang, X., Xu, F. & Zhang, H. J. I. T. o. N. N. & Systems, L. L1/2 regularization: a thresholding representation theory and a fast solver. **23**, 1013–1027 (2012).
55. Khashei, M., Bijari, M. & Ardali, G. A. R. J. N. Improvement of Auto-Regressive Integrated Moving Average models using Fuzzy logic and Artificial Neural Networks (ANNs). **72**, 956–967 (2009).
56. Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A. & Mougiakakou, S. J. I. T. o. M. I. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural. *Network*. **35**, 1207–1216 (2016).

Acknowledgements

This study was financially supported by the Natural Science Foundation of China (416015694 and 41601422). We thank the donor of the datasets and the participants who contributed to the data collection of HFMD cases in this study.

Author contributions

J.Y.G., H.Q.S. and Y.F.K. were involved in conception and design of this study, J.Y.G. wrote the whole manuscript, H.Q.S. and Y.F.K. suggested the method to be used and revised the whole manuscript. L.Z.L. provided the original data. R.M., Y.E.H. and N.H. participated in data organization, J.J.L., J.J.Z. and Y.Z. participated in revision and submission of this research. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-54495-2.

Correspondence and requests for materials should be addressed to H.S. or Y.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.