Influenza Outbreak Forecast in New York City with Weather Data

Zhengxing Li, Ziying Tao, Guangjun Xue, Yichen Zhang

Abstract— **Influenza is a highly contagious acute respiratory illness recognized since ancient times. Better epidemic predictions would set up more appropriate public health prevention and intervention strategies in temperate cities, like the New York City. After several literature works, we found epidemics occur mainly during the season months with abnormal changing of temperature, precipitation, UV radiation, and wind speed. Thus, we built some models, Random Forest, Linear Regression, Gradient Boosting, K-Nearest Neighbors Algorithm, Deep Neural Network, to predict the break of influenza. The result shows that our model could predict the influenza to some extent, with more enough time we also want to use more weather stations data to train and adjust the models, making the results become better.**

I. INTRODUCTION

In the United States, influenza is one of the most significant diseases in humans, generating worldwide annual epidemics, which result in about three to five millions cases of severe illness, and about 250,000 to 500,000 deaths [1]. Therefore, improving influenza knowledge about key epidemiological parameters such as survival, transmission and reproduction in hosts is essential to upgrade surveillance network and to develop more accurate predicting models. Better epidemic predictions would set up more appropriate public health prevention and intervention strategies.

Epidemics occur mainly during the winter season months in temperate cities (like the New York City) [2-4] unlike in tropical and sub-tropical cities where they generally happen during the rainy season [5-8]. These differences suggest a climate impact on influenza spread. Climate might affect influenza diffusion (onset, duration, size) by impacting individuals contact rates (frequency and duration), population immunity and virus survival outside human body. Various climatic factors such as temperature, humidity, rainfalls, UV radiation, sunshine duration and wind speed might have an impact on influenza spread. In temperate countries, humidity and temperature might play an important role in influenza spread. Several laboratory works showed that a cold and dry weather promotes a higher virus survival outside human body and a better transmission [9]. Another theory suggests a link between vitamin D secretion and influenza immunity, which is supported by experiments [10, 11]. As UV radiation is involved in vitamin D production, a lack of UV radiation in winter, for temperate countries, leads to a reduction of vitamin D production and might boost influenza epidemics [12, 13]. Dowell also suggested a role of dark/light cycles and photoperiod on the immune systems caused by melatonin fluctuations [14]. Thereby UV radiation and sunshine duration might have an indirect effect on influenza infections. Finally, in China, Xiao et al. [15] proposed that a low wind speed contribute to influenza spread. In fact, a strong wind speed may have a dispersive effect on influenza in the environment limiting its diffusion.

Our goals are to predict the influenza incidents in the future years of New York City basing on the weather data. For instance, we are interested in the number of incidents in the next year. And also, we want to know which month in the year would have possible outbreaks. Using these predictions, we can make some recommendations to the government and do some prevention.

II. RELATED WORK

There are three immunological types of influenza virus: A, B, C. The type A virus is highly variable and shows continuous antigenic variation and is a major cause of frequent epidemics and periodic pandemics. It also infects animals and birds. Type B virus shows antigenic variation to a lesser degree which results in epidemics, whereas type C appears to be antigenically stable and causes sporadic upper respiratory tract illness. It is estimated that annually around 0.5-1 million people die and 600 to 1,200 million people become sick due to influenza epidemics worldwide (Layne et al, 2001). Thus the disease affects a large segment of the world population resulting in significant mortality, morbidity and economic loss. The World Health Organization has established a global network of 112 national influenza centers in 83 countries and regularly reports on the global influenza situation and recommends current updated strains for use in the influenza vaccine. Presently antigenic variant strains of influenza type A(H1N1), A(H3N2) and type B viruses are causing frequent epidemics in humans globally. Surveillance is essential for identifying the new variants of these types and subtypes for the selection of vaccine strains.

As a part of the influenza program, a study was initiated at the National Institute of Virology (NIV) Pune City, Maharashtra State, India, since 1976 which has been recognized as the National Influenza Center since 1980 by the WHO. During the course of this continuous surveillance of influenza in Pune City between 1976 to 2002, NIV investigated several outbreaks of influenza and isolated 43 antigenic variant strains of influenza types A and B, which included many global epidemics strains (Rao et al, 1979, 1982; Rao and Banerjee 1993; Rao, 2003). The present communication reports the variant strains of influenza type A (H3N2) and type B isolated during influenza outbreaks in the year 2003 from Pune City.

III. DATASETS AND FEATURES

A. Datasets and Prepossessing

We used daily weather data from National Oceanic and Atmospheric Administration(NOAA), which contains global surface summary of the day from New York City Central Park weather station. The influenza data we used is the weekly influenza test report from Centers for Disease Control and Prevention. To investigate the most recent potential relationship, we choose to use datasets for the time period from 2011 to 2018.

For the prepossessing of the raw data, we first treat with the missing data, replacing them with mean values. Then we aggregated the daily weather data frame into weekly data frame, and merged the weekly weather data and influenza data into one.

The prepossessed data contains:

- **year:** from 2011 to 2018
- **week:** from 1 to 52 or 53
- **temp_weeklyMin:** weekly minimum temperature
- **temp_weeklyMax:** weekly maximum temperature
- **temp_weeklyMean:** weekly mean temperature
- **slp_weeklyMean:** weekly mean sea level pressure
- **wdsp_weeklyMedian:** weekly median wind speed
- **wdsp_weeklyMax:** weekly maximum wind speed
- **prcp_weeklyMean:** weekly mean precipitation
- **uv_time:** weekly mean uv time
- **ILITOTAL:** weekly reported influenza infections

B. Features and Visualizations

Before we get started to develop models, we created visualizations of the features to explore the potential relationship between them. The pair plot of all features is shown in Figure 1. We can find that some features are highly linearly related, so that we could choose to use one of them in our models.

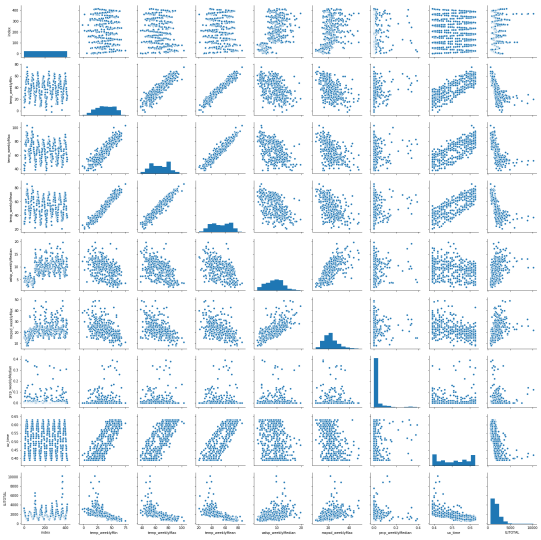

Fig. 1: Pairplot of all features

We also plotted a heatmap of features and influenza infections to find the correlation influence between different features and influenza infection numbers. From fig 2, we can find that the influenza infection numbers are more related to features: slp_weeklyMean, wdsp_weeklyMedian, wdsp_weeklyMax, prcp_weeklyMean.

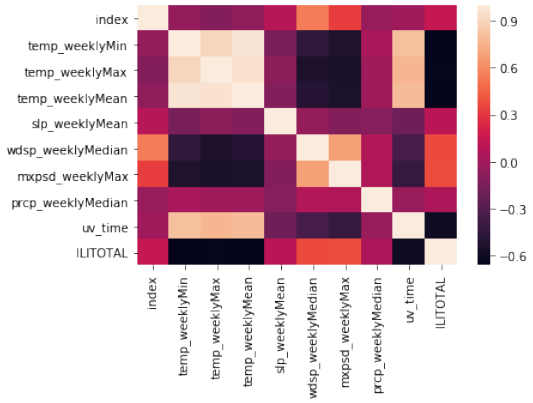

Fig. 2: Correlation heatmap of features and influenza infections

IV. METHODS

A. Linear Regression

We tried linear regression model, which is a simple but effective model compared with the baseline, to predict rating from a combination of various features: $feature_1$, $feature_2$, ..., $feature_n$. We created an one dimension array for the prediction. The first entry was the bias θ_0, each other entry, θ_1, θ_2, ... θ_n, represented the parameters for other features.

$$f = \theta_0 + \theta_1 * feature_1 + \theta_2 * feature_2 + ... + \theta_n * feature_n \tag{1}$$

In this method, we intended to fit a model, minimizing the residual sum of squared distance between the predicted values and true values in the training data. After inputted our training data, we are able to calculate the optimal parameters in the equation above.

B. Random Forest Regression

Random forest is an ensemble model: using decision trees as individual models and bagging as the ensemble method. We randomly chose samples from the training data to build trees, and then randomly selected subsets of features to generate the best split. Given a training set X = x_1, x_2, ..., x_n, with output Y = y_1, y_2, ..., y_n. The random forest was built, consisting of B multiple such random decision trees: for b = 1,...,B, we train a regression tree f_b on X_b, Y_b. To predict the influenza infections, we output the mean predicted regression outputs of the trees in the forest.

$$\widehat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{2}$$

C. Gradient Boosting

The idea of gradient boosting is an ensemble of week prediction models, typically decision trees. We set up the differentiable loss function:

$$L(y,F(x)) = \sum (y_i - F(x_i))^2 \tag{3}$$

For gradient boosting, we started with a initialize decision tree on data:

$$F_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, \gamma) \tag{4}$$

For iteration time from m = 1 to M, we set the measurement of residuals r_{im} as in eq.5, and fitted the updated tree $h_m(x)$ to r_m.

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{5}$$

Then we solve for the multiplier γ_m to update the model:

$$\gamma_m = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{6}$$

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{7}$$

D. K-Nearest Neighborhood

For the k-nearest neighborhood(knn) model, we did the prediction through finding out the closest relationship between training data and testing data. The measurement of relationship or similarity is based on the inverse of their distance.

For the training data, we have pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), given the measurement of similarity, we reordered the training data according to the distance between the target test data.

We set k as the k nearest point to the target data: the k most similar data point. After trying to set k from 1 to 20, we found that when k=8, we can get the lowest RMSE in the validation data. In our knn model, we set k=8: we picked the 8 most similar weekly weather data compare to the target one and calculate the average infection number of these similar weeks as in eq.8.

$$f(y = j \mid X = x) = \frac{1}{k} \sum_{i \in A} I(y^i = j) \tag{8}$$

E. Deep Neural Network

In our deep neural network model, we designed a network consists of one input layer, two hidden layer and the output layer as shown in fig. 3.

The dimension of input x of the neural network model is 415, and we trained the two layer of the model with ReLU function:

$$f(x) = max(x, 0) \tag{9}$$

Let W_1 be the weight matrix of the first layer, b_1 be the bias of the first layer, W_2 be the weight matrix of the second layer, and b_2 be the bias of the second layer. The parameters W_1, W_2 are learned with stochastic gradient descent. We derived them with chain rule. We should compute the output of each layer

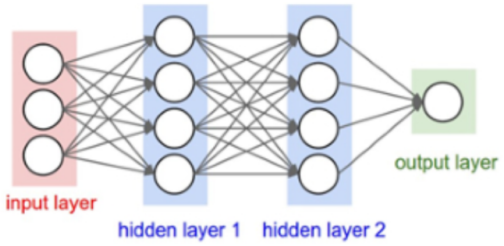

Fig. 3: Deep Neural Network Model

$$h_{W,b}(x) = ReLU(Wx + b) \tag{10}$$

We trained our neural network using batch gradient descent descent: first perform a feedforward pass, computing the activation for layers until the output layer. And we repeatly took the partial derivatives for the cost function $J(W,b)$ to reduce the cost function.

$$\frac{\partial}{\partial W_{ij}} J(W,b;x,y) = a_j^l \delta_i^{l+I} \tag{11}$$

$$\frac{\partial}{\partial b_i} J(W,b;x,y) = \delta_i^{l+I} \tag{12}$$

V. RESULTS

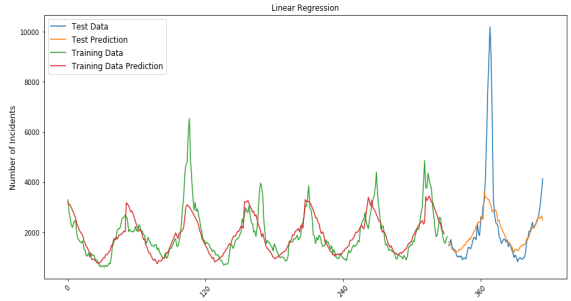

Fig. 4: Regression result of Linear Regression model

Linear regression has the worst performance on test set. This is reasonable because the prediction can not be linear.

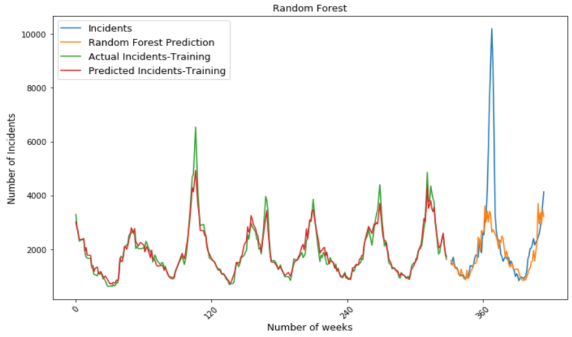

Fig. 5: Regression result of Random Forest model

Random forest regression has a really decent result. It fits the training data pretty good. Even though it did not predict the future accurately, it predicted the outbreak successfully.

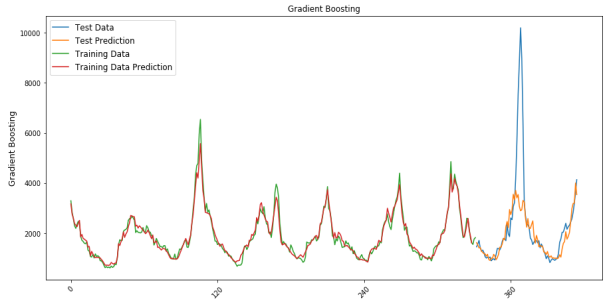

Fig. 6: Regression result of Gradient Boosting model

Gradient Boosting has the best result on both test set and training set. It has the lowest RMSE and fits the data best.

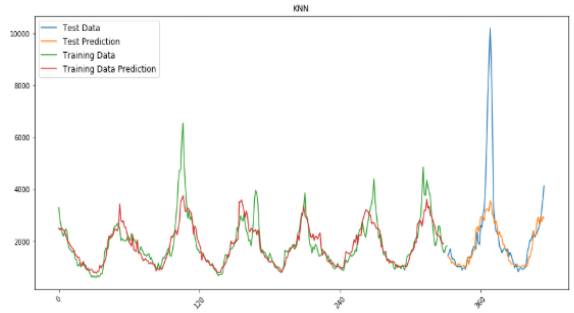

Fig. 7: Regression result of KNN model

The performance of the KNN model is not bad. Although it has a larger RMSE value than Gradient Boosting and random forest, it predicts the happening of each outbreaks successfully.

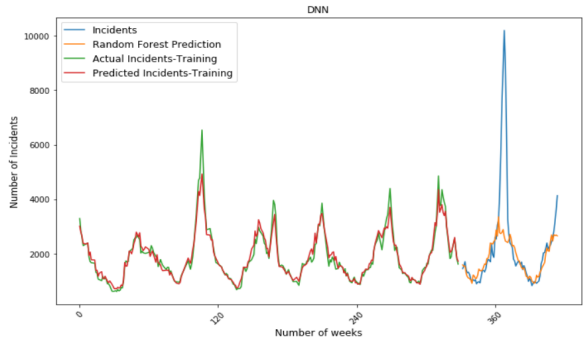

Fig. 8: Regression result of DNN model

The result of DNN model fits the data well. However, it failed to predict the exact number of incidents near week 360 accurately.

Overall, it turns out that all the models predict the outbreak of influenza successfully. And Gradient Boosting has the best result on test set.

Models	Training Data	Testing Data
Linear Regression	558.823	1403.405
Random Forest	222.476	1461.625
Gradient Boosting	169.714	1392.070
KNN(k=8)	601.175	1500.434
DNN	527.139	1470.413

TABLE I: RMSE of models

VI. DISCUSSION

Linear Regression model, using the combination of all the features, is a simple but effective way in predicting. For that flu infection is the result of interaction of weather features, KNN model, which is based more on similarity, is not very suitable for this case.

Our weather data are regularly changing with seasons changing, which likes the data are combined by several lines just with positive or negative slopes and linearly at every small locality. Then, during reading references, we find the New York City is a typical temperature city where the UV radiation and precipitation play the significant role. So we use the models to test and train the data without UV radiation and precipitation respectively, the consequence shows that the RMSE without UV/Precip is four times of the RMSE with UV/Precip, which agrees with the reference and we approve these by using the machine learning.

VII. CONCLUSION

Above all, we, through the whole results, find the influenza could be controlled the weather. Because with the changing of weather, like temperature, UV radiation, precipitation, wind speed and snow down, the patients of influenza will decrease or increase regularly. Besides, after training several models, we find the test consequences could meet our expectation. The test result can predict the break of the influenza in the New York City just under the combination with the different weather data, which make our goal come true and really could help the government to do some preparation before the break of influenza.

VIII. FUTURE WORK

We will use universal weather data, like finding different meteorological stations data, and using the longer span of year, or do more literature research to find a better way, to do test. Besides, we will also improve our model and try several new models to see the consequence which is more reasonable.

IX. CONTRIBUTION

This is a term work, thanks for all four of our group members. Zhengxing Li finds and collect the data of weather and influenza. Ziying Tao collects data and builds the models. Guangjun Xue collects the codes and builds the models. Yichen Zhang collects the data, designs and edits the poster. we worked together in the part of conclusion and results analysis. Everyone contributes to the project greatly.

REFERENCES

[1] World Health Organization: Influenza (Seasonal). 2014

[2] Viboud C, Flahault A. Influenza Epidemics in the United States, France, and Australia.

[3] Tamerius JD, Viboud C. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates.

[4] Finkelman BS, Grenfell BT. Global Patterns in Seasonal Activity of Influenza A/H3N2, A/H1N1, and B from 1997 to 2005: Viral Coexistence and Latitudinal Gradients.

[5] Moura FEA, Perdigo ACB, Siqueira MM. Seasonality of influenza in the tropics: a distinct pattern in Northeastern Brazil.

[6] Rao BL, Banerjee K. Influenza surveillance in Pune, India, 197890.

[7] Rao BL, Yeolekar LR, Kadam SS, Pawar MS, Kulkarni PB, More BA, Khude MR. Influenza surveillance in Pune, India, 2003.

[8] Dosseh A, Ndiaye K, Spiegel A, Sagna M, Mathiot C. Epidemiological and virological influenza survey in Dakar, Senegal: 1996-1998.

[9] Fuhrmann C. The effects of weather and climate on the seasonality of influenza: what we know and what we need to know.

[10] Lowen AC, Steel J, Mubareka S, Palese P. High temperature (30 C) blocks aerosol but not contact transmission of influenza virus. J Virol.

[11] Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on relative humidity and temperature.

[12] McDevitt J, Rudnick S, First M, Spengler J. Role of absolute humidity in the inactivation of influenza viruses on stainless steel surfaces at elevated temperatures. Appl Environ Microbiol.

[13] Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M. Absolute humidity and the seasonal onset of influenza in the continental United States.

[14] Dowell SF. Seasonal variation in host susceptibility and cycles of certain infectious diseases. Emerg Infect Dis.

[15] Xiao H, Tian H, Lin X, Gao L, Dai X, Zhang X, Chen B, Zhao J, Xu J. Influence of extreme weather and meteorological anomalies on outbreaks of influenza A (H1N1) Chin Sci Bull.

[16] Radhika Y, Shashi M. Atmospheric temperature prediction using support vector machines. International Journal of Computer Theory and Engineering. 2009 Apr; 1(1):1793 8201.

[17] Han J, Kamber M. Data mining: Concepts and techniques. Morgan and Kaufmann; 2000.

[18] Smith BA, McClendon RW, Hoogenboom G. Improving air temperature prediction with artificial neural networks. International Journal of Computational Intelligence. 2007; 3(3):17986.