Research Article

Learning from Large-Scale Wearable Device Data for Predicting Epidemics Trend of COVID-19

Guokang Zhu, Jia Li, Zi Meng, Yi Yu, Yanan Li, Xiao Tang, Yuling Dong, Guangxin Sun, Rui Zhou, Hui Wang, Kongqiao Wang*, and Wang Huang

Huami Corporation, Hefei, China

Correspondence should be addressed to Kongqiao Wang; kongqiao.wang@huami.com

Received 27 March 2020; Accepted 15 April 2020; Published 5 May 2020

Guest Editor: Jinchang Ren

Copyright © 2020 Guokang Zhu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The pandemics of COVID-19 triggered out an alarm on the public health surveillance. The popularity of wearable devices enables a new perspective for the precaution of the infectious diseases. In this study, we propose a framework, which is based on the heart rate and sleep data collected from the wearable devices, to predict the epidemic trend of COVID-19 in different countries and cities. On top of a physiological anomaly detection algorithm defined based on wearable device data, an online neural network prediction modelling methodology combining both detected physiological anomaly rate and historical COVID-19 infection rate is explored. 4 models are trained separately according to geographical segmentation, i.e., North China, Central China, South China, and South-Central Europe. The de-identified sensor data from about 1.3 million wearable device users are used for verification. Experiment results indicate that the prediction models can be utilized to alert the outbreak of COVID-19 in advance, which sheds light on a health surveillance system with wearable device.

1. Introduction

With the pandemics of the novel coronavirus, now named Corona Virus Disease 2019 (COVID-19), more than 300,000 people have been infected in at least 127 countries as of March 23, 2020, according to the WHO’s report [1]. COVID-19 spreads from person to person and has killed thousands [2–5]. Over the last few weeks since COVID-19 outbreak, several studies have been done to forecast the epidemics trend of COVID-19 in China [6–8]. For example, Wu et al. built a Susceptible-Exposed-Infectious-Recovered (SEIR) model to simulate the epidemics across the major cities in China [7]. Yang et al. applied the Long Short Term Memory (LSTM) model to predict the number of the newly infected COVID-19 cases by utilizing the data of Severe Acute Respiratory Syndrome (SARS) in 2003 [6]. Although the models from those studies could simulate the outbreak trend of the disease, they relied heavily on the officially reported statistics; therefore, the timeliness of the models could be affected. On the contrary, big data analysis, such as analysis of Internet data, may provide real-time surveillance and improve the timeliness of the forecasting [9–16]. For instance, Google invented the influenza epidemic prediction tool Google Flu Trend (GFT) to estimate the level of influenza activity based on the individual web search queries from different regions [9–11]. They assumed that more individuals might search online for the information about specific diseases, if the influenza disease risk is higher in a certain region. Therefore, Google built a database containing 50 millions of the most common web search queries on all influenza related topics and constructed the risk prediction model GFT with those search query data as input [9]. Google showed that GFT could help predict the influenza-like illness outbreak 7–10 days before the Centers for Disease Control and Prevention (CDC) report [10]. In fact, the surveillance report from CDC usually has a lag around 1-2 weeks. Therefore, the result from Google indicated that big data analysis could promote the timeliness for public health surveillance. However, search queries can be greatly influenced by the social hotspots, which weakens the correlation between the search queries and the influenza-like illness diseases [17].
With the popularity of fitness band and smartwatch, physiological signs, such as heart rate, activity, sleep, and so on, can be conveniently acquired from the wearable biosensors [18–20]. As of 2019, more than 100 million consumers owned Huami wearable devices, and the number continues growing. In contrast with the big data from web search engine, data from wearable devices can provide more objective information on the health status of the users. Once the users are infected with the influenza-like illness, for example, their physiological signs would be altered. Radin et al. explored the relationship between physiological anomaly rate of wearable device users and the influenza-like illness rate reported from the US CDC [21] to build the regression models to predict the influenza-like illness cases at different states of America. They utilized the data of heart rate and sleep from the wearable devices to improve the traditional models. The prediction results have strong correlation with the official data. Li et al. also investigated the role of the physiological change measured with wearable devices on the diagnosis and analysis of disease [22]. The researchers came up with a personalized disease detection framework, which identifies the abnormal physical signs, e.g., Lyme disease and other inflammatory responses, from the longitudinal data of the individuals. All these studies mentioned above enlighten the way to use wearable device data for public health surveillance.

According to the clinical studies [23–25], the common symptoms at the onset of COVID-19 are fever, cough, and fatigue, which are closely related to the physiological signs measured by the wearable devices. Therefore, it might be a good way to predict the epidemic trend of COVID-19 by building a prediction model based on the wearable device data.

The main purpose of this study is to provide a novel framework for predicting the trend of COVID-19 outbreak at different countries and cities, using big data collected from the wearable devices. There are two major contributions of this study: (1) a physiological anomaly detection method is developed and can identify the anomalous signs reflected by the physiological data from wearable sensors; (2) an online learning framework is proposed for public health emergency surveillance.

2. Methodology

2.1. Physiological Anomaly Detection. According to the study on fever and cardiac rhythm [26], heart rate increases by 8.5 beats per minute, on average, for every 1°C increase in body temperature, so the elevated RHR might be related to fever caused by COVID-19 or influenza-like illness. The basic anomaly detection method is based on the elevated RHR. Because the shortened sleep length also causes an increase of RHR [27], we weaken this factor in the physiological anomaly detection method.

RHR and sleep length are directly acquired with the corresponding sensors of Huami wearable devices. Both kinds of synchronized data from the accelerometer (ACC) sensor and the photoplethysmography (PPG) sensor are used to analyze sleep status (including sleep recognition and stage) for getting sleep length. During sleep, the PPG data is used to compute the RHR. For each user, overall mean and Standard Deviation (SD) of RHR and sleep length throughout the entire period are calculated. A daily RHR is defined as anomaly if it is larger than the average RHR plus 1.5 SD, and meanwhile, the daily sleep is longer than the average sleep minus 0.5 SD. Considering that COVID-19 or influenza-like illness always persists days, we define the detection standard of physiological anomaly as continuous anomaly for at least 5 consecutive days.

2.2. Online Prediction of COVID-19 Infection Rate. The physiological anomaly detected by our method is an indication of fever, which in fact can be caused by COVID-19 or influenza-like illness. Thus, the key point for COVID-19 infection rate prediction is to distinguish COVID-19 anomaly from the whole physiological anomaly. To this end, as shown in Figure 1, a heterogeneous neural network [28] regression model combining sparse categorical features and dense numerical features (CDNet) is proposed.

CDNet concatenates 2 subnetworks: CatNN and DenNN. The inputs of the CatNN are sparse categorical features, i.e., holiday activity, season, and weather. The inputs of the DenNN are historical physiological anomaly rate, active user density, and historically officially reported COVID-19 rate, where the historically detected physiological anomaly rate is calculated with dividing the number of users detected as physiological anomaly by the number of total active users. The output layer of CDNet normalized by a Sigmoid function outputs the predicted physiological anomaly rate. The detailed inputs and outputs are summarized in

\[
R'_{t+1,k} = \text{CDNet}\left(\{R_{t-j,k}, r_{t-j,k}, C_{t-j,k}, c_{t-j,k} : j = 0, \ldots, 6\}, \right.
\]

\[
\left. R_{t,k}, D_{t,k}\right),
\]

where, for country or city \(k\), the output \(R'_{t+1,k}\) is the predicted physiological anomaly rate in the next period, \(R_{t-j,k}\) is the physiological anomaly rate the \(j\)-th period earlier, \(r_{t-j,k}\) is the physiological anomaly rate in the same period of \(R_{t-j,k}\) last year, \(C_{t-j,k}\) and \(c_{t-j,k}\) are the corresponding categorical information with the same temporal definition as \(R_{t-j,k}\) and \(r_{t-j,k}\), respectively, \(R_{t,k}\) is the officially reported COVID-19 rate (ratio of confirmed COVID-19 patient number to the number of residents in the country or city) in the current period, \(D_{t,k}\) is the current active user density (ratio of active user number to the number of residents in the country or city). To distinguish regional disparity, 4 different CDNet models are trained separately for North China, Central China, South China, and South-Central Europe.

In order to get the predicted anomaly rate caused by COVID-19 for the next period, the predicted physiological anomaly rate with \(R'_{t+1,k}\) and without \((R'_{t+1,k}|RC_{t,k} = 0)\) the supervision of officially reported data is calculated separately. As shown in Figure 2, the supervision is removed by setting \(RC_{t,k}\) as 0. Then, the predicted anomaly rate caused
by COVID-19 for the next period $P_{t+1,k}$ can be calculated as the difference between $R_{t+1,k}$ and $R_{t+1,k} | RC_{t,k} = 0$:

$$P_{t+1,k} = R_{t+1,k} - \left\{ R_{t+1,k} | RC_{t,k} = 0 \right\}.$$  

(2)

To consecutively predict the epidemics trend of COVID-19, the CDNet model is trained in an online learning way. As shown in Figure 3, the initial CDNet model $M_0$ is trained with the input of $\{ R_{t-j,k}, T_{t-j,k}, C_{t-j,k}, c_{t-j,k}, j = 1, 2, \ldots, 7 \}$, and with the target as $R_{t,k}$. The weights of CDNet are updated step by step with the transmission of COVID-19, using the arriving data of newly officially reported COVID-19 rate and detected physiological anomaly rate. The step size of the sliding window for online learning is set as 1 week.

3. Experiments

3.1. Dataset. De-identified sensor data of about 1.3 million users who wore Huami devices from July 1, 2017, to April 8, 2020 were obtained according to appropriate security control. All the users are notified that their de-identified data could potentially be used for academic research under the Huami Privacy Policy.

All the users wore their Huami devices for at least 100 days throughout the entire period. Daily measures include Resting Heart Rate (RHR), activity, and sleep length, which are the bases of physiological anomaly detection. Data with missing RHR or sleep length are excluded. The daily COVID-19 infection rate data come from CDC of the corresponding countries.

We build separate models for different countries and cities listed in Table 1, according to the geographical segmentation considering the regional and life style differences. Taking North China as an example, we utilize data from 5 representative cities (Beijing, Shijiazhuang, Jinan, Taiyuan, and Tianjin) to do analysis and model building. The detailed summaries of the active user numbers are also listed in Table 1. The users enrolled in the study are chosen from 19 cities of central, southern, and northern China and 7 South-central European countries to sufficiently reveal the regional disparity.

3.2. Analysis Result in China. The consecutive 3-year physiological anomaly rate curves in Wuhan together with the predicted physiological anomaly rate curves with and without the supervision of the officially reported COVID-19 infection rate in 2020 are illustrated in Figure 4. They are aligned by the time of Chinese Spring Festival in the temporal axis. In the figure, all the 5 curves peak around the time of Chinese Spring Festival. Besides, the predicted physiological anomaly rate with the supervision of official data in 2020 fits well with the rate calculated by the anomaly detection algorithm, which validates the prediction performance of the CDNet. Additionally, the physiological anomaly rate curve excluding COVID-19 in 2020 overlaps with both the predicted and the detected physiological anomaly rate curves including COVID-19 in 2020 before the outbreak of COVID-19, which verifies the basic reliability of the model. After that, all these 3 curves rise rapidly, which indicates that the outbreak of influenza-like illness is together with COVID-19. The predicted outbreak period coincides with the real situation. In addition, we also predict the physiological anomaly rate curve from 2018 to 2019 with the prediction model and find that the predicted curve fits well with the total anomaly rate curve during the 2 years. This may indicate that the obvious separation happening around the Chinese Spring Festival between the predicted anomaly rate curves with and without the supervision of the officially reported COVID-19 infection rate in 2020 results from the outbreak of COVID-19.

Figure 5 illustrates the predicted COVID-19 infection rate across 5 Chinese cities and the officially reported accumulating COVID-19 infection cases in Wuhan. In the figure, there is an obvious outbreak period in the predicted infection rate curve for each city, which may correspond to that of the newly confirmed cases. Taking Wuhan as an example, the predicted infection rate peaks around January 28, while the officially reported newly confirmed infection rate in Wuhan reached largest on February 8 (the data after February 12 in Wuhan is omitted since the COVID-19 diagnostic criteria changed on that day, which causes a sudden sharp increase of 13,436 newly confirmed cases). The predicted disease peak is ahead of the officially reported peak by 11 days. The early arrived peak may indicate that health surveillance with the wearable sensors can play an important role.
role in infectious disease alert and in time public health management. In fact, Wu and McGoogan also found there was a lag between the start of the illness and the diagnosis of COVID-19 by viral nucleic acid testing [2]. The newly infected cases were actually peaked around January 28 if determined by the onset of the symptoms, which happens to be consistent with our findings. In addition, Figure 5 also shows that the predicted infection rate in Wuhan gradually decreases since January 28 and reaches a local minimum on February 1, which may correspond to the plateau in the
3.3. Analysis Result in Italy and Spain. Figures 6(a) and 6(b) illustrate the predicted COVID-19 infection rate and the officially reported accumulating COVID-19 infection rate in Italy and Spain, respectively. The predicted infection rate in Italy rises rapidly from February 23, 2020, which coincides with the outbreak of COVID-19 in this country. As for Spain, the predicted infection rate starts to increase from February 29, which is 6 days later than Italy, and the predicted rate increases quickly since that. This is consistent with the real situation that the outbreak of COVID-19 was later in Spain.

As shown in Figure 6, the principal peak in the predicted COVID-19 infection curve of either Italy or Spain has arrived as of April 8. In correspondence to the largest number of newly confirmed infection cases, which are reported officially by Italy on March 21 and Spain on March 25, the predicted principal peaks for the two countries happen around the time of March 13 and March 18, respectively. Both predicted principal peaks are ahead of the officially reported data by at least 1 week.

3.4. Correlation Analysis. To evaluate the reasonableness of predicting COVID-19 infection rate from physiological anomaly rate, we choose 19 Chinese cities to calculate the correlation between the officially reported COVID-19 infection rate and the detected physiological anomaly rate using Pearson’s correlation coefficient shown in equation (3). In the equation, $t_0$ represents the start of the COVID-19 outbreak, $t_1$ stands for the end of the study period, and $X$, $Y$ represent the officially reported COVID-19 infection rate and the physiological anomaly rate, respectively. The correlation analysis is performed with two steps. In the first step, we find the point, corresponding to the outbreak peak point of the officially reported COVID-19 infection curve, on the physiological anomaly rate curve. In the second step, we align the curves by the two points, and calculate the correlation coefficient.

$$\rho_{X,Y} = \frac{\sum_{t=0}^{t_1} (X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=0}^{t_1} (X_t - \bar{X})^2 \sum_{t=0}^{t_1} (Y_t - \bar{Y})^2}}$$  

Pearson’s correlation coefficients $\rho$ for different cities in China are listed in Table 2. The average $\rho$ value reaches around 0.68, which is strong correlation that further supports the opinion that physiological signs are useful for public health emergency alert. However, some cities do not show strong correlation, which may be due to the following reasons. Firstly, the officially reported cases of infection in some cities, e.g., Wuhan, were adjusted on certain days resulting in sudden changes. Secondly, the number of active users in some cities, e.g., Nanning, are relatively small that influences the performance of the model; therefore, the $\rho$ value can be further improved when the number of active users increases. Finally, some cities, e.g., Beijing, have unstable user population and data noise due to the population shift.

3.5. Retention Effect. In the above correlation analysis, it is noticeable that there might be some retention effect in the detected physiological anomaly rate. To be specific, some anomalous people may continue to wear their devices so that they are calculated as anomaly in multiple days. This results in statistic error during the correlation analysis.

In order to analyze the impact, we calculate the retention rates of people being detected as anomaly for several consecutive days. As shown in Figure 7, if a person is detected as anomaly on a certain day, the possibility of wearing the device is decreasing gradually from 3.5% down to 0.2% in the following 4 days. This indicates that the retention effect may have very limited influence on the correlation analysis.
In this study, a prediction model of COVID-19 epidemic trend has been realized using physiological data collected by wearable devices. The results show that prediction with dynamic physiological data may have the advantage to alert the infection outbreak in advance. However, the detection method for calculating the physiological anomaly rate has some limitations. 

Firstly, on holidays, e.g., Chinese Spring Festival, Christmas, etc., transportation and population shift, social activities, and alcohol drinking might greatly influence the physiological signs of the users. For example, the elevated RHR due to heavy drinking on holidays might persist several days and greatly influence the physiological anomaly rate to be detected. Especially for China, the outbreak time of COVID-19 and influenza-like illness overlap with Chinese Spring Festival. Thus, it is necessary to distinguish the elevated RHR cases induced by holiday activities from infection.

Secondly, the anomaly rate is the statistical description of wearable device users’ physiological signs in anomaly. The validity of the statistical description depends on both the user scale and diversity. For example, for a city with 0.1% COVID-19 number
officially reported infection rate of COVID-19, if the number of active users in the city is less than 10,000, there might be only 10 people infected among them. Such scale of data cannot support a convincing inference. Regarding to the diversity, the prediction accuracy can be greatly improved if the distribution of active users is consistent with the natural distribution. For example, since elderly people and people with other diseases, e.g., Cardiovascular Disease (CVD), are more susceptible to COVID-19 [2, 3], the statistical performance of the model will be influenced if there is not enough coverage of such people.

Thirdly, although the current study provides the population evolution model for public health surveillance, it may be more meaningful for medical workers as well as individuals to take early precautions, if individualized health status prediction model is available. In the future, such prediction model based on wearable device data will be explored by incorporating more individual features, such as age, gender, BMI (Body Mass Index), etc.

5. Conclusions

The public health emergency could cause severe damage to the health and prosperity of our society. The popularity of wearable device provides the opportunity for the researchers to utilize big health data for public health emergency surveillance. In this study, a COVID-19 prediction framework using the health data from wearable devices was put forward. The proposed model could well predict the epidemic trend of COVID-19 outbreak in various countries and cities. The results from the study may shed light on a nationwide solution for the infectious disease surveillance system.

Data Availability

The concerned sensor data cannot be shared due to user privacy. For academic purposes, de-identified region-level statistics can be shared under agreement.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Huami Corporation.

References

[1] World Health Organization, Coronavirus Disease (COVID-2019) Situation Reports, World Health Organization, Geneva, Switzerland, 2020, https://www.who.int/emergencies/diseases/novelcoronavirus-situation-reports.
[2] Z. Wu and J. M. McGoogan, “Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China,” JAMA, vol. 323, 2020.
[3] W. Guan, Z. Ni, Y. Hu et al., “Clinical characteristics of 2019 novel coronavirus infection in China,” New England Journal of Medicine, vol. 395, pp. 1708–1720, 2020.
[4] J. F.-W. Chan, S. Yuan, K.-H. Kok et al., “A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster,” The Lancet, vol. 395, no. 10223, pp. 514–523, 2020.
[5] N. Chen, M. Zhou, X. Dong et al., “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study,” The Lancet, vol. 395, no. 10223, pp. 507–513, 2020.
[6] Z. Yang, Z. Zeng, K. Wang et al., “Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions,” Journal of Thoracic Disease, vol. 12, 2020.
[7] J. T. Wu, K. Leung, and G. M. Leung, “Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study,” The Lancet, vol. 395, no. 10225, pp. 689–697, 2020.
[8] S. Zhao, Q. Lin, J. Ran et al., “Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak,” International Journal of Infectious Diseases, vol. 92, pp. 214–217, 2020.
[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” Nature, vol. 457, no. 7232, pp. 1012–1014, 2009.
[10] New York Times, Google Uses Searches to Track flu’s Spread, New York Times, New York, NY, USA, 2008, http://www.nytimes.com/2008/11/12/technology/internet/12flu.html?_r=1.
[11] A. F. Dugas, Y.-H. Hsieh, S. R. Levin et al., “Google Flu Trends: correlation with emergency department influenza rates and crowding metrics,” Clinical Infectious Diseases, vol. 54, no. 4, pp. 463–469, 2012.
[12] M. J. Paul, M. Dredze, and D. Broniatowski, “Twitter improves influenza forecasting,” PLoS Currents, vol. 6, 2014.
[13] Q. Yuan, E. O. Nsoesie, B. Lv et al., “Monitoring influenza epidemics in China with search query from Baidu,” PLoS One, vol. 8, no. 5, 2013.
[14] K. Liu, T. Wang, Z. Yang et al., “Using Baidu search index to predict Dengue outbreak in China,” Scientific Reports, vol. 6, pages, 2016.
[15] M. Santillana, A. T. Nguyen, M. Dredze et al., “Combining search, social media, and traditional data sources to improve influenza surveillance,” PLoS Computational Biology, vol. 11, no. 10, 2015.
[16] Q. Xu, Y. R. Gel, L. L. R. Ramirez et al., “Forecasting influenza in Hong Kong with Google search queries and statistical model fusion,” PLoS One, vol. 12, no. 5, 2017.
[17] D. Lazer, R. Kennedy, G. King, and A. Vespignani, “The parable of Google Flu: traps in big data analysis,” *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[18] Y. Liu, H. Wang, W. Zhao, M. Zhang, H. Qin, and Y. Xie, “Flexible, stretchable sensors for wearable health monitoring: sensing mechanisms, materials, fabrication strategies and features,” *Sensors*, vol. 18, no. 2, p. 645, 2018.

[19] D. Dias and J. Paulo Silva Cunha, “Wearable health devices—vital sign monitoring, systems and technologies,” *Sensors*, vol. 18, no. 8, p. 2414, 2018.

[20] T. Arakawa, “Recent research and developing trends of wearable sensors for detecting blood pressure,” *Sensors*, vol. 18, no. 9, p. 2772, 2018.

[21] J. M. Radin, N. E. Wineinger, E. J. Topol, and S. R. Steinhubl, “Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study,” *The Lancet Digital Health*, vol. 2, no. 2, pp. e85–e93, 2020.

[22] X. Li, J. Dunn, D. Salins et al., “Digital Health: tracking physiomes and activity using wearable biosensors reveals useful health-related information,” *PLoS Biology*, vol. 15, no. 1, Article ID e2001402, 2017.

[23] C. Huang, Y. Wang, X. Li et al., “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,” *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[24] D. Wang, B. Hu, C. Hu et al., “Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China,” *JAMA*, vol. 323, 2020.

[25] Z. Xu, L. Shi, Y. Wang et al., “Pathological findings of COVID-19 associated with acute respiratory distress syndrome,” *The Lancet Respiratory Medicine*, vol. 8, 2020.

[26] J. Karjalainen and M. Viitasalo, “Fever and cardiac rhythm,” *Archives of Internal Medicine*, vol. 146, no. 6, pp. 1169–1171, 1986.

[27] L. Faust, K. Feldman, S. M. Mattingly et al., “Deviations from normal bedtimes are associated with short-term increases in resting heart rate,” *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[28] G. Ke, Z. Xu, J. Zhang et al., “DeepGBM: a deep learning framework distilled by GBDT for online prediction Tasks,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 384–394, Anchorage, AK, USA, August 2019.