IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 1, 2020, accepted March 23, 2020, date of publication March 27, 2020, date of current version April 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983799

Using Partial Least Squares Regression to Fit Small Data of H7N9 Incidence Based on the Baidu Index

RUIJING GAN[1], JIYONG TAN[1], LIYING MO[1], YU LI[1], AND DAIZHENG HUANG[1,2]

[1] School of Preclinical Medicine, Guangxi Medical University, Nanning 530021, China
[2] The Laboratory of Biomedical Photonics and Engineering, Guangxi Medical University, Nanning 530021, China

Corresponding author: Daizheng Huang (daizheng-huang@qq.com)

This work was supported in part by the BaGui Scholar Program of Guangxi Province, National Training Program of Innovation and Entrepreneurship for Undergraduates under Grant 201710598017 and Grant 201710598018, in part by the Guangxi Natural Science Foundation under Grant 2018GXNSFAA281133, and in part by the National Natural Science Foundation of China under Grant 81860604.

ABSTRACT The internet search data will help the disease control department to estimate the disease in advance. The H7N9 epidemic that occurred in Guangxi Province was used as an example to demonstrate its association with Baidu search data. At first,16 search terms which have high correlation with H7N9 disease were selected by expert determination and calculation. At the same time, the number of disease cases were downloaded from the website of Guangxi CDC. The partial least square regression was choosed to estimate after comparing the regression models for the number of epidemic cases is very less than baidu searches data. To filter independent variables, cross validation and variable importance in projection were applied. The results show that: 1.the proposed method is suitable for fitting the data of H7N9 disease with few samples, and the fitting degree is perfect. 2.it will help to screen out the important searching index which are more relate to H7N9 epidemic by using cross validation and variable import in project. 3.compared with the PCA methods, the proposed method presented great advantages in performance index, especially with the help of cross validation and variable importance in projection.

INDEX TERMS Partial least squares regression, H7N9, Baidu index, variable importance in projection.

I. INTRODUCTION

People are becoming more inclined to search for health information online before seeking medical services with the development of internet [1]. Search terms can be downloaded and analyzed to detect patterns in relation to disease rates. Moreover, they can be used to test the hypothesis that increases in specific search terms may be related to increases in disease rates [2]. Compared with traditional healthcare-based surveillance systems, surveillance systems built on internet search engine data can provide a wider range of population surveys and terms searched by individuals in a certain area. The search words directly reflect the intent of the query person and have tendentiousness. At the same time, the data of internet search engines provide real-time statistics and can keep pace with the disease outbreak completely. Therefore, the estimation of infectious diseases with the help of internet search data has attracted considerable research interest in recent years. One of the most representative is Google Flu Trends(GFT) [3]–[7].

Baidu is the most popular search engine in China [8]. Research between Baidu search data and some infectious diseases has been an ongoing effort [9]–[12].

Seasonal epidemics of H7N9 have been observed every winter and spring in some provinces of China, especially in the coastal areas of southeast China. H7N9 epidemics bring great harm to the health and safety of individuals. In recent years, several research efforts have been made to develop the spread of the H7N9 epidemic. Bayesian phylogeography was used to identify and compare migration patterns and factors predictive of H5N1 and H7N9 diffusion rates in China [13]. Tao used semi-quantitative risk assessment to assess human infection with H7N9 epidemic in Zhejiang Province [14]. Virlogeux constructed an ecological model to evaluate the animal-to-human and human-to-human transmission of H7N9 in China [15]. Standard deviational ellipse analysis was conducted to examine the directional trend of disease spreading, and retrospective space-time permutation

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

scan statistic was then used to identify the spatio-temporal cluster patterns of H7N9 outbreaks in humans [16].

However, no studies have focused on H7N9 epidemic based on Baidu search data in spite of the data of internet search can keep pace with the disease outbreak. In this study, the H7N9 epidemic that occurred in Guangxi Province from January 1, 2017, to July 31 was used as an example to demonstrate its association with Baidu search data.

Researchers have put forward various models in the fitting estimation of epidemic diseases. Different fitting models were selected according to data kinds which mainly include linear model [17]–[19] and nonlinear model [20], [21]. In this example, we found that the number of epidemic cases is very less than baidu searches data. It is the condition which has many independent variables, but few observed data. The Partial Least Square Regression(PLSR) was choosed to estimate after comparing the regression models. PLSR is a method which is suitable either where there are many correlated independent variables or where the number of independent variables are much more than the number of observations for the dependent variable [22]–[24]. This method combines principal component analysis, canonical correlation analysis, and multiple linear regression analysis [25]–[27]. The advantages of PLSR are as follows: (1) it can process regression modeling on the condition that the multiple independent variables have strong multi-collinearity. (2) it can carry out regression modeling when the number of sample points is less than the number of variables. (3) the modeling contains all original variables. and (4) The regression coefficient of every independent variable can be explained easily.

PLSR has been introduced in many literatures [28]–[30], but most of them focused on the application of multi-collinearity or had a small sample size. In this study, Variable Importance in Projection(VIP) was used for independent variable filtering [31]. It is a filtration method based on PLSR and filters variables more quickly than ordinary stepwise regression. Moreover, it can be used in cases where the number of independent variables is higher than that of samples. For a certain independent variable, VIP not only reflects the effect of the variable itself on the dependent variable but also takes into account the influence of other variables on the dependent variable indirectly through this variable [32].

The rest of the paper is organized as follows. Section 2 contains the methodology. Section 3 contains the results and discussion. Section 4 contains the conclusions.

II. METHODOLOGY

The H7N9 epidemic was managed as a statutory class B infectious disease by the National Health and Family Planning Commission of China in November 1, 2013. China has experienced five outbreaks of H7N9 epidemics since the discovery of the virus. The cases were distributed in 27 provinces. Most of the cases were sporadic and concentrated in the coastal area of East China. The number of human infections with H7N9 virus from 2013 to June 2018 is shown in Figure 1. As shown in the figure, the occurrence time of the H7N9 epidemic starts from winter to spring. From October 1, 2016, to the beginning of 2017, the fifth outbreak of H7N9 epidemic occurred with a large number of cases. With the implementation of prevention and control measures, such as closure of the live poultry trading market in Shanghai, Anhui, and Zhejiang provinces, the number of H7N9 cases significantly decreased in May 2017. The number of human infections with H7N9 virus in China from January 1, 2017, to July 1, 2018, is shown in Figure 2.

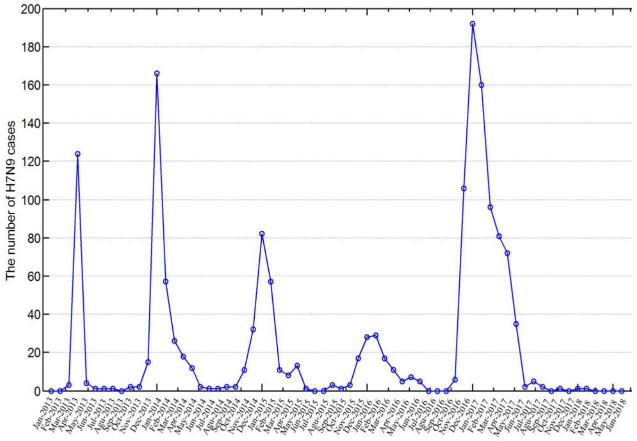


FIGURE 1. Cases of human infections with H7N9 from 2013 to June 2018 in China. Five outbreaks occurred, and the occurrence time starts from winter to spring.

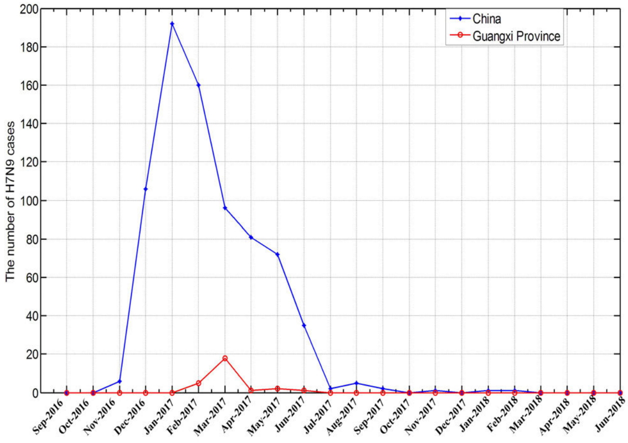


FIGURE 2. Cases of human infections with H7N9 from January 2017 to June 2018 in Guangxi and the whole country of China. Winter in Guangxi will be delayed by approximately two months compared with the rest of China.

Guangxi, officially known as the Guangxi Zhuang Autonomous Region, is located in southern China (east longitude 104° 28′ − 112° 04′ and north latitude 20° 54′ − 26° 24′) [33]. It occupies an area of 236,700 km^2 with a population of over 47.96 million people in 2015. Guangxi has 14 prefecture-level cities, and its capital is Nanning. In general, the H7N9 epidemic would occur from winter to spring. In terms of climate, winter in Guangxi is delayed by two months compared with the rest of China.

The number of human infections with H7N9 virus from January 1, 2017, to July 1, 2018, of Guangxi Province is also shown in Figure 2. As can be seen from the Figure, 1) the incidence data of H7N9 outbreak in China and Guangxi Province have the same time and trend, increased and then slowly declined (from the data of five national H7N9 outbreaks in China, the H7N9 virus outbreak in Guangxi was delayed by about two months compared with the national average). 2) the incidence data of H7N9 outbreak in China and Guangxi Province have the same onset time. Because of the geographical location, climatic characteristics, and trend characteristics of disease data, the H7N9 epidemic that occurred in Guangxi has a certain representativeness. To demonstrate regression of H7N9 epidemic associated with Baidu search data, the epidemic in Guangxi Province was used as an example.

Therefore, the H7N9 epidemic that occurred in Guangxi and China can be considered as having the same time, trend, and onset time.

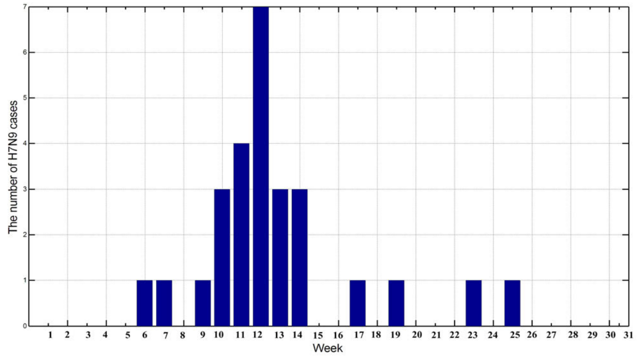


FIGURE 3. Interval of human infections with H7N9 in Guangxi from the 1st to 31st week in 2017.

As no H7N9 outbreaks occurred in Guangxi before January 2017 and no new H7N9 cases were found after July 2017, human infections with H7N9 from January 1, 2017, to July 31, 2017,were considered for analysis. The incidence data of H7N9 from January 1, 2017, to July 31, 2017,were collected weekly from the Guangxi Health Information Network (http://www.gxhfpc.gov.cn/xxgks/yqxx/ssyqdt/) and are shown in Figures 3 and 4. The Baidu search engine (http://index.baidu.com) data from the Guangxi region during the same period were collected. The search terms were carefully selected and should be related to influenza epidemics. By consulting two senior doctors knowledgeable about patient performance and a preventive medicine specialist for vocabulary advice, 30 words were searched via the Baidu index. Words with very low correlation with H7N9 disease were eliminated by calculations. Sixteen search terms via the Baidu index were selected in the end (Table 1). These terms reflect the clinical symptoms of H7N9 infection or prevention of H7N9. As Chinese are used to searching in Baidu, the Chinese words were only analyzed in the study.

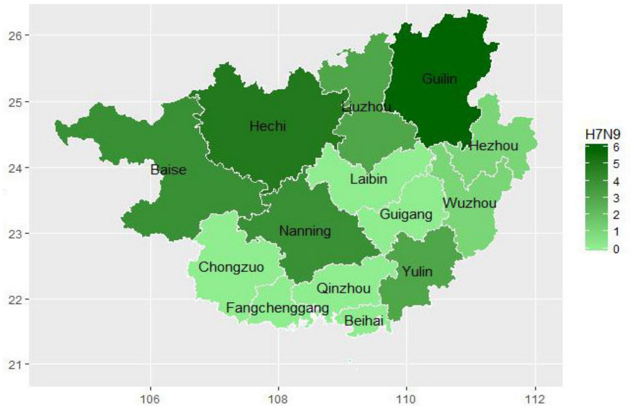


FIGURE 4. Spatial distribution maps of H7N9 epidemics that happened in 14 prefecture-level cities of Guangxi from January 1, 2017, to July 31, 2017.

TABLE 1. Chinese search words using the Baidu index and their representative English meaning and abbreviation number.

Chinese	English	Abbreviation
H7N9	H7N9	x_1
发烧	low fever（37°-38.5°）	x_2
发热	high fever (>38.5°)	x_3
肺炎	pneumonia	x_4
呼吸困难	dyspnea	x_5
肌肉酸痛	soreness	x_6
咳嗽	cough	x_7
脓毒症	sepsis	x_8
禽流感	avian influenza	x_9
预防禽流感	prevention of avian influenza	x_{10}
流感症状	influenza symptoms	x_{11}
痰液	sputum	x_{12}
头痛	headache	x_{13}
休克	shock	x_{14}
意识障碍	conscious disturbance	x_{15}
禽流感预防	avian influenza prevention	x_{16}

To determine the association of H7N9 epidemic with Baidu search data and filter independent variables in PLSR, VIP was applied. The method was performed as follows:

Step1. The relationship between search terms and infectious disease was determined using the Pearson Correlation Coefficient(PCC) [34], Spearman Correlation Coefficient (SCC) [35], and Kendall Correlation Coefficient(KCC) [36].

Step2. VIF was calculated by establishing a multiple regression equation between search terms and infectious disease to test the multi-collinearity of search terms [37].

Step3. PLSR was performed.

The proposed research model is show in Figure 5.

A. PLSR

PLSR is a multivariate regression method that projects the input–output data down in to a latent space, extracting a number of principal factors with an orthogonal structure, while capturing most of the variance in the original data. PLS derives its usefulness from its ability to analyze data

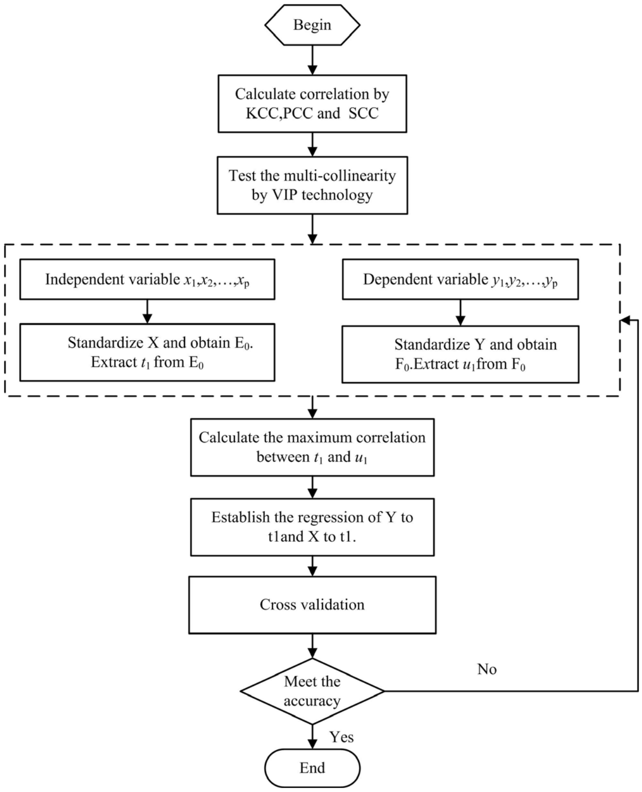

Begin
Calculate correlation by KCC,PCC and SCC
Test the multi-collinearity by VIP technology
Independent variable $x_1,x_2,\ldots,x_p$
Dependent variable $y_1,y_2,\ldots,y_p$
Standardize X and obtain E_0. Extract t_1 from E_0
Standardize Y and obtain F_0.Extract $u_1$from F_0
Calculate the maximum correlation between t_1 and u_1
Establish the regression of Y to t1and X to t1.
Cross validation
Meet the accuracy
No
Yes
End

FIGURE 5. The flow chart of the proposed algorithm.

with strongly collinear, noisy and numerous variables in the predictormatrix X and responses Y [38].

Suppose that there area single dependent variable Y and a set of independent variable matrix $X=(x_1,x_2,\ldots,x_p)$. The steps of PLSR are as follows:

Step1. Standardize X and obtain E0. Standardize Y and obtain F_0.

Step2. Extract t_1 from E_0 and u_1 from F_0 (here $u_1 = F_0$ and t_1 is the linear combination of $(x_1,x_2,\ldots,x_p)$). t_1 and u_1 must carry the most variable information of their own data, and the degree of relevance between t_1 and u_1 should be maximal, namely, $\mathrm{Var}(t_1) \to \max$, $\mathrm{Var}(u_1) \to \max$, and $r(t_1, u_1) \to \max$.

Step3. Establish the regression of Y to t_1 and X to t_1.

Step4. Evaluate the accuracy of regression equations. The algorithm will stop if the accuracy of regression equations becomes satisfactory. Otherwise, residual vector will be calculated, and t_1 and u_1 will be replaced in step2. Steps 2 to 4 will be repeated until the accuracy of regression equations becomes satisfactory.

Step5. Establish the regression equation of the original variable Y to X. Cross validation is usually used to determine the extraction number of components in PLSR. It is assumed that components of m have been extracted from X to establish the regression equation. The regression equation of Y to $(x_1,x_2,\ldots,x_p)$ will be established by transformation in the end.

B. CROSS VALIDATION

For cross validation, only the important components are needed instead of all components for PLSR modeling. To judge whether one component is necessary or not, the component will be added to the model, and the effect of the model prediction is observed. The component is important when the effect of the model prediction is obviously improved. Cross validation is usually used to determine the number of components extracted from PLSR. The cross validation is defined as follows:

$$Q_h^2 = 1 - \frac{PRESS_h}{SS_{h-1}}. \tag{1}$$

where y_i is the original data of dependent variable. $t_1, t_2, \cdots, t_m$ are the extracted components. $\hat{y}_{hi}$ is the fitting value of the ith sample points after using all sample points and regression by taking the components of $t_1 \sim t_h$. $\hat{y}_{h(-i)}$ is the fitting value of y_i computed by the same PLSR after deleting the ith sample points and regression by taking the components of $t_1 \sim t_h$.

$$SS_h = \sum_{t=1}^{n} (y_i - \hat{y}_{hi})^2. \tag{2}$$

$$PRESS_h = \sum_{t=1}^{n} (y_i - \hat{y}_{h(-i)})^2. \tag{3}$$

When Q_h^2 satisfies $Q_h^2 \geq 0.0975$, the new component of t_h can significantly improve the model's prediction ability.

C. VIP

VIP indicates the importance of the independent variable when it is used to interpret the dependent variable, which is defined as follows:

$$VIP_j = \sqrt{P \sum_{h=1}^{m} R(y, t_h) w_{hy}^2 / R(y, t_1, t_2, \cdots, t_m)}. \tag{4}$$

where P is the number of independent variables. m is the number of components extracted from the original variable. t_h represents the hth component. The capability of the component t_h to explain for the dependent variable y is represented by $R(y, t_h)$. W_{hj} is the jth component of the axis W_h. In general, the independent variable plays an important role in the dependent variable when VIP is larger than 1. The effect of the independent variable on the dependent variable is not obvious when the value of VIP ranges from 0.5to1. When the value of VIP is less than 0.5, the independent variable is basically meaningless to the dependent variable. Therefore, the independent variable is eliminated when the value of VIP is less than 0.5.

D. PERFORMANCE INDEXES FOR FITTING

The Relative Error (RE), Root Mean Square Rrror (RMSE), Mean Absolute Percentage Error(MAPE) and Sum of Squared Error (SSE) were used to evaluate the results of fitting. $\hat{y}_i$ and y_i were set as the fitted and the observed value, respectively [39]–[43]. The definitions are as follows:

$$RE_i = \frac{|\hat{y}_i - y_i|}{y_i}, i = 1, 2, \ldots, n. \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}. \tag{6}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|. \tag{7}$$

$$SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2. \tag{8}$$

III. RESULTS AND DISCUSSION

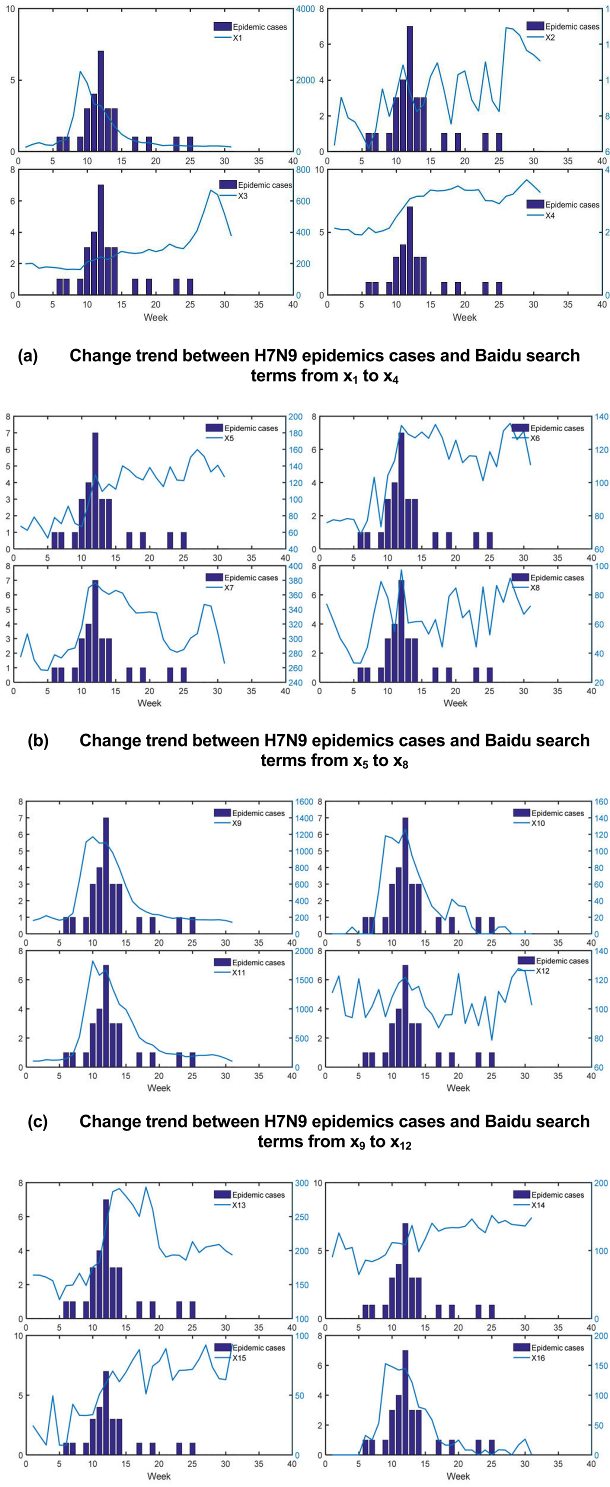
To investigate the change trend between search data and H7N9 infectious diseases, the H7N9 infectious diseases that occurred from January 1, 2017, to July 31, 2017 (intervals of a week and the sum data of a week were taken as the time series), and search data are drawn in Figure 6. As can be seen from the figure, most search data and infectious diseases almost synchronously changed with time, especially in peak number. The search terms, x1,x9,x10,x11,x13, and x16, have the most significant relationship with H7N9 cases.

The relationship between search terms and infectious disease was determined by PCC,SCC, and KCC. The results are shown in Table 2. PCC measures the degree of linear correlation between variables. Non-parametric rank can be measured by SCC and KCC. Moreover, SCC and KCC can judge whether two variables have the same change trend.

TABLE 2. Correlation between search terms and H7N9 epidemics.

Search terms	PCC	SCC	KCC
x_1	0.6181	0.6026	0.4859
x_2	-0.0788	-0.1665	-0.1215
x_3	-0.2436	-0.2959	-0.2371
x_4	0.0096	-0.1797	-0.1242
x_5	-0.0484	-0.1718	-0.1411
x_6	0.2197	0.0764	0.0508
x_7	0.4880	0.3515	0.2879
x_8	0.1767	-0.0739	-0.05933
x_9	0.7738	0.6159	0.5023
x_{10}	0.7702	0.4866	0.4037
x_{11}	0.8063	0.5891	0.4854
x_{12}	0.1756	-0.1350	-0.0989
x_{13}	0.2391	0.1163	0.0904
x_{14}	-0.1779	-0.2449	-0.1863
x_{15}	-0.0278	-0.1429	-0.1158
x_{16}	0.7566	0.5639	0.4651

In Table 2, six words, including H7N9, avian influenza, avian influenza prevention, prevention of avian influenza, avian flu symptoms, and cough, have larger PCC, SCC, and KCC values than others. That is to say, these six terms have higher degree of linear correlation and better change trend with infectious disease than others. However, some words, such as fever and shock, have poor correlation. The reason maybe that common people have public knowledge instead of professional knowledge, so they are more accustomed to using public vocabulary for Web search.



(a) Change trend between H7N9 epidemics cases and Baidu search terms from x_1 to x_4

(b) Change trend between H7N9 epidemics cases and Baidu search terms from x_5 to x_8

(c) Change trend between H7N9 epidemics cases and Baidu search terms from x_9 to x_{12}

(d) Change trend between H7N9 epidemics cases and Baidu search terms from x_{13} to x_{16}

FIGURE 6. Change trend between H7N9 epidemics and search data in Guangxi from the 1st to 31st week in 2017. (Note: The left axis is the number of epidemic cases and the right axis is the number of search term.)

TABLE 3. VIF of search terms.

Search terms	VIP
x_1	49
x_2	6
x_3	7
x_4	19
x_5	18
x_6	20
x_7	8
x_8	4
x_9	118
x_{10}	49
x_{11}	54
x_{12}	6
x_{13}	7
x_{14}	4
x_{15}	9
x_{16}	53

The results of VIF are listed in Table 3. The larger the VIF, the stronger the multi-collinearity. The variables are considered to have multi-collinearity when the size of VIF is >10. As seen in Table 3,the VIF size of eight variables is more than 10. Furthermore, the VIF size of some variables is even more than 100, which indicates strong multi-collinearity among variables. Therefore, the multiple linear regression model is obviously not suitable to fit these circumstances.

As mentioned in Section II, the process will not repeat until the cross validation is less than 0.0975. The cross validity of each principal component is shown in Table 4.

TABLE 4. Cross validity of principal components.

Number of Principal Components	Q_h^2	Critical Value
1	1.0000	0.0975
2	-0.1854	0.0975

In Table 4, the cross validity of the first component is 1. Of course, the first component can obviously improve the prediction effect of the model. However, the cross validity of the second component is -0.1854, which is less than 0.0975.Therefore, one component should be selected. The coefficients were calculated and are listed in Table 5.

The infectious disease and fitting values are shown in Figure 7. To analyze the deviation between the fitting value and the observed value, the RE and residual squares are shown in Figure 8.

We hypothesized that the fewer words used in the Baidu index searches, the better. However, the extracted independent variables t_h must carry the most variable information of X and relate to dependent variable Y as much as possible. Through calculations, the interpretation abilities of the principal component t_h to independent variable X and dependent variable Y were 34.05% and 60.99%, respectively.

TABLE 5. Coefficients of fitting by 16 Search terms.

Search terms	Coefficient
x_1	-0.000128180144025612
x_2	-0.0132684607128705
x_3	-0.000801145827143626
x_4	-0.000514671628893644
x_5	0.00439766820009298
x_6	0.00847721770024412
x_7	0.00535751844388172
x_8	-0.00799418329332998
x_9	0.000773092809466925
x_{10}	0.00607475742198735
x_{11}	0.000736624789732710
x_{12}	0.00905224363237950
x_{13}	0.000771834493711511
x_{14}	-0.00556196902969274
x_{15}	0.000397719488601921
x_{16}	0.00420936443490208
Constant	-1.5043

TABLE 6. VIF of search terms.

Search terms	VIP	Describe
x_1	1.360671609	important
x_2	0.173541498	can remove
x_3	0.53610854	less important
x_4	0.021216289	can remove
x_5	0.106606052	can remove
x_6	0.483518099	unimportant
x_7	1.074115288	important
x_8	0.389026509	unimportant
x_9	1.703122173	important
x_{10}	1.695131575	important
x_{11}	1.774601242	very important
x_{12}	0.386564336	unimportant
x_{13}	0.526376011	less important
x_{14}	0.391478787	unimportant
x_{15}	0.061188684	can remove
x_{16}	1.665402232	important

VIP was used to measure the importance of each independent variable X in explaining the dependent variable Y. The VIP was calculated for each variable and is listed in Table 6.

The size relationship of VIP was ordered as follows: $x_{11} > x_9 > x_{10} > x_{16} > x_1 > x_7 > x_3 > x_{13} > x_6 > x_{14} > x_8 > x_{12} > x_2 > x_5 > x_{15} > x_4$. The independent variable plays an important role in interpreting dependent variables when VIP is larger than 1. As seen in Table 6, x_{11}, x_9, x_{10}, x_{16}, x_1, and x_7 have an important role in interpreting the variable set of Y, especially x_{11}.

To compare the fitting effects of independent variables on dependent variables, three groups of independent variables with VIP>1 (including x_1, x_7, x_9, x_{10}, x_{11}, and x_{16}), VIP>0.5(including x_1, x_3, x_7, x_9, x_{10}, x_{11}, x_{13}, and x_{16}), and VIP>0.3 (including x_1, x_3, x_6, x_7, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, and x_{16}) were selected to fit dependent variables. The extraction of principal components was performed according to the criterion of minimum error.

TABLE 7. Coefficients of fitting by PCAR.

Search terms	Coefficient
x_1	0.0002
x_2	0.0010
x_3	-0.0011
x_4	-0.0001
x_5	-0.0000
x_6	0.0041
x_7	0.0052
x_8	-0.0011
x_9	0.0006
x_{10}	0.0053
x_{11}	0.0004
x_{12}	0.0076
x_{13}	0.0028
x_{14}	-0.0068
x_{15}	0.0008
x_{16}	0.0040
Constant	-2.3525

TABLE 8. Performance index comparison.

Model	RMSE	MAPE	SSE
PLSR6	0.8213	0.0068	20.9106
PLSR8	0.7780	0.0061	18.7639
PLSR11	0.9722	0.0095	29.2999
PLSR16	0.9747	0.0096	29.4491
PCA	0.9696	0.0095	29.1455

In order to compare the results fitted by the proposed method in the paper, Principal Component Analysis (PCA) regression is performed with all independent variables. The coefficient of regression equations are listed in Table 7.

The fitting results obtained are shown in Figure 7, the RE is shown in Figure 8, the performance indexes of fitting are listed in Table 8.

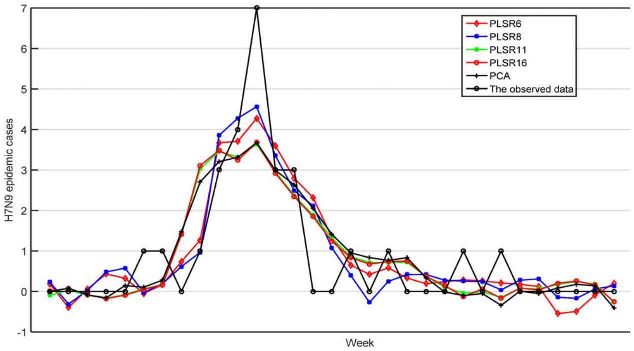


FIGURE 7. Fitting data with different independent variables, PCA method and observed data of H7N9 epidemics in Guangxi from the 1st to 31st week in 2017.

The degree of variation between the fitting and observed data is evaluated by MSE and RMSE. The smaller the values of MSE and RMSE, the smaller the degree of variation, and the closer the fitting are to the observed. As can be seen from table 6, the MSE and RMSE of the proposed method is smaller than that of the other two methods. For a good fitting equation, SSE should be as small as possible. The fitting equation would be have strong interpretation ability when SSE is small. It is obvious from the calculated results that the proposed approach has performance advantages in SSE. And MAPE is a measure of the degree of deviation between each fitting and the observed, which reflects the distribution of the fitting around the observed. The smaller the value, the better the fitting. We also find that MAPE of the proposed method is smaller than that of the other two methods. The size of all performance index were ordered as follows: PLSR8<PLSR6<PLSR11<PLSR16<PCA. That is to say, the model of PLSR8 is the best regression model in this case. Meanwhile, when independent variables exhibit multi-collinearity (VIP>0.5), perfect fitting will be obtained. Therefore, the fitting accuracy is improved by using VIP to select variables.

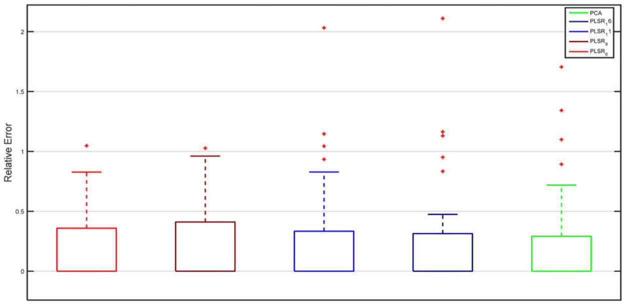


FIGURE 8. The RE comparison using different variables and PCA to fit H7N9 epidemics.

Statistical of the obtained results was investigated by using R,T-test and p-value. The results are listed in Table 9.

TABLE 9. The results of statistical comparison.

Model	R	P-value
PLSR6	0.9713	0.0068
PLSR8	0.9780	0.0061
PLSR11	0.9022	0.0095
PLSR16	0.8747	0.0096
PCA	0.9696	0.0095

R indicates the correlation between the fitting and the observed data. The greater the R value, the better the correlation. The size relationship of R was ordered as follows: PLSR8> PLSR6> PCA> PLSR11> PLSR16. A p-value of < 0.05 was considered significant. $p<0.01$ are for all models obviously, which reveals that the difference is statistically significant between the fitting and the observed.

For a certain independent variable, VIP not only reflects the effect of the variable itself on the dependent variable but also takes into account the influence of other variables on the dependent variable indirectly through this variable.

In general, the independent variable can be eliminated directly if its VIP is very small. If two independent variables have a strong correlation and have a large VIP value, which will have a greater impact on the dependent variables, VIP analysis will need to retain the two variables at the same time. To release the pressure of multi-collinearity, both the R and the professional knowledge should be taken into account simultaneously to judge which variable should be retained in the model. Although the analysis process is more complex than stepwise regression, it is more practical.

IV. CONCLUSION

Factors such as season, climate, and biological effect are considered in previous regression models of H7N9 epidemic disease. In this study, the regression model was based on the network retrieval data by patients. A good result was achieved by using GFT to predict influenza disease. However, people usually use Baidu to search information in China. In this study, the regression using the Baidu index for H7N9 infectious disease was determined.

The technologies of VIP and cross validation were used for independent variable filtering in order to get better fitting value. Results show that PLSR can well fit small sample data, and a good relationship between search data and H7N9 epidemic was observed. All models ultimately serve the profession, and a good model will not be accepted if it does not conform to the actual major. Therefore, the search data through the Baidu index can provide monitoring for the outbreak of H7N9.

Nevertheless, PLSR has some defects. The coefficient of LSR is generally interpreted as the average change of the dependent variable caused by one unit of independent variable change. But the physical meaning of coefficients of PLSR is difficult to explain. PLSR is not applicable when the number of independent variables is small for the cumulative contribution rate of the components to the dependent variable is required to reach 85%. In order to ascertain performance of the proposed method and possible factors that will impact on the model performance in practice, more infectious diseases should be considered in the future.

ACKNOWLEDGMENT

(Ruijing Gan and Jiyong Tan contributed equally to this work.)

REFERENCES

[1] Y. Gu, F. Chen, T. Liu, X. Lv, Z. Shao, H. Lin, C. Liang, W. Zeng, J. Xiao, Y. Zhang, C. Huang, S. Rutherford, and W. Ma, "Early detection of an epidemic erythromelalgia outbreak using Baidu search data," *Sci. Rep.*, vol. 5, no. 1, Oct. 2015, Art. no. 12649.

[2] A. K. Johnson, T. Mikati, and S. D. Mehta, "Examining the themes of STD-related Internet searches to increase specificity of disease forecasting using Internet search terms," *Sci. Rep.*, vol. 6, no. 1, Dec. 2016, Art. no. 36503.

[3] H. A. Carneiro and E. Mylonakis, "Google trends: A Web–based tool for Real–Time surveillance of disease outbreaks," *Clin. Infectious Diseases*, vol. 49, no. 10, pp. 1557–1564, Nov. 2009.

[4] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi, "Assessing Google flu trends performance in the united states during the 2009 influenza virus a (H1N1) pandemic," *PLoS ONE*, vol. 6, no. 8, 2011, Art. no. e23610.

[5] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen, "Reassessing Google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales," *PLoS Comput. Biol.*, vol. 9, no. 10, 2013, Art. no. e1003256.

[6] G. J. Milinovich, G. M. Williams, A. C. A. Clements, and W. Hu, "Internet-based surveillance systems for monitoring emerging infectious diseases," *Lancet Infectious Diseases*, vol. 14, no. 2, pp. 160–168, Feb. 2014.

[7] L. J. Martin, B. Xu, and Y. Yasui, "Improving Google flu trends estimates for the united states through transformation," *PLoS ONE*, vol. 9, no. 12, 2013, Art. no. e109209.

[8] China Internet Network Information Center. (2014). *2013 Chinese Search Engine Market Research Report*. [Online]. Available: http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/ssbg/201401/P020140127366465515288.pdf

[9] K. Liu, T. Wang, Z. Yang, X. Huang, G. J. Milinovich, Y. Lu, Q. Jing, Y. Xia, Z. Zhao, Y. Yang, S. Tong, W. Hu, and J. Lu, "Using Baidu search index to predict dengue outbreak in China," *Sci. Rep.*, vol. 6, no. 1, Dec. 2016, Art. no. 8040.

[10] Z. Li, T. Liu, G. Zhu, H. Lin, Y. Zhang, J. He, A. Deng, Z. Peng, J. Xiao, S. Rutherford, R. Xie, W. Zeng, X. Li, and W. Ma, "Dengue Baidu search index data can improve the prediction of local dengue epidemic: A case study in guangzhou, China," *PLOS Neglected Tropical Diseases*, vol. 11, no. 3, 2017, Art. no. e0005354.

[11] L. Vaughan and Y. Chen, "Data mining from Web search queries: A comparison of Google trends and Baidu index," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 1, pp. 13–22, Jan. 2015.

[12] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein, "Monitoring influenza epidemics in China with search Query from Baidu," *PLoS ONE*, vol. 8, no. 5, 2013, Art. no. e64323.

[13] C. M. Bui, D. C. Adam, E. Njoto, M. Scotch, and C. R. MacIntyre, "Characterising routes of H5N1 and H7N9 spread in China using Bayesian phylogeographical analysis," *Emerg. Microbes Infections*, vol. 7, no. 1, pp. 1–8, Dec. 2018.

[14] T. Yang, F. Li, and F. He, "Semi-quantitative Risk Assessment of Human Infection With H7N9 Avian Influenza Epidemic in Zhejiang Province," *J. Zhejiang Univ.*, vol. 47, no. 2, pp. 131–136, 2018.

[15] V. Virlogeux, L. Feng, T. K. Tsang, H. Jiang, V. J. Fang, Y. Qin, P. Wu, X. Wang, J. Zheng, E. H. Y. Lau, Z. Peng, J. Yang, B. J. Cowling, and H. Yu, "Evaluation of animal-to-human and human-to-human transmission of influenza a (H7N9) virus in China, 2013–15," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 552.

[16] W. Dong, K. Yang, Q. Xu, L. Liu, and J. Chen, "Spatio-temporal pattern analysis for evaluation of the spread of human infections with avian influenza A(H7N9) virus in China, 2013–2014," *BMC Infectious Diseases*, vol. 17, no. 1, p. 704, Dec. 2017.

[17] Y. H. Hu, S. C. Yu, X. Qi, W. J. Zheng, Q. Q. Wang, and H. Y. Yao, "An overview of multiple linear regression model and its application," *Chin. J. Preventive Med.*, vol. 53, no. 6, pp. 653–656, 2019.

[18] T. Britton and D. Ouédraogo, "SEIRS epidemics with disease fatalities in growing populations," *Math. Biosci.*, vol. 296, pp. 45–59, Feb. 2018.

[19] A. Milan, L. Furlanis, F. Cian, R. Bressan, R. Luzzati, C. Lagatolla, M. L. Deiana, A. Knezevich, E. Tonin, and L. Dolzani, "Epidemic dissemination of a carbapenem-resistant acinetobacter baumannii clone carrying armA two years after its first isolation in an italian hospital," *Microbial Drug Resistance*, vol. 22, no. 8, pp. 668–674, Dec. 2016.

[20] R. Gan, N. Chen, and D. Huang, "Comparisons of forecasting for hepatitis in guangxi province, China by using three neural networks models," *PeerJ*, vol. 4, Nov. 2016, Art. no. e2684.

[21] P. Haddawy, A. H. M. I. Hasan, R. Kasantikul, S. Lawpoolsri, P. Sa-Angchai, J. Kaewkungwal, and P. Singhasivanon, "Spatiotemporal Bayesian networks for malaria prediction," *Artif. Intell. Med.*, vol. 84, pp. 127–138, Jan. 2018.

[22] A. J. Tamhankar, S. S. Karnik, and C. S. Lundborg, "Determinants of antibiotic Consumption–development of a model using partial least squares regression based on data from india," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 6421.

[23] C. Koch, A. E. Posch, H. C. Goicoechea, C. Herwig, and B. Lendl, "Multi-analyte quantification in bioprocesses by Fourier-transform-infrared spectroscopy by partial least squares regression and multivariate curve resolution," *Analytica Chim. Acta*, vol. 807, pp. 103–110, Jan. 2014.

[24] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xua, E. Correa, M. L. Turner, and R. Goodacre, "Tutorial review: Metabolomics and partial least squares-discriminant analysis-a marriage of convenience or a shotgun wedding," *Anal Chim Acta.*, vol. 879, pp. 10–23, Jun. 2015.

[25] V. Bonfatti, F. Tiezzi, F. Miglior, and P. Carnier, "Comparison of Bayesian regression models and partial least squares regression for the development of infrared prediction equations," *J. Dairy Sci.*, vol. 100, no. 9, pp. 7306–7319, Sep. 2017.

[26] E. Luedeling and A. Gassner, "Partial least squares regression for analyzing walnut phenology in california," *Agricult. Forest Meteorol.*, vols. 158–159, pp. 43–52, Jun. 2012.

[27] M. D. Dyar, M. L. Carmosino, E. A. Breves, M. V. Ozanne, S. M. Clegg, and R. C. Wiens, "Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples," *Spectrochimica Acta Part B, At. Spectrosc.*, vol. 70, pp. 51–67, Apr. 2012.

[28] S. R. Karunathilaka, B. J. Yakes, S. Farris, T. J. Michael, K. He, J. K. Chung, R. Shah, and M. M. Mossoba, "Quantitation of saccharin and cyclamate in tabletop formulations by portable Raman and NIR spectrometers in combination with partial least squares regression," *Food Anal. Methods*, vol. 11, no. 4, pp. 969–979, Apr. 2018.

[29] C. J. Meunier, E. C. Mitchell, J. G. Roberts, J. V. Toups, G. S. McCarty, and L. A. Sombers, "Electrochemical selectivity achieved using a double voltammetric waveform and partial least squares regression: Differentiating endogenous hydrogen peroxide fluctuations from shifts in pH," *Anal. Chem.*, vol. 90, no. 3, pp. 1767–1776, Feb. 2018.

[30] Z. Zhu, J. Li, Y. Guo, X. Cheng, Y. Tang, L. Guo, X. Li, Y. Lu, and X. Zeng, "Accuracy improvement of boron by molecular emission with a genetic algorithm and partial least squares regression model in laser-induced breakdown spectroscopy," *J. Anal. At. Spectrometry*, vol. 33, no. 2, pp. 205–209, 2018.

[31] S. Mostafapour and H. Parastar, "N-way partial least squares with variable importance in projection combined to GC×GC-TOFMS as a reliable tool for toxicity identification of fresh and weathered crude oils," *Anal. Bioanal. Chem.*, vol. 407, no. 1, pp. 285–295, Jan. 2015.

[32] Z. Zhang and G. S. Feng, "Application of variable importance for projection in the variables selection," *Prev. Med.*, vol. 39, no. 22, pp. 5813–5815, 2012.

[33] *Guangxi Zhuang Autonomous Region*. [Online]. Available: https://baike.baidu.com/item/%E5%B9%BF%E8%A5%BF/162679?fromtitle=%E5%B9%BF%E8%A5%BF%E5%A3%AE%E6%97%8F%E8%87%AA%E6%B2%BB%E5%8C%BA&fromid=163178&fr=aladdin

[34] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A new sampling method in particle filter based on pearson correlation coefficient," *Neurocomputing*, vol. 216, pp. 208–215, Dec. 2016.

[35] Y. Duan and C. Song, "Relevant modes selection method based on spearman correlation coefficient for laser signal denoising using empirical mode decomposition," *Opt. Rev.*, vol. 23, no. 6, pp. 936–949, Dec. 2016.

[36] P. Grzegorzewski, "Kendall's correlation coefficient for vague preferences," *Soft Comput.*, vol. 13, no. 11, pp. 1055–1061, 2009.

[37] R. Salmerón Gómez, J. García Pérez, M. D. M. López Martín, and C. G. García, "Collinearity diagnostic applied in ridge estimation through the variance inflation factor," *J. Appl. Stat.*, vol. 43, no. 10, pp. 1831–1849, Jul. 2016.

[38] K. P. Singh, S. Gupta, A. Kumar, and S. P. Shukla, "Linear and nonlinear modeling approaches for urban air quality prediction," *Sci. Total Environ.*, vol. 426, pp. 244–255, Jun. 2012.

[39] W.-C. Hong, M.-W. Li, J. Geng, and Y. Zhang, "Novel chaotic bat algorithm for forecasting complex motion of floating platforms," *Appl. Math. Model.*, vol. 72, pp. 425–443, Aug. 2019.

[40] Z. Zhang and W.-C. Hong, "Electric load forecasting by complete ensemble empirical mode decomposition adaptive noise and support vector regression with quantum-based dragonfly algorithm," *Nonlinear Dyn.*, vol. 98, no. 2, pp. 1107–1136, Oct. 2019.

[41] Y. Dong, Z. Zhang, and W.-C. Hong, "A hybrid seasonal mechanism with a chaotic cuckoo search algorithm with a support vector regression model for electric load forecasting," *Energies*, vol. 11, no. 4, p. 1009, 2018.

[42] H. Kundra and H. Sadawarti, "Hybrid algorithm of cuckoo search and particle swarm optimization for natural terrain feature extraction," *Res. J. Inf. Technol.*, vol. 7, no. 1, pp. 58–69, Jan. 2015.

[43] W. Song, C. Cattani, and C.-H. Chi, "Multifractional brownian motion and quantum-behaved particle swarm optimization for short term power load forecasting: An integrated approach," *Energy*, vol. 194, Mar. 2020, Art. no. 116847.

RUIJING GAN received the bachelor's and master's degrees major in clinical medicine from Guangxi Medical University, China. She is currently an Associate Professor with Guangxi Medical University. Her research interests include epidemiological prediction and immunology.

JIYONG TAN received the bachelor's, master's, and Ph.D. degrees in epidemiology from Guangxi Medical University, China. His current research interest includes epidemiological prediction.

LIYING MO received the bachelor's degree major in biotechnology from Guangxi Medical University, China. She is currently pursuing the master's degree.

YU LI received the bachelor's degree major in computer information management from Guangxi Medical University, China. She is currently pursuing the master's degree.

DAIZHENG HUANG received the M.S. degree in optical engineering from the Huazhong University of Science and Technology, in 2006, and the Ph.D. degree in electric power system and automation from Guangxi University, in 2015. He is currently a Professor with the Department of Biomedical Engineering. His research interest includes artificial intelligence algorithm.