

Construction of Influenza Early Warning Model Based on Combinatorial Judgment Classifier: A Case Study of Seasonal Influenza in Hong Kong*

Zi-xiao WANG^{1,2,3}, James NTAMBARA⁴, Yan LU¹, Wei DAI¹, Rui-jun MENG¹, Dan-min QIAN^{1,5#}

¹Department of Medical Informatics, School of Medicine, Nantong University, Nantong 226001, China

²Department of Computer Science, College of Engineering and Computing Sciences, New York Institute of Technology, New York 10023, USA

³Department of Computer Science, College of Overseas Education, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

⁴Department of Epidemiology, School of Public Health, Nantong University, Nantong 226019, China

⁵Artificial Intelligence Laboratory Center, De Montfort University of Leicester, Leicester LE1 9BH, United Kingdom

© Huazhong University of Science and Technology 2022

[Abstract] Objective: The annual influenza epidemic is a heavy burden on the health care system, and has increasingly become a major public health problem in some areas, such as Hong Kong (China). Therefore, based on a variety of machine learning methods, and considering the seasonal influenza in Hong Kong, the study aims to establish a Combinatorial Judgment Classifier (CJC) model to classify the epidemic trend and improve the accuracy of influenza epidemic early warning. **Methods:** The characteristic variables were selected using the single-factor statistical method to establish the influencing factor system of an influenza outbreak. On this basis, the CJC model was proposed to provide an early warning for an influenza outbreak. The characteristic variables in the final model included atmospheric pressure, absolute maximum temperature, mean temperature, absolute minimum temperature, mean dew point temperature, the number of positive detections of seasonal influenza viruses, the positive percentage among all respiratory specimens, and the admission rates in public hospitals with a principal diagnosis of influenza. **Results:** The accuracy of the CJC model for the influenza outbreak trend reached 96.47%, the sensitivity and specificity change rates of this model were lower than those of other models. Hence, the CJC model has a more stable prediction performance. In the present study, the epidemic situation and meteorological data of Hong Kong in recent years were used as the research objects for the construction of the model index system, and a lag correlation was found between the influencing factors and influenza outbreak. However, some potential risk factors, such as geographical nature and human factors, were not incorporated, which ideally affected the prediction performance to some extent. **Conclusion:** In general, the CJC model exhibits a statistically better performance, when compared to some classical early warning algorithms, such as Support Vector Machine, Discriminant Analysis, and Ensemble Classifiers, which improves the performance of the early warning of seasonal influenza. **Key words:** influenza prediction; data-driven; Support Vector Machine; Discriminant Analysis; Ensemble Classifier

Annual influenza epidemics are a burden to public health due to the characteristics of high infectivity, high incidence and sudden onset, and the impact on the socio-economic and organizational aspects of life^[1-3]. Therefore, timely information on the impending epidemic is essential both for optimizing work

organization and drug accumulation, in order to develop new strategies to control the disease^[4]. There has been an emergence of various global outbreaks, such as Spanish flu, the United States “swine flu”, the Russian flu, and influenza^[5-8]. China has had several outbreaks of influenza, including the H1N1 flu pandemic in 2009^[9, 10] and the H7N9 flu epidemic in 2013^[11, 12]. These epidemics have caused great losses to national life and economy, despite the preventive efforts, and influenza epidemics are accountable for the substantial morbidity and mortality in China. In early 2019, there was an outbreak of winter influenza in Hong Kong, China, which lasted for approximately 14 weeks. A total of 357 deaths were recorded. The scale of the onset and intensity of the spread of influenza were larger than

Zi-xiao WANG, E-mail: zwang62@nyit.edu

#Corresponding author, E-mail: qdm11@163.com

*This project was supported by grants from the Ministry of Education Humanities and Social Sciences Research Fund Project (No. 17YJCZH140), the Jiangsu Philosophy and Social Sciences Fund Project (No. 18SHB004), and the Jiangsu University Philosophy and Social Sciences Research Fund Project (No. 2017SJB1211).

those of previous influenza epidemics^[13]. Thus, it is imperative to build an early warning model of influenza with high accuracy, and appropriate interventions should be made, in order to prevent or at least prepare for catastrophic epidemics^[14]. The seasonal outbreaks of influenza infection can cause respiratory illness globally or even death for all age groups^[15]. Hence, the present study constructed an early warning model of influenza based on the influenza epidemic situation in Hong Kong in the past eight years. Vaccination and other possible precautionary strategies to reduce disease burden have been implemented^[16]. However, the establishment of early detection and warning systems is a crucial step to set up effective control measures, and combat upcoming influenza-related epidemics^[17, 18]. These systems rely primarily on reliable and timely sources of data. In previous years, electronically and routinely collected data has emerged as a convenient source of surveillance data^[19, 20]. Monitoring systems worldwide play a vital role in supporting data-driven guidelines and principles. Accordingly, based on the data-driven principle, it has become an effective means to thoroughly analyze the trend of an epidemic by thoroughly studying the internal relationship between meteorological factors and the spread of the influenza virus^[21, 22]. The machine learning algorithm, such as the support machine vector (SVM) and neural network, is an effective tool for analyzing these data, and improving the accuracy of the prediction^[23, 24].

In the present study, MATLAB and R were used for the statistical analysis and calculations. Based on descriptive statistical analysis methods, the development trend of influenza epidemics under the combined influence of several meteorological factors was presented, and a common assessment classifier was constructed as a benchmark for targeted epidemic prevention.

Various studies have established concepts for the detection and clinical prediction of influenza using different factors^[22, 25–27]. A predictive influenza model in Beijing was built based on meteorological and influenza virus activity, and confirmed the correlation between the positivity rate of each influenza virus subtype for the previous two weeks, and weather factors and influenza-like illnesses (ILI) per week^[28]. The effects of climate parameters on the modeling and forecasting of seasonal influenza transmission were investigated in Abidjan, Cote d'Ivoire^[29]. Furthermore, Soebiyanto *et al* modeled influenza cases in Hong Kong using land surface temperature (LST), precipitation, and relative humidity as covariables^[30]. In addition, the ARIMA model and Holt-Winters model were discussed and compared in the prediction of influenza in the context of Wuhan, which provided a scientific basis for the prevention and control of influenza^[31]. On other hand, the Maxent ecologic niche model was

used to study the relationship between H7N9 infectious diseases and environmental factors, and predict the potential risk area of the avian influenza epidemic^[32]. Emerging biomarker prediction methods are effective for influenza outbreak prediction, such as that in the study conducted by Gao *et al*, in which protein dynamic network biomarkers were used to predict an influenza pandemic outbreak^[33]. In addition, Arturo López Pineda *et al* compared the accuracy of different algorithms for influenza detection, and revealed that logistic regression, the Artificial Neural Network (ANN), and SVM have relatively high accuracy^[34].

Derived from this research, it remains difficult to achieve appropriate precision using the unique linear regression model for predicting ILI^[20]. The connection between climate factors and influenza also does not provide a systematic warning for epidemics at present. Extensive research into the pathogen is necessary to obtain an accurate prediction on an epidemic scale. However, due to the complexity between biological and social systems, there are higher demands for scientific research conditions^[15]. Despite these considerable research efforts, no reliable disease model has been developed to accurately detect early-warning signals^[35, 36]. Therefore, the investigators proposed a classification method to predict the development trend of an epidemic situation. Factors for the influenza epidemic were determined based on the influenza statistics recorded by the Department of Health of Hong Kong Center for Health Protection. In order to improve the accuracy of the prediction, the Hong Kong seasonal influenza early warning model was developed, which provides benchmarks for the prevention of epidemics.

1 MATERIALS AND METHODS

1.1 Data Collection

The data were collected based on the National Influenza Surveillance Plan (2017 version) (hereinafter referred to as “the Plan”), in order to set the requirements for the influenza epidemic scale. Combined with the Hong Kong Department of Health’s description of the epidemic intensity of influenza outbreaks (Moving Epidemic Method, MEM), two influenza development trends were defined, namely, recession and outbreak. Data obtained from the 1st week of 2014 to the 33rd week of 2021 (December 29, 2013, to August 14, 2021) was used to investigate the development trend of influenza.

The influenza data for Hong Kong, which was released by the Center for Health Protection of the Department of Health of the Government of the Hong Kong Special Administrative Region in 2021, and contains the weekly ILI, the admission rate of patients diagnosed with influenza in public hospitals, and the number of positive samples of seasonal influenza

virus, was retrieved per day (<https://www.chp.gov.hk/sc/resources/29/304.html>). The weather data were obtained from the Hong Kong Observatory website (<https://www.hko.gov.hk/sc/wxinfo/pastwx/mws/mws.html>). This included the average atmosphere pressure, air temperature, dew point temperature, humidity cloud amount and rainfall.

1.2 Statistical Analysis

The research process is presented in fig. 1. SPSS 26.0 was used for the statistical analysis, and one-way ANOVA was performed on independent samples from the training set. Ultimately, high correlation variables ($P < 0.05$) were selected as characteristic variables to improve the prediction performance of the model. The meteorological data were processed by week, and all data were normalized and dimensioned. Then, MATLAB was used to construct the influenza early warning model. Several classical algorithms for modeling were selected, including non-ensemble algorithms, such as Discriminant Analysis, SVM and Ensemble Classifier. Next, the combinatorial judgment prediction method was introduced to determine the

final classification model. Finally, MATLAB was used to compare the recall of the combined judgment model and other general classifiers, in order to facilitate the evaluation of each model.

2 RESULTS

2.1 Construction of Early Warning Model of Influenza

2.1.1 Research Objects The meteorological and epidemic scale of influenza in Hong Kong was selected as the research object during the period from the earliest recorded date (December 29, 2013) to the present year (August 14, 2021). These data were divided by week (seven days), with a total of 397 groups.

2.1.2 Research Variables Based on the existing literature and index system established by domestic and foreign scholars, the index system was supplemented and improved by combining this with the expert interview method, and the influencing factors of the disease were summed up to build the index system of influencing factors for the outbreak trend of influenza (table 1).

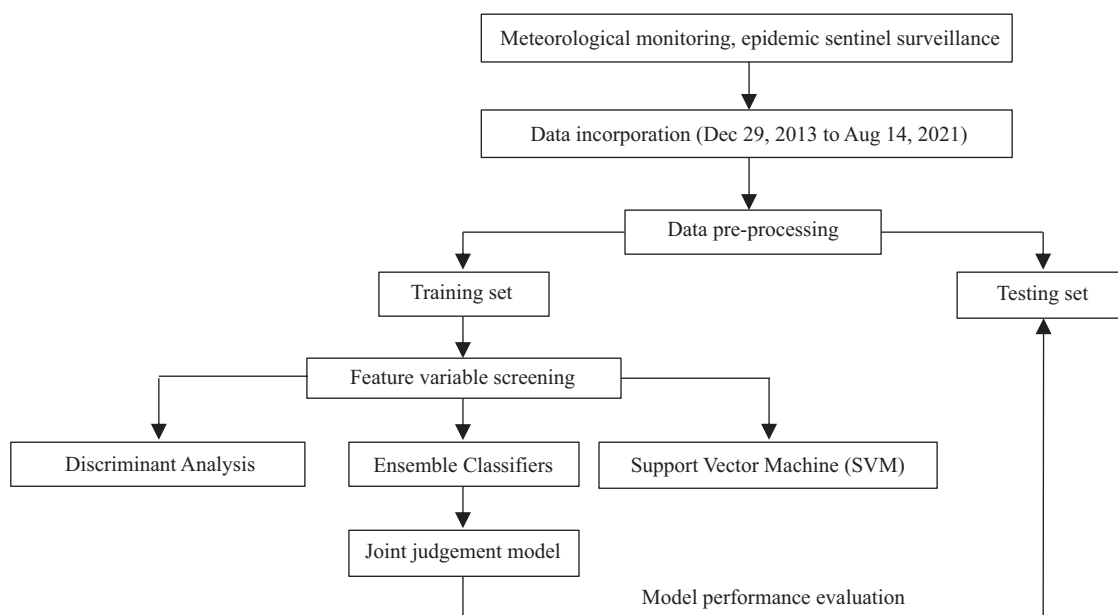


Fig. 1 Flow chart for the research

The influenza prediction model based on multiple machine learning methods

Table 1 Index system table of influencing factors for the influenza outbreak trend

Level 1	Level 2	Level 3
Meteorological monitoring U1	Atmosphere pressure U11	Average pressure U111
	Air temperature U12	Absolute maximum temperature U121
		Average temperature U122
		Absolute minimum temperature U123
		Average dew point temperature U124
Epidemic situation monitoring U2	Relative humidity U13	Average relative humidity U131
	Cloud amount U14	Average cloud amount U141
	Rainfall U15	Total rainfall U151
	Laboratory surveillance U21	Number of positive detections of seasonal influenza viruses U211
	Hospital spot detection U22	Positive percentage among all respiratory specimens U212
	Admission rates in public hospitals with principal diagnosis of influenza U221	

CDC: Chinese Center for Disease Control and Prevention

2.1.3 Outcome Variable According to “the Plan”, when 30 or more influenza-like cases occur in the same school, kindergarten, or other collective units within one week, these should be reported through the public health emergency management information system within two hours. Combined with the MEM influenza outbreak intensity level adopted by the Department of Health in Hong Kong, the categories of recession (0) and outbreak (1) were defined. Weekly ILI outbreaks of ≥ 30 were set as 1, while other outbreaks (< 30) were set as 0, in order to obtain the outbreak trend data from 2014 to the 33rd week of 2021.

2.1.4 Characteristic Variable Screening After the one-way ANOVA analysis, some variables were retained, including the average atmosphere pressure, absolute maximum air temperature, mean air temperature, absolute minimum air temperature, mean dew point temperature, the number of positive detections of seasonal influenza viruses, the positive percentage among all respiratory specimens, and the admission rates in public hospitals with principal diagnosis of influenza. All of these were recorded in the previous week.

2.1.5 Data Preprocessing Meteorological data were averaged by week. According to the established index system, the weekly meteorological data was investigated. Considering that the meteorological data were in stable fluctuation in the short term, EXCEL was used to take the weekly value of the meteorological data. The mapminmax function in MATLAB was used to normalize the data and process scales.

2.1.6 Construction of Classification Prediction Model and Prediction Results

Selection of the classification model Common machine classification and prediction models were established. According to the classification results, Discriminant Analysis, SVM and Ensemble Classifier were selected, as well as the commonly used influenza prediction model—the Autoregressive Integrated Moving Average model (ARIMA)^[37–39].

Algorithm implementation Initially, the preprocessed data set in MATLAB were imported, and the “Classification Learner” application was selected. Then, “SVM”, “Ensemble Classifier” and “Discriminant Analysis” were checked for training, Holdout Validation was adopted for verification, and the percentage setting was set to 40%. Finally, the best subclass was selected for each model to represent the performance of the model. The KernelFunction of the best sub-model was “polynomial”, and Solver was “SMO”. The DiscrimType of the best sub-model for Discriminant Analysis was “linear”, and priors=(0.904, 0.096). The method of the best sub-model for Ensemble Classifier was “Subspace”, and the model parameters were NLearn=30 and LearnRate=1.

Joint decision classification model Based on

the training data, the classification prediction model was established, and the ratio for the training data and testing data was set to 6:4. Overall, the prediction accuracy of these classifiers was independently evaluated. Considering that there are advantages and disadvantages for every single model, the present study selected three models with the highest accuracy, mixed these models, and constructed a CJC.

Feature selection Variables were initially included after the one-way ANOVA. The analysis included the following: mean atmospheric pressure, absolute maximum temperature, mean temperature, absolute minimum temperature, mean dew point temperature, several positive detections of seasonal influenza viruses, the positive percentage among all respiratory specimens, and admission rates in public hospitals with a principal diagnosis of influenza.

1) Regression and classification based on the Discriminant Analysis. The Discriminant Analysis model was trained, and the linear discriminant (Model 1.1) and quadratic discriminant (Model 1.2) sub-models were established, respectively (fig. 2). The accuracy of model 1.1 on the validation set was 95.6%, while that of Model 1.2 was 92.4%. At the same time, the verification of the confusion matrix revealed that Model 1.2 had a higher false positive rate, and was more likely to produce a positive misjudgment of the outbreak trend. In addition, the receiver operating characteristic (ROC) curve revealed that model 1.1 was superior to model 1.2, in terms of classification effect. Therefore, based on accuracy and sensitivity, linear discrimination (Model 1.1) was chosen for the subsequent prediction analysis.

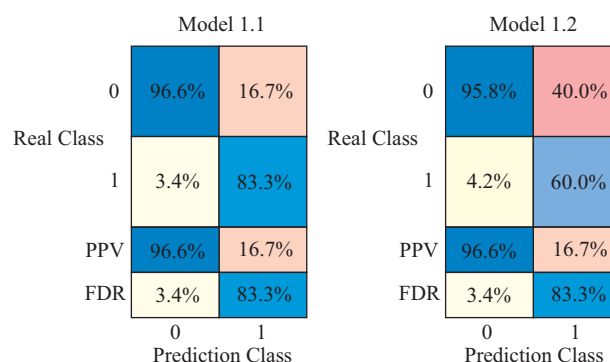


Fig. 2 Discriminant Analysis result
Confusion matrices were built to evaluate the performance of the Discriminant Analysis

2) Selection of ensemble classifiers. The data sets are trained with integrated classification algorithms to obtain the performance of different integrated classification algorithms (fig. 3). The accuracy of model 2.1 (Boosting Tree), model 2.2 (Bagged Tree), model 2.3 (Subspace Discrimination), model 2.4 (Subspace KNN) and model 2.5 (RUSBoosted Tree) validation

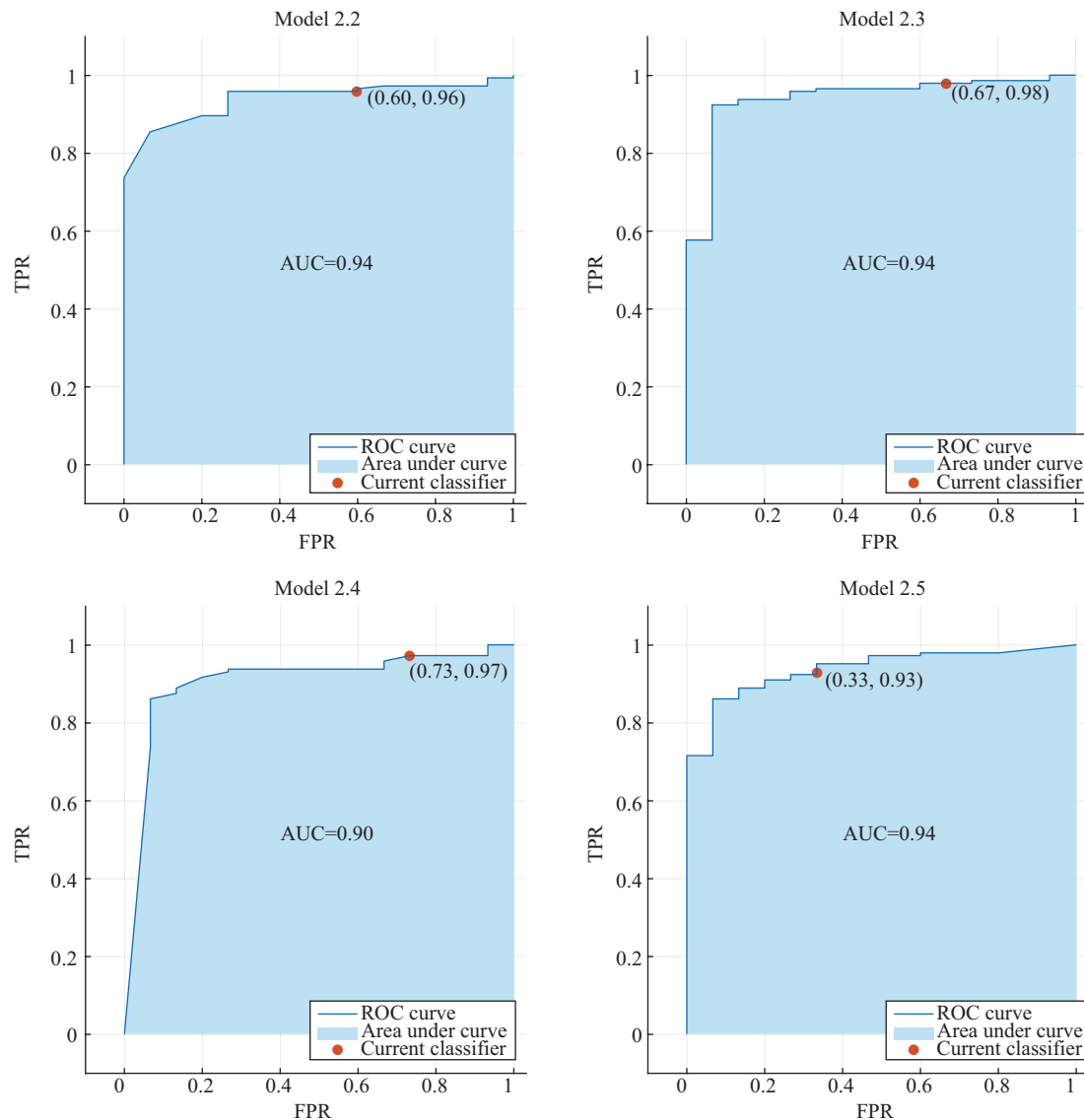


Fig. 3 The receiver operating characteristic (ROC) curve based on the ensemble classification. Based on the area under the curve (AUC), the Bagged Tree, Subspace Discriminant, Subspace KNN and RUSBoosted Tree were described, respectively.

sets was 92%, 94%, 94%, 90% and 94%, respectively. The Subspace Discriminant (Model 2.3) with better training performance was selected as the component of the classifier by integrating the two evaluation indexes of accuracy and area under the ROC curve (AUC).

3) Regression and classification based on SVM. A variety of SVM functions were set up (table 2), the classification performance of the SVM models under different functions on the training set was compared, functions with high accuracy and recall were selected, the SVM model training was conducted, and the quadratic SVM models were exported.

4) CJC-based epidemic development prediction method. In response to the prediction of the trends of influenza, the present study adopted a stable classifier named, CJC. Through a variety of model validations, the data before and after the outbreak was selected for testing. The accuracy of the Discriminant Analysis,

Table 2 SVM classification performance

SVM	Accuracy	Precision	Recall	AUC
Liner	95.6%	83.33%	66.67%	0.96
Quadratic	95.6%	78.57%	73.33%	0.95
Cubic	93.0%	62.5%	66.67%	0.93
Fine Gaussian	90.5%	0	0	0.92
Medium Gaussian	92.40%	63.64%	46.67%	0.92
Coarse Gaussian	93.70%	77.78%	46.67%	0.96

AUC: area under the receiver operating characteristic curve

Ensemble Classifier and SVM in the test sets was higher than 90%, showing that these have good consistent judgment on the trend of an influenza outbreak in two states. Building on these, a combined prediction method of epidemic development was established. The CJC classification results are presented in table 3, the classification result for CJC was based on the most frequent value of the above three methods.

Table 3 CJC classification results

Classification model	Classification results of each classifier (0/1)							
	0	1	0	0	1	1	0	1
Subspace Discriminant	0	1	0	0	1	1	0	1
Ensemble Classifiers	0	0	1	0	1	0	1	1
SVM	0	0	0	1	0	1	1	1
CJC	0	0	0	0	1	1	1	1

In the binary model, the classification result would be 0 or 1. The CJC model will select the majority of the above three classifiers as its result.

2.2 Classification Model Prediction Performance Evaluation

The commonly used evaluation indexes for dichotomy are accuracy and recall. The experiment introduced accuracy and specificity to evaluate the performance of the model. The test set data was used to objectively evaluate the accuracy of the CJC, in which the outbreak trend was set as a positive category and the recession trend was set as a negative category. The accuracy and recall were calculated by the confusion matrix, in order to measure the prediction performance of the model (fig. 4). Among the 397 samples in the training set, there were 38 positive samples and 359 negative samples. Among these, there were 27 true positive samples (TP), 11 false-negative samples (FN), three false-positive samples (FP), and 356 true negative samples (TN). The accuracy, precision, recall and specificity were 96.47%, 90.0%, 71.05% and 99.16%, respectively.

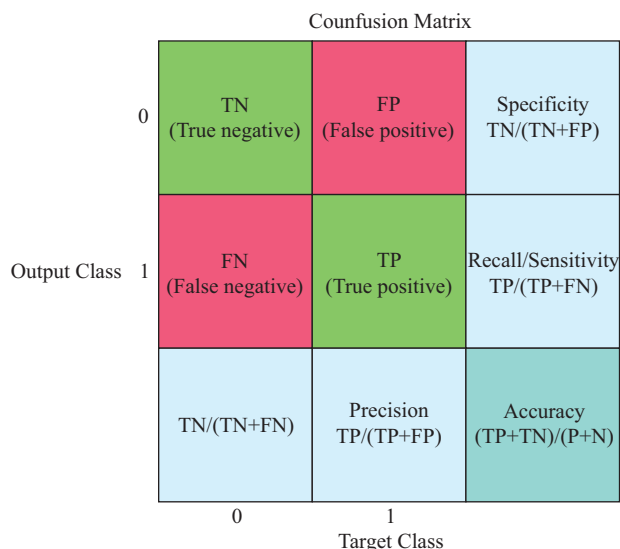


Fig. 4 Confusion matrix of the binary classification
The performance of the classifier was determined by the above nine indicators.

2.3 Comparative Experiment

2.3.1 Forecast Results of Commonly Used Influenza Prediction Models

ARIMA is a commonly used influenza prediction model. The trend, seasonal and random deviations of the time series of ILI outbreaks were decomposed from the 1st week of 2014 to the 33rd week of 2021 using the “decompose” function in the R language (the decomposition result is presented in fig. 5).

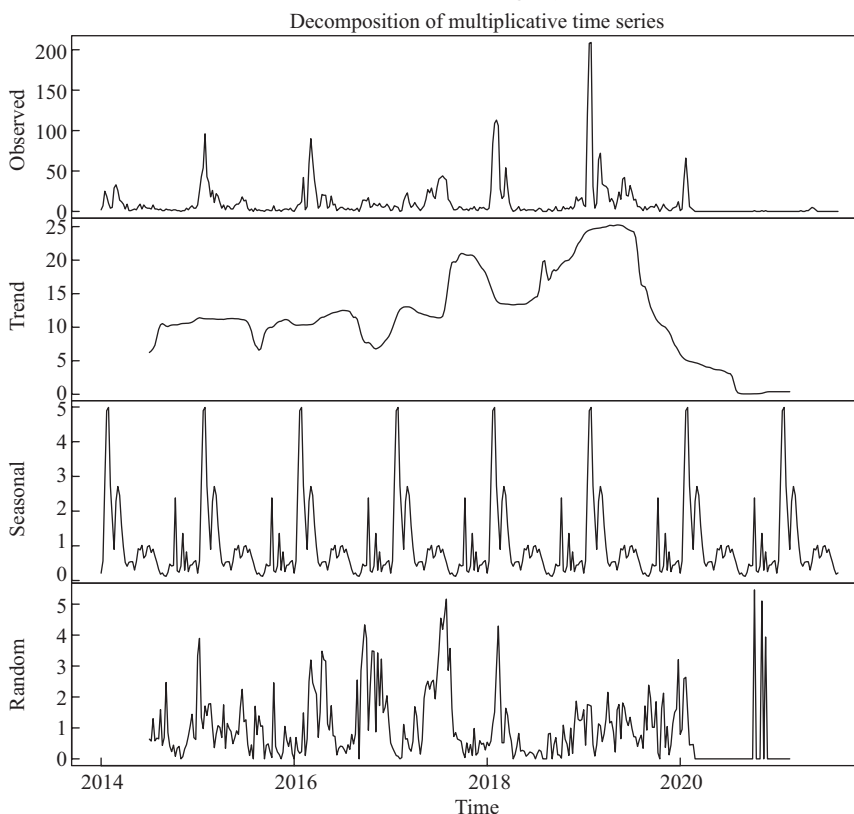


Fig. 5 Decomposition of influenza trend, seasonality and random errors
The decompose function in the R language was used to perform the time series analysis from the first week of 2014 to the 33rd week of 2021.

The number of ILI outbreaks in schools/institutions had obvious seasonal characteristics. The trend decreased in 2015 and 2016, but quickly resumed and reached a peak in 2018–2019, indicating that this series is a non-stationary time series, and that the random error surged in the high influenza season. Although the ARIMA model exhibited a good performance on epidemic scale fitting, the accuracy was far from sufficient (figs. 6 and 7).

When using ARIMA to predict the number of ILI outbreaks, it was found that although the ARIMA model fits the prediction accuracy, this was not ideal for the winter influenza prediction in 2019. Furthermore, although the ARIMA classification reduced the requirements of training data, and considering that external factors greatly fluctuate during the influenza season, other factors (weather, etc.) can accelerate or restrain the spread of influenza. Furthermore, a prediction method based on a single indicator (ILI outbreak number of schools/institutions) is not completely suitable for seasonal influenza under climatic fluctuations, and the use of impact indicators provides a theoretical basis for the influenza outbreak

status. Compared to the research methods of general small indicators, the prediction performance obtained based on the meteorological conditions and fixed-point surveillance of influenza was excellent.

2.3.2 Comparison of Results with Common Machine Learning Classification Algorithms Many algorithms used for disease surveillance are well-established. However, each method has some context, and is disease-specific^[40]. The reason is due to the differences in surveillance purpose, the disease’s epidemiologic features, or the approach in calculating the alarm threshold. In this case, extending multiple models for integrated calculation proved to be a fast and effective forecasting method^[41]. With that said, in comparison with common machine learning classification algorithms, the best sub-category in each model was initially selected, and the model prediction performance was quantified through the evaluation of the classification model. Then, the prediction performance of the three methods (SVM, Ensemble Classifier, and Discriminant Analysis) and CJC was compared, and the model prediction accuracy was calculated, according to the test set (tables 4 and 5).

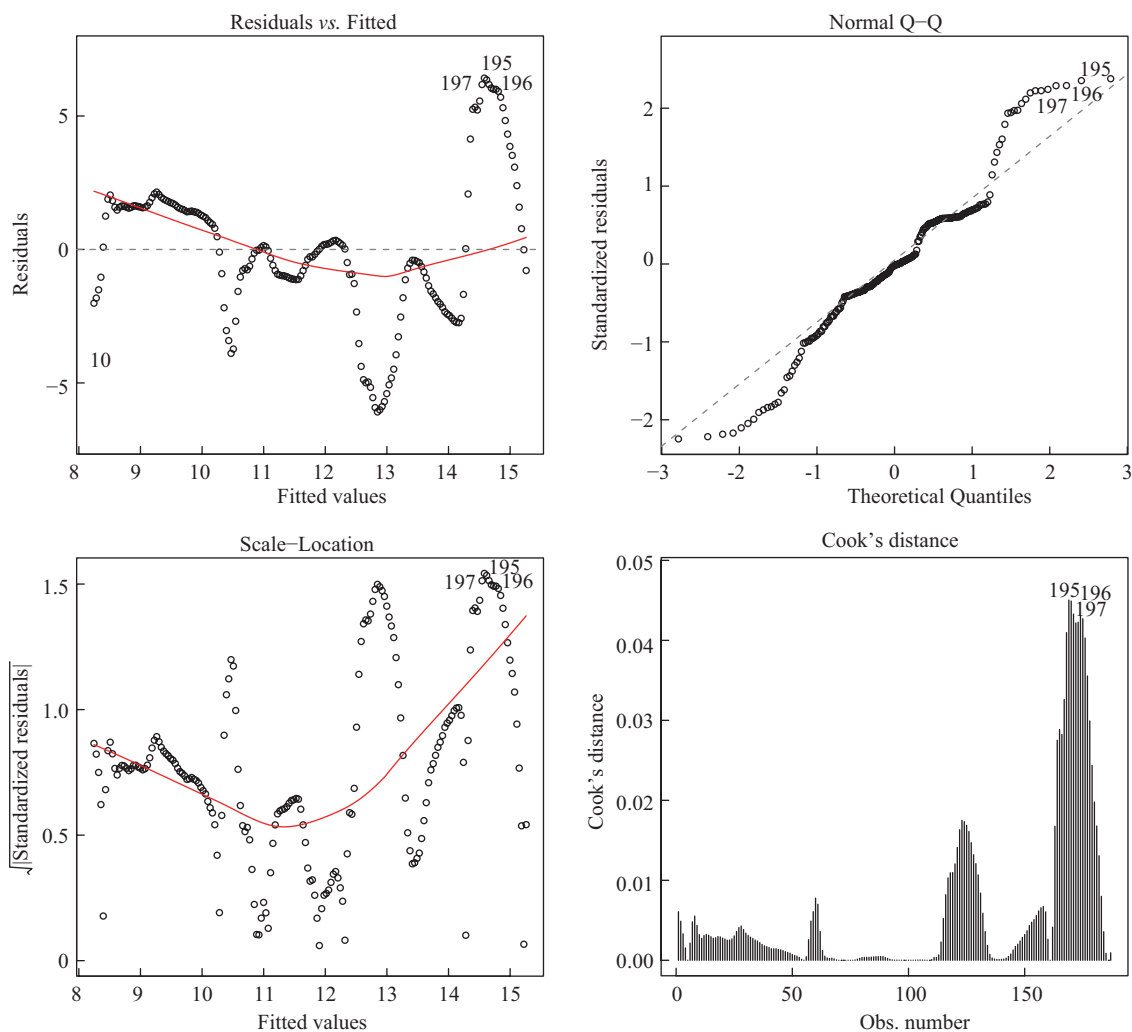


Fig. 6 Statistical graph for the regression model fitting results
The trending term in the model was extracted and fitted with a linear model.

Table 4 Training performance of multiple classification algorithms

Algorithm	Training set prediction performance			
	Classification accuracy (A1)	Accuracy (P1)	Recall/sensitivity (RS1)	Specificity (SP1)
SVM	94.96%	73.91%	73.91%	97.21%
Discriminant analysis	96.22%	81.82%	78.26%	98.14%
Ensemble classifier	96.64%	85.71%	78.26%	98.60%
CJC	96.64%	85.71%	78.26%	98.60%

Table 5 Testing performance of multiple classification algorithms

Algorithm	Testing set prediction performance				
	Classification accuracy (A2)	Accuracy (P2)	Recall/sensitivity (RS2)	Specificity (SP2)	F1-score
SVM	94.97%	81.82%	60.00%	98.61%	53.00%
Discriminant analysis	96.23%	100.00%	60.00%	100.00%	75.00%
Ensemble classifier	95.60%	90.00%	60.00%	99.31%	52.00%
CJC	96.23%	100.00%	60.00%	100.00%	75.00%

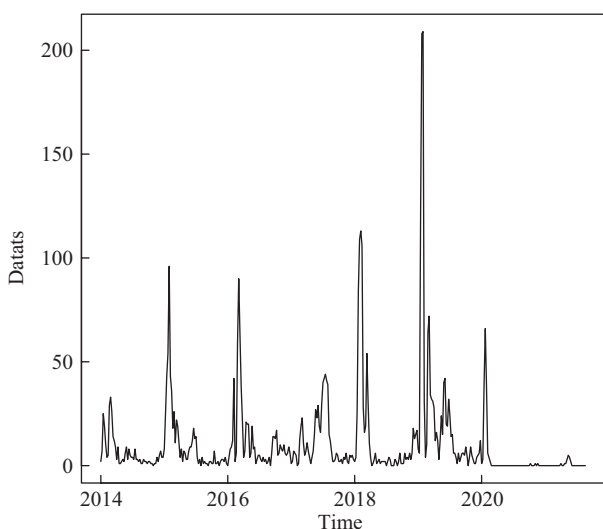


Fig. 7 The actual scale of the pandemic
The scale of the outbreak from 2014 to the 33rd week of 2021 (number of outbreaks of influenza-like illness in school buildings)

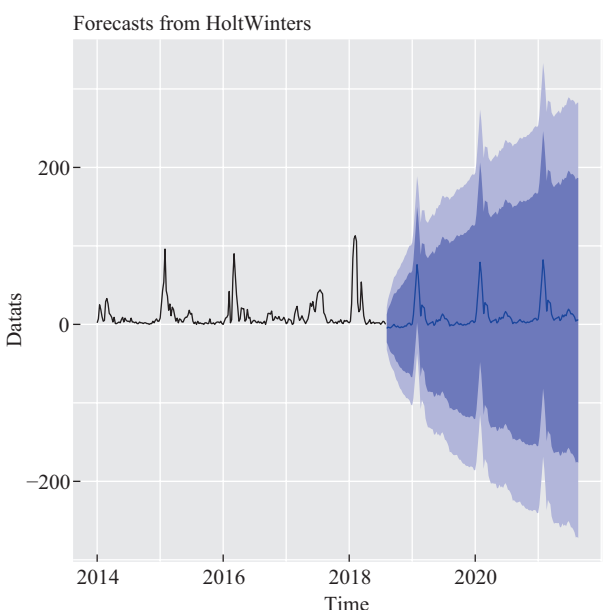


Fig. 8 Diagram for the ARIMA model fitting to the epidemic scale
The fitted value is close to the true value, but the predicted value is far below the expectations in 2019.

By comprehensively considering the single performance value and performance ratio value, the predicting performance of the four algorithms was compared. In tables 4 and 5, the single performance value of CJC performed better than that of the Discriminant Analysis on the training sets. This had the same effect on test sets for Ensemble Classifier, but was slightly lower than that for SVM. However, from the perspective of performance ratio values, and according to formula 1 and 2, the SVM sensitivity change rate was 0.28, and the SVM specificity change rate was 0.01. Furthermore, the CJC sensitivity change rate was 0.23, and the CJC specificity change rate was 0.01. These indicate that the prediction performance of CJC was relatively stable.

$$\text{Sensitivity change rate} = \frac{[RS1-RS2]}{RS2} \quad (\text{Formula 1})$$

$$\text{Specificity change rate} = \frac{[SP1-SP2]}{SP2} \quad (\text{Formula 2})$$

Briefly, the CJC classification has been statistically proven to have better prediction performance, when compared to SVM, Discriminant Analysis, and Ensemble Classifier. This can properly predict upcoming influenza epidemics, both in the long term and short term, providing early and close warning alarms for the start of epidemic periods.

3 DISCUSSION

Annual influenza epidemics place a heavy burden on the health system, and have become a major public health problem in some regions, such as Hong Kong (China)^[40, 42]. Based on the ILI counts of various numbered seasons, a predictive model was created, taking into account the annual seasonal trend. This was found to have predictive ability for a week and a season in advance. The early detection of epidemics is a key element in preventing loss of quality of life, and major economic and health effects. Therefore, the combinatorial judgment prediction method was introduced to determine the final classification

model that could improve prediction performance, and complement other developed models, in order to improve accuracy, precision, sensitivity and specificity^[43]. By performing these, several classical algorithms were carefully chosen for the development, including non-ensemble algorithms, such as SVM, Discriminant Analysis, and Ensemble Classifier. Subsequently, the combinatorial judgment prediction method was created to ascertain the final classification model. Then, MATLAB was used to compare the recall of the CJC, and other frequently used classification models and influenza prediction algorithms, in order to assess each model.

Undoubtedly, the presently developed model has been statistically shown to have better projection performance, when compared to SVM, Discriminant Analysis, and Ensemble Classifier. This can adequately predict both long-term and short-term influenza outbreaks by providing early and accurate warnings for the onset of epidemic periods. Unlike the previous ARIMA prediction model, this produced random errors that resulted in poor accuracy mainly during peak influenza seasons. When the single performance ratio value was determined, the present CJC model exhibited a higher performance, when compared to its counterparts of Discriminant Analysis and Ensemble Classifier, but its performance was slightly lower than that of SVM. In addition, in the context of ratios, the prediction performance of CJC was comparatively steady. Several studies have reported different models. For instance, a previous study^[34] was conducted on the detection of early warning signals of influenza outbreak based on a dynamic network marker, and reported that the dynamical network marker (DNM) method has potential and power in detecting the early-warning signals for influenza outbreaks, which may lead a new way of public real-time surveillance for epidemic diseases. Compared to the present constructed model of CJC, which perfectly exhibited its timing, precision and accuracy, this can serve as a solid complement to the previous one. More importantly, this can be easily applied by different ministerial governments and independent researchers. In comparison to the ongoing COVID-19 pandemic, which has greatly hit and affected the world, a recent study^[15] developed the minimum-spanning-tree-based DNM (MST-DNM). When this was piloted and applied in Italy, the MST-DNM model was able to monitor the whole process of COVID-19 transmission, and successfully identify the early-warning signals. This confirms that when a proven model performs well, this can help to early detect diseases, and prevent and curb the transmission.

The present study is in line with the previous one, in which, theoretically, the more alarms are generated, the more precise the prediction model becomes^[40]. When alarms are generated, this means that the incident

is more irregular, and real, when compared to a data error. Therefore, an accurate predictive model would not only reduce the number of false alarms, but also prevent the triggering of alarms in a period of apparently high incidence, thereby preventing unnecessary further mobilization of sources in practice.

The present study has some limitations. Although it was hoped that all data and algorithms would be suitable for the present research, and since the weather and epidemic data used for the present research was released by the Hong Kong District Government, which did not consider regional differences, these may have affected the accuracy of the forecast. Furthermore, the algorithms used to build the CJC model lacks a lot of experimental evidence to show whether these are the best algorithms to build the combinatorial model.

There are several ways to improve the research. First, data from different districts can be selected for further experiments. Thus, other geographical differences can be further detected, which is not possible at the level of national surveillance. Hong Kong is a subtropical coastal city in the northern hemisphere. As it is known, the weather greatly affects the influenza forecast. Hence, one option is to choose the epidemic data and weather data from cities of different types of climate. For example, the epidemic data of Europe and Beijing greatly differ from that in Hong Kong^[44, 45]. Another improvement to consider would be the use of more algorithms and models.

In the present research, SVM, Discriminant Analysis, and Ensemble Classifier were used to build the CJC model. Naïve Bayes and recurrent neural networks (RNNs) are also effective algorithms for predicting influenza^[34, 46]. These algorithms could be added to the original CJC model, in order to develop a new CJC model, and compare the accuracy of the new and old models. In addition, the same algorithm exhibited different precisions when the data obtained from different domains was analyzed. This is also worthy of further investigation.

In the present study, based on meteorological factors and the fixed-point surveillance of influenza, the annual data of seasonal influenza outbreaks in Hong Kong was used as the research object, and a CJC model was constructed to predict the influenza outbreak trends. The prediction performance of this newly developed model was higher than that for Discriminant Analysis and Ensemble Classifier, in terms of classification accuracy, precision, sensitivity and specificity. Furthermore, the prediction stability has exhibited better convincing results, when compared to that of previous models. This would play a supplementary role in the early warning of influenza outbreaks, and allows for the timely preparation of healthcare systems, and benefits for patients, healthcare workers, and the society.

Conflict of Interest Statement

The authors declare no conflict of interest.

REFERENCES

- 1 Trucchi C, Paganino C, Orsi A, *et al.* Hospital and economic burden of influenza-like illness and lower respiratory tract infection in adults ≥ 50 years-old. *BMC Health Serv Res*, 2019,19(1):585
- 2 Molinari NA, Ortega-Sanchez IR, Messonnier ML, *et al.* The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine*, 2007, 25(27):5086-5096
- 3 Keilman LJ. Seasonal Influenza (Flu). *Nurs Clin North Am*, 2019,54(2):227-243
- 4 Dong M, Zhang X, Yang K, *et al.* Forecasting the COVID-19 transmission in Italy based on the minimum spanning tree of dynamic region network. *PeerJ*, 2021, 9:e11603
- 5 Spinney L. The Spanish flu: an interdisciplinary problem. *Lancet*, 2018,392(10164):2552
- 6 Baudon E, Peyre M, Peiris M, *et al.* Epidemiological features of influenza circulation in swine populations: A systematic review and meta-analysis. *PLoS One*, 2017, 12(6):e0179044
- 7 Gregg MB, Hinman AR, Craven RB. The Russian flu. Its history and implications for this year's influenza season. *JAMA*, 1978,240(21):2260-2263
- 8 Labella AM, Merel SE. Influenza. *Med Clin North Am*, 2013,97(4):621-645, x
- 9 Cai J, Xu B, Chan KKY, *et al.* Roles of Different Transport Modes in the Spatial Spread of the 2009 Influenza A(H1N1) Pandemic in Mainland China. *Int J Environ Res Public Health*, 2019,16(2):222
- 10 Sullivan SJ, Jacobson RM, Dowdle WR, *et al.* 2009 H1N1 influenza. *Mayo Clinic Proc*, 2010,85(1):64-76
- 11 Su S, Gu M, Liu D, *et al.* Epidemiology, Evolution, and Pathogenesis of H7N9 Influenza Viruses in Five Epidemic Waves since 2013 in China. *Trends Microbiol*, 2017,25(9):713-728
- 12 Tanner WD, Toth DJ, Gundlapalli AV. The pandemic potential of avian influenza A(H7N9) virus: a review. *Epidemiol Infect*, 2015,143(16):3359-3374
- 13 Yang W, Lau EHY, Cowling BJ. Dynamic interactions of influenza viruses in Hong Kong during 1998-2018. *PLoS Comput Biol*, 2020,16(6):e1007989
- 14 Chen Y, Yang K, Xie J, *et al.* Detecting the outbreak of influenza based on the shortest path of dynamic city network. *PeerJ*, 2020,8:e9432
- 15 Chen P, Chen E, Chen L, *et al.* Detecting early-warning signals of influenza outbreak based on dynamic network marker. *J Cell Mol Med*, 2019,23(1):395-404
- 16 Chiu SS, Kwan MY, Feng S, *et al.* Early season estimate of influenza vaccination effectiveness against influenza hospitalisation in children, Hong Kong, winter influenza season 2018/19. *Euro Surveill*, 2019,24(5):1900056
- 17 Wen A, Wang LW, He H, *et al.* An aberration detection-based approach for sentinel syndromic surveillance of COVID-19 and other novel influenza-like illnesses. *J Biomed Inform*, 2021,113:103660
- 18 Stroup DF, Williamson GD, Herndon JL, *et al.* Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Stat Med*, 1989,8(3):323-329; discussion 331-332
- 19 Vandendijck Y, Faes C, Hens N. Eight years of the Great Influenza Survey to monitor influenza-like illness in Flanders. *PLoS One*, 2013,8(5):e64156
- 20 Michiels B, Thomas I, Van Royen P, *et al.* Clinical prediction rules combining signs, symptoms and epidemiological context to distinguish influenza from influenza-like illnesses in primary care: a cross sectional study. *BMC Fam Pract*, 2011,12:4
- 21 Sooryanarain H, Elankumaran S. Environmental role in influenza virus outbreaks. *Annu Rev Anim Biosci*, 2015,3:347-373
- 22 Liu W, Dai Q, Bao J, *et al.* Influenza activity prediction using meteorological factors in a warm temperate to subtropical transitional zone, Eastern China. *Epidemiol Infect*, 2019,147:e325
- 23 Kim J, Ahn I. Weekly ILI patient ratio change prediction using news articles with support vector machine. *BMC Bioinformatics*, 2019,20(1):259
- 24 Chen P, Liu R, Aihara K, *et al.* Autoreservoir computing for multistep ahead prediction based on the spatiotemporal information transformation. *Nat Commun*, 2020,11(1):4568
- 25 Vuichard-Gysin D, Mertz D, Pullenayegum E, *et al.* Development and validation of clinical prediction models to distinguish influenza from other viruses causing acute respiratory infections in children and adults. *PloS One*, 2019,14(2):e0212050
- 26 Alessa A, Faezipour M. A review of influenza detection and prediction through social networking sites. *Theor Biol Med Model*, 2018,15(1):2
- 27 Barry MA, Florent Arinal F, Cheikh Talla C, *et al.* Performance of case definitions and clinical predictors for influenza surveillance among patients followed in a rural cohort in Senegal. *BMC Infect Dis*, 2021,21(1):31
- 28 Wang XL, Wu SS, Wu XN, *et al.* A study on the prediction of influenza based on the climate factors and influenza viral activity. *Int J Virol*, 2017,4(5):296-299
- 29 N'Gattia AK, Coulibaly D, Nzussouo NT, *et al.* Effects of climatological parameters in modeling and forecasting seasonal influenza transmission in Abidjan, Cote d'Ivoire. *BMC Public Health*, 2016,16(1):972
- 30 Soebiyanto RP, Adimi F, Kiang RK. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PLoS One*, 2010,5(3):e9450
- 31 Wang P, Peng Y, Yang XB. Application of ARIMA model and Holt-Winters exponential smoothing method to predict influenza-like cases in Wuhan. *Modern Prevent Med*, 2018,45(3):385-389
- 32 Xu M, Cao C, Li Q, *et al.* Ecological Niche Modeling of Risk Factors for H7N9 Human Infection in China. *Int J Environ Res Public Health*, 2016,13(6):600
- 33 Gao J, Wang K, Ding T, *et al.* Forecasting influenza A pandemic outbreak using protein dynamical network biomarkers. *BMC Syst Biol*, 2017,11(Suppl 4):85
- 34 Lopez Pineda A, Ye Y, Visweswaran S, *et al.* Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *J Biomed Inform*, 2015,58:60-69
- 35 Chen L, Liu R, Liu ZP, *et al.* Detecting early-warning signals for sudden deterioration of complex diseases by

- dynamical network biomarkers. *Sci Rep*, 2012,2(1):342
- 36 Liu R, Zhong J, Hong R, *et al.* Predicting local COVID-19 outbreaks and infectious disease epidemics based on landscape network entropy. *Sci Bull*, 2021,66:2265-2270
- 37 He Z, Tao H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. *Int J Infect Dis*, 2018,74:61-70
- 38 Paul S, Mgbere O, Arafat R, *et al.* Modeling and Forecasting Influenza-like Illness (ILI) in Houston, Texas Using Three Surveillance Data Capture Mechanisms. *Online J Public Health Inform*, 2017,9(2):e187
- 39 Gao H, Wong KK, Zheteyeva Y, *et al.* Comparing Observed with Predicted Weekly Influenza-Like Illness Rates during the Winter Holiday Break, United States, 2004-2013. *PLoS One*, 2015,10(12):e0143791
- 40 Michiels B, Nguyen VK, Coenen S, *et al.* Influenza epidemic surveillance and prediction based on electronic health record data from an out-of-hours general practitioner cooperative: model development and validation on 2003-2015 data. *BMC Infect Dis*, 2017,17(1):84
- 41 Manitz J, Hohle M. Bayesian outbreak detection algorithm for monitoring reported cases of campylobacteriosis in Germany. *Biom J*, 2013,55(4):509-526
- 42 Campe H, Heinzinger S, Hartberger C, *et al.* Clinical symptoms cannot predict influenza infection during the 2013 influenza season in Bavaria, Germany. *Epidemiol Infect*, 2016,144(5):1045-1051
- 43 Yang K, Xie J, Xie R, *et al.* Real-Time Forecast of Influenza Outbreak Using Dynamic Network Marker Based on Minimum Spanning Tree. *Biomed Res Int*, 2020,2020:7351398
- 44 Koppeschaar CE, Colizza V, Guerri C, *et al.* InfluenzaNet: Citizens Among 10 Countries Collaborating to Monitor Influenza in Europe. *JMIR Public Health Surveill*, 2017,3(3):e66
- 45 Wang X, Wu S, MacIntyre CR, *et al.* Using an adjusted Serfling regression model to improve the early warning at the arrival of peak timing of influenza in Beijing. *PLoS One*, 2015,10(3):e0119923
- 46 Venna SR, Tavanaei A, Gottumukkala RN, *et al.* A Novel Data-Driven Model for Real-Time Influenza Forecasting. *IEEE Access*, 2019,7:7691-7701

(Received Jul. 12, 2021; accepted Oct. 26, 2021)