scientific reports

OPEN

# Machine learning based regional epidemic transmission risks precaution in digital society

Zhengyu Shi[1], Haoqi Qian[2,3,4]✉, Yao Li[5], Fan Wu[6,7] & Libo Wu[4,8,9]✉

**The contact and interaction of human is considered to be one of the important factors affecting the epidemic transmission, and it is critical to model the heterogeneity of individual activities in epidemiological risk assessment. In digital society, massive data makes it possible to implement this idea on large scale. Here, we use the mobile phone signaling to track the users' trajectories and construct contact network to describe the topology of daily contact between individuals dynamically. We show the spatiotemporal contact features of about 7.5 million mobile phone users during the outbreak of COVID-19 in Shanghai, China. Furthermore, the individual feature matrix extracted from contact network enables us to carry out the extreme event learning and predict the regional transmission risk, which can be further decomposed into the risk due to the inflow of people from epidemic hot zones and the risk due to people close contacts within the observing area. This method is much more flexible and adaptive, and can be taken as one of the epidemic precautions before the large-scale outbreak with high efficiency and low cost.**

The international society was caught off guard by the unexpected outbreak of COVID-19 since the beginning of 2020[1]. Quite a few regions or even the whole country had adopted the lockdown policies in order to make the pandemic under control, and these non-pharmacological interventions had been demonstrated to be effective[2]. Meanwhile, these strict interventions had caused severe economic and social welfare losses as well. The world GDP growth rate has declined by 3.4% in 2020[3] and around 81 percent of the global workforce was affected due to government responses to the pandemic[4]. However, when there lacks enough knowledge and information about the accurate transmission risks of COVID-19, the best policy response that policy makers may make is to try to cut off all possible transmission paths immediately.

Similar to other infectious diseases, spread of the COVID-19 is mainly resulted from direct, indirect and close contacts between people at the micro-level[5–7], as well as from regional population flows at the macro-level[8–10]. Existing literature has shown that transmission risks are predictable for infectious diseases such as Severe Acute Respiratory Syndrome (SARS)[11], Middle East Respiratory Syndrome (MERS)[12], Ebola[13] and flu[14] by using various behavioral data such as search engine[15,16], social media[17] and wearable devices[18,19]. Assumptions behind these predictions focus more on the macro-level so that people's contact behaviors are simplified as homogeneous parameters in traditional epidemiological models such as SI, SIR, SEIR and etc[14,20–23]. In such models, it is necessary to estimate the classic reproduction number accurately, that is, the number of secondary cases of infected individuals in the susceptible group[24–26]. But it is difficult to get the reproduction number which can fit the epidemic transmission evolution perfectly in the real world. The deviation between theory and reality can be explained by the differences of individual behaviors[27,28]. The neglect of the micro heterogeneity may lead to misestimating the regional epidemic risks. Moreover, the greater the difference of individual characteristics within the group, the greater the estimated error[29].

Most of the early studies are limited to the small-scale with few personnel and low population flow such as families[30], flights[31] and hospitals[32], but the premise assumption of this setting is too peculiar, so it is difficult to extend to the city and even the national level. The development of 5G and Internet of Things technology

[1]School of Data Science, Fudan University, Shanghai 200433, China. [2]Institute for Global Public Policy, Fudan University, Shanghai 200433, China. [3]LSE-Fudan Research Centre for Global Public Policy, Fudan University, Shanghai 200433, China. [4]MOE Laboratory for National Development and Intelligent Governance, Fudan University, Shanghai 200433, China. [5]Shanghai Ideal Information Industry (Group) Co., Ltd, Fudan University, Shanghai 200120, China. [6]Shanghai Public Health Clinical Center, Fudan University, Shanghai 200032, China. [7]Key Laboratory of Medical Molecular Virology, Fudan University, Shanghai 200032, China. [8]School of Economics, Fudan University, Shanghai 200433, China. [9]Institute for Big Data, Fudan University, Shanghai 200433, China. ✉email: qianhaoqi@fudan.edu.cn; wulibo@fudan.edu.cn

guarantees collection of individual trajectory data[33,34]. Therefore, many scholars try to obtain large-scale information about people contacts through wearable devices or mobile phones[35,36], which creates more opportunities for further researches on epidemic transmission path and risk, seasonal fluctuation and spatial evolution and so on[37–39]. However, those previous epidemic studies focused more on the observed population migration between cities, base stations or some grid units, as well as the population density within a certain region[10,40]. This will inevitably lead to the fact that people in the same space are supposed to be homogeneous and static when considering regional risks, while ignoring the actual situation of dynamic contact among them.

In order to provide enough evidence at high-resolution level for policy makers to take targeted measures, heterogeneous individual level contact behaviors have been put more emphases on[41], and some intelligent technology like big-data analytics[42], artificial intelligence[43], cloud computing[44] and machine learning[45] may provide better solutions. During the spreading period of COVID-19, many countries have tried to launch individual tracing systems and monitor potential transmission risks through smart phones[46,47], and Apple and Google also developed COVID-19 Alert App jointly[48]. All of those applications need users be willing to install or use, otherwise they cannot offer the pandemic information for users. Since data coverage is more crucial for epidemic prevention, this invasive data acquisition way will reduce the prevention efficiency. Except for the individual trajectory, the contact topology is also quite important. As a straightforward scientific tool, complex network can effectively describe the dynamic contact topology between different individuals[49,50], so as to obtain more micro scale discovery of epidemic transmission. For example, individual level contact network tends to show small-world and nonrandom graph properties[51,52]. These features reflect the fact that more complicated models[53,54] are anticipated to investigate the micro mechanisms of infectious disease transmission. Then regional transmission risks can be more precisely identified by adopting comprehensive population flow pattern data. This type of bottom-up transmission risk modelling techniques has shown increasing importance in the policy making procedure in the public health field[55]. Another advantage of using big data in practice is that it can reduce unnecessary intrinsic risks in the traditional epidemiological surveys[56,57]. These risks are commonly caused by missing some part of objective information due to memory biases or dishonesties.

To overcome the aforementioned issues, we construct a novel contact network structure based on mobile phone signaling. It establishes a weighted contact topology network in a non-intrusive way and can reflect the difference of social interaction better. Since the data in the real world often suffer from the highly imbalanced distribution, the traditional methods can hardly deal with that[58–60]. For this reason, if conventional neural network is used to identify rare disease patients or high-risk virus carriers from a large number of negative people, the results will be seriously biased. Thus, after reconstructing the individual-centered contact feature, the neural network prediction of extreme events is carried out for each individual. In this study, we estimate the town-level transmission risks for COVID-19 in Shanghai based on a high-resolution contact network compiled from nearly 7.5 million mobile phone users. Individual level contact behaviors are modelled by using the machine learning method. Results show that this machine learning based bottom-up technique has great potential for identifying regional transmission risks. The interesting conclusions provide policy implications that unnecessary economic and welfare losses can be avoided by controlling the spread of infectious diseases in advance.

## Methods and data

### Contact strength.

There are 94,733 Telecom base stations in Shanghai with an average coverage of 0.0669 square kilometers. We define that if two mobile phone signals interact with one base station at the same time slice τ, then the two individuals' trajectories have a coincidence. In this paper, the time slice τ is set to 1/12 hour. If individuals coincide with high-risk group, the risk of infection will increase, and consequently such contacts are called effective contacts; while the mutual contacts within the general group do not generate new risks of infection, such contacts are invalid. In order to simplify the contact analysis, it is necessary to concentrate on the effective contacts when identifying regional transmission risks of infectious diseases.

Furthermore, we constructed the contact strength to quantify the influence of effective contacts. Effective contact frequency is one of the determinants to increase the infectious transmission risks. The longer an individual has been exposed to the high-risk group, the more likely to be infected. Nevertheless, only considering the duration of effective contact is not enough. Since the individuals in high-risk group have been to different epidemic hot zones, the possibilities of carrying virus are distinct and we use a dynamic virus carrying risk coefficient to distinguish one from another. Thus, the contact strength can be calculated by the product of virus carrying risk coefficient from high-risk individual $h$ and effective contact frequency,

$$\omega_{h\to i,d} = t_{h,i,d} \times \gamma_{h,d} \tag{1}$$

where, $\omega_{h\to i,d}$ represents the contact strength between individual $i$ and individual $h$ on day $d$, which will be the weight of corresponding edge in the $d$ th contact network. $t_{h,i,d}$ is the times of effective contacts between individual $i$ and $h$ on the $d$ th day.

The virus carrying risk coefficient $\gamma_h$ of individual $h$ is determined by the epidemic hot zone with the highest risk coefficient in the recent viral incubation period $T_{virus}$. First of all, we define the epidemic infection density $\rho_c$ of epidemic hot city $c$ as the proportion of the cumulative number of confirmed cases in the permanent population,

$$\rho_c = \frac{Nc_c}{Np_c} \tag{2}$$

where, $Nc_c$ is the cumulative number of confirmed cases in city $c$, and $Np_c$ is the permanent resident population of city $c$ (unit: 10,000 people). We set infection density $\rho_o$ of the city with the transmission risk to be estimated as the baseline and adjust the other cities' infection densities, so as to obtain the risk coefficient for travelling or living in city $c$,

$$r_c = \frac{\rho_c}{\rho_o} \tag{3}$$

where, $r_c$ is the risk coefficient of travelling or living in city $c$, and $r_o$ is the risk coefficient of the city to be estimated. Obviously, $r_o = 1$. Therefore, the $\gamma_{h,d}$ is equal to the maximum value of $r_c$ in the historical trajectory of individual $h$ counting down $T_{virus}$ from day $d$.

**Contact networks.** In order to simulate the risks of spread infectious diseases in the crowd better, we proposed a growing network based on the microscopic spatiotemporal contact details among individuals, which called contact network. In this contact network, every mobile phone user is a node. Only when the effective contact occurs, the corresponding nodes will form an edge and the weight of the edge is their contact strength. As shown in Fig. 1a, the red dots indicate individuals of high-risk group and green dots indicate individuals of general group. At time $T$, there are two high-risk individuals under Station 1 and they have effective contacts (red line) with other people under the same station; while the other contacts are invalid (green line). And under Station 2, all people are belonging to the general group, so there is no effective contact. Thus, each base station forms a sub-network. After a time slice τ, some individuals move from one station to another, and then each base station generate a new sub-network following by the latest contacts. With people moving across the base stations during one day, such sub-network will be generated continuously. At the end of the day, all of the effective contacts and the nodes to which they are connected eventually form a daily contact network. Obviously, people who do not have contacted with the high-risk group are not included in the contact network.

Because the contact network describes the possible path of epidemic spreading in detail, we can further learn the transmission risk based on artificial neural network. The purpose of transmission risk learning is to identify individuals with higher potential infectious risk and estimate the corresponding probabilities. Here we mainly consider the first layer of virus transmission risks, that is, the infection between adjacent nodes in contact network. Therefore, as shown in Fig. 1b, all contact networks within nearly $T_{virus}$ days are transformed into individual-centered single-layer networks. $T_{virus}$ is the latent period of the infection and the potential risk of carrying virus can be taken into account by selecting the contact networks during the $T_{virus}$. And then, we extract contact feature sequences from those single-layer networks as the input of artificial neural network. Each contact feature sequence consists of two element sequences: $TF$, which represents the total contact strength, and $K$, which indicates whether the individual has contacted with the confirmed cases,

$$TF_{i,j,d} = \sum_{h \in H_{i,j,t}} \omega_{h \to i,d} \tag{4}$$

$$K_{i,j,d} = \begin{cases} 1 & if\ \ there\ are\ confirmed\ cases\ in\ H_{i,j,d} \\ 0 & otherwise \end{cases} \tag{5}$$

where $i$ is an individual, $j$ is the municipal district of the city to be estimated and $d$ is the time. Thus, $TF_{i,j,d}$ indicates the contact intensity between individual $i$ and high-risk group in area $j$ on day $d$, which is the sum of edge weights of corresponding nodes in contact network. $H_{i,j,d}$ is the subset of high-risk group who had contact with individual $i$ in area $j$ on day $d$ effectively. If there is a confirmed case in subset $H_{i,j,d}$, then $K_{i,j,d}$ equals 1, otherwise it is 0.

**Artificial neural network of extreme events.** Artificial neural network is used to learn epidemic transmission risk. After completing the feature transformation of contact network nodes, we calculate the cross term of contact intensity $TF$ and contact tag $K$. These three variables are standardized and then used as the input variables of the neural network. And then, we label the high-risk people by the potential risks' sources. Those isolated people are divided into two categories according to whether they had a sojourn to epidemic hot zone. If people have not been to the epidemic hot zone, their infection risks come from the contact in the observing area. In contrast, people who have been to the epidemic hot zone, the regional transmission risk comes from the epidemic hot zone people inflow. The rest individuals of the high-risk group are labeled as the third category.

As shown in Fig. 2, the basic framework of the network is fully-connected and adopts leaky ReLU as activation function to reduce the silent neurons. However, isolation is an extreme event, that is, the proportion of positive-marked data in the dataset is very low. The high-risk group accounts for a very small number of the total population, let alone those who are isolated. Due to the imbalance of three kinds of people, it is necessary to adjust the neural network in the multi-classification training[61–63]. Therefore, in order to avoid the prediction error of the true positive cases caused by imbalanced data training, the neural network adopts a weighted cross entropy $L(Y, P)$ as the loss function for extreme event learning,

$$L(Y,P) = -\frac{1}{N}\sum_i \left( w_k \sum_k y_{i,k} \log p_{i,k} \right) \tag{6}$$

where $N$ is the size of training sample, $k$ marks different classes. $y_{i,k}$ indicates whether the individual $i$ belongs to class $k$, if so, it is 1; otherwise, it is 0. $p_{i,k}$ is the probability that the model predicts individual $i$ belonging to class $k$ and $w_k$ is the weight of class $k$.

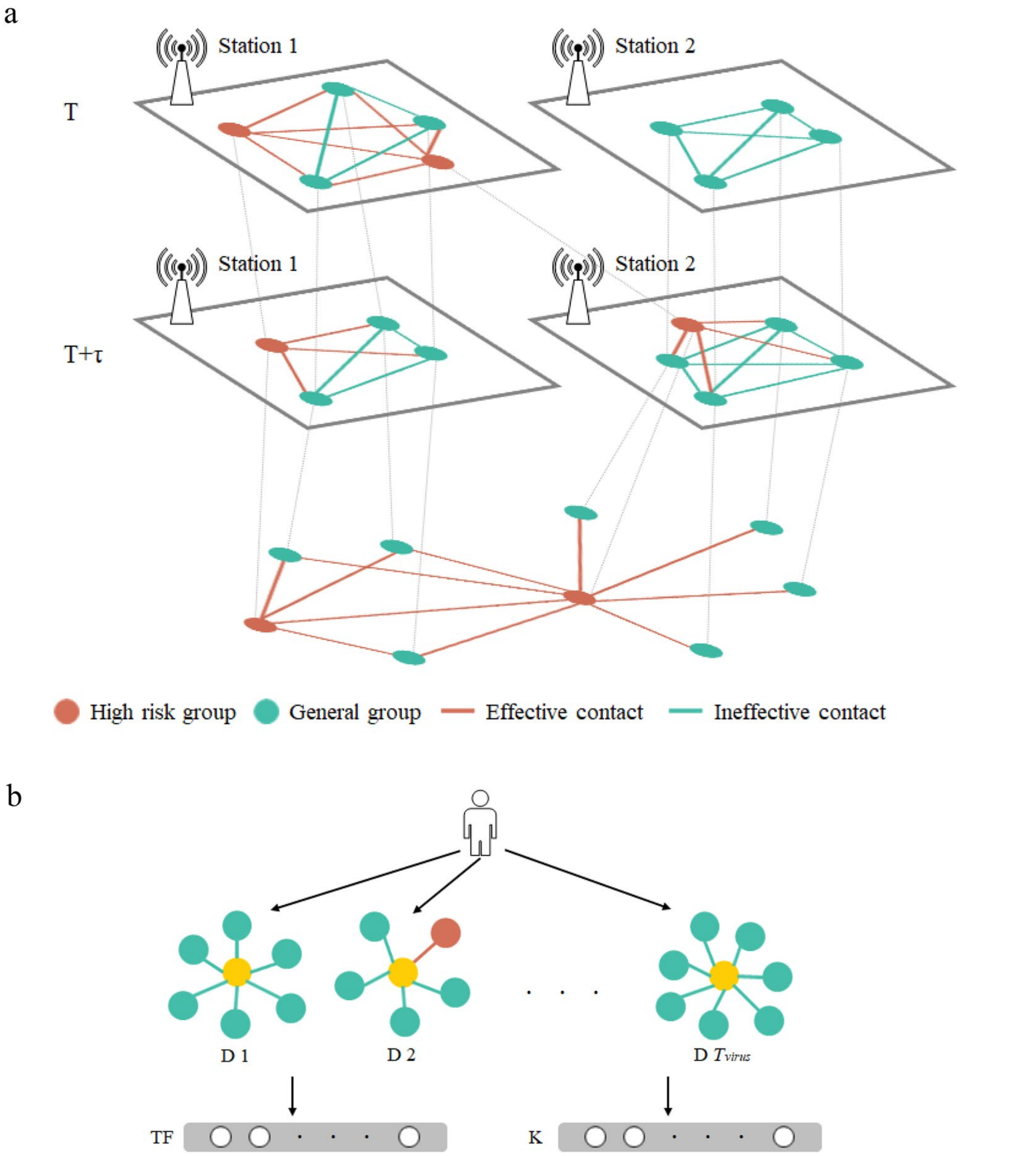

**Figure 1.** Contact networks structure. (**a**) Schematic diagram of sub-networks and contact network, taking two base stations as an example. Note that this figure only shows the trajectory simulation of two high-risk individuals and seven general individuals during two time slices, but in fact, each contact network is composed of $24/\tau$ sub-networks of all base stations. (**b**) Visualization of individual-centered contact feature sequence transformation. Before learning the transmission risk, the model takes each individual as an observation object and extracts the contact information of the adjacent nodes from the contact networks within $T_{virus}$.

This loss function can give larger weight to the rarer categories, that is to say, the corresponding $w_k$ of the isolated groups are larger in order to increase the misclassified cost of these two rare categories, so that the neural network can learn useful information more effectively and achieve better prediction results.

After normalizing the initial learning results of neural network by *Softmax*, the probability that individual $i$ belongs to each class can be obtained. The class with the largest probability is the prediction class of individual $i$.

**Estimation of regional transmission risk.** The main residence of each individual is determined by their most frequently located region for mobile phone signals during the night. Thus, we can divide those people into different group in terms of their residences. The risks of infectious disease transmission will come from the activities of people living there.

Since we have labeled the high-risk people as three categories and used multi-classification learning to fit how likely these people are to belong to the certain category, risk due to epidemic hot zones people inflow and risk due to close contacts are the average probability of corresponding-labeled individuals settled here,

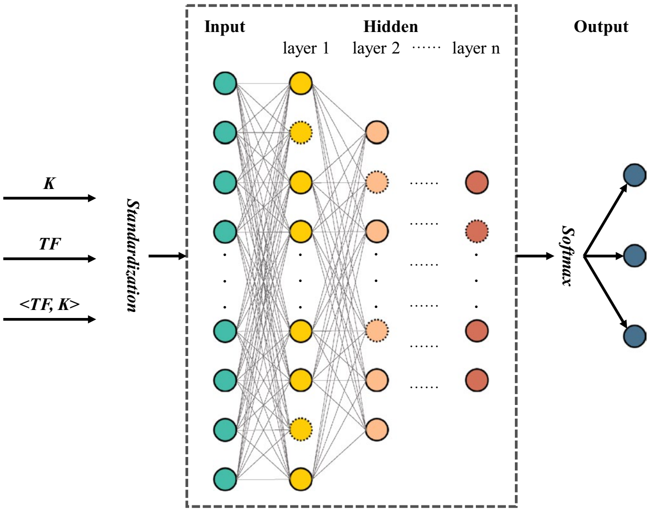

**Figure 2.** Artificial neural network structure. Visualization of neural network learning. After max–min scaling the input variables, the contact features can be learned by a fully-connected neural network. During model training, some neurons (dotted dots) are temporarily discarded from the network according to a certain probability, so that the network can avoid over fitting and be generalized better.

$$TR_s^{(inflow)} = \frac{1}{N_s}\sum_{i=0}^{N_s} p_{i,s}^{(ehz)} \tag{7}$$

$$TR_s^{(contact)} = \frac{1}{N_s}\sum_{i=0}^{N_s} p_{i,s}^{(non)} \tag{8}$$

where $s$ is the region of risk to be assessed, $N_s$ is the number of individuals settled in $s$. $p_{i,s}^{(ehz)}$ is the probability of disease transmission from epidemic hot zones caused by individual $i$ and $TR_s^{(inflow)}$ represents the risk caused by the inflow people from epidemic hot zones. Similarly, $p_{i,s}^{(non)}$ is the probability of disease transmission caused by individual $i$ who have not been to the epidemic hot zones and $TR_s^{(contact)}$ represents the risk caused by the close contacts within the observing region. It is obvious that $TR_s^{(inflow)}$ and $TR_s^{(contact)}$ are between 0 and 1, and larger values mean higher regional transmission risks.

Because of the properties of the *Softmax* function, the probabilities of no risk and other two risks are additive, and the sum of them is equal to one. Thus, the total transmission risk can be derived from $TR_s^{(inflow)}$ and $TR_s^{(contact)}$,

$$TR_s = TR_s^{(inflow)} + TR_s^{(contact)} = \frac{1}{N_s}\sum_{i=0}^{N_s}\left(p_{i,s}^{(ehz)} + p_{i,s}^{(non)}\right) \tag{9}$$

$TR_s$ ranges likewise from zero to one. Because this is a bottom-up indicator, the regional transmission risk will rise if individuals are more likely to be classified into potential isolated group. In the contrast, if most individuals are predicted as the non-isolated group, the regional transmission risk will decrease.

**Data.** We intercepted China Telecom's mobile signaling data in Shanghai from January 22 to February 4, 2020 to capture the users' real-time trajectories. We divided these 7,451,621 mobile phone users into high-risk group and general group according to their epidemiological diagnosis and historical action trails. High-risk group includes four kinds of people: the confirmed cases, the suspected cases, the medical isolators other than the first two and the people who once had a sojourn to epidemic hot zone. Considering the features of epidemic transmission and population flow in the early stage of COVID-19, forty-eight cities in China, including Wuhan and Wenzhou, were marked as the high-risk epidemic hot zones (see more details in Supplementary Information). And then, we identified 735,546 high-risk users in Shanghai based on mobile phone tracking during this period. In addition to the high-risk group, the rest of mobile phone users belonged to the general group.

As of February 4, 2020, there were 22,501 people in the isolation list provided by Shanghai Center for Disease Control and Prevention, covering eight districts in Shanghai. This eight districts include Baoshan, Chongming, Hongkou, Huangpu, Minhang, Pudong, Songjiang and Xuhui. Among them, 2459 isolators on the list were effectively matched, accounting for only 0.3343% of the Telecom high-risk users. In these matched isolators, 1742 isolators had epidemic hot zone sojourn and 717 isolators did not leave Shanghai during the observing period, accounting for 0.2368% and 0.0975% of the Telecom high-risk users respectively. For those users, being isolated was indeed a rare event.

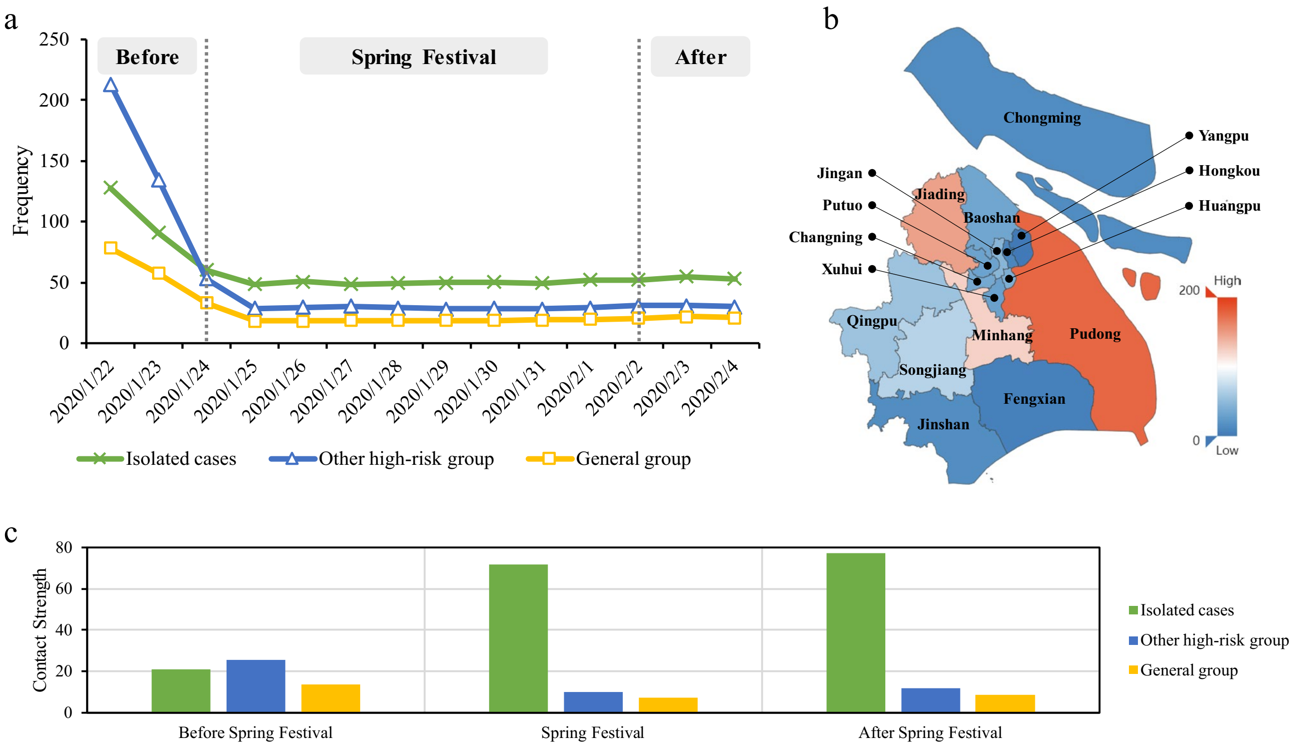

**Figure 3.** Crowd contact features of COVID-19 in the early stage. (**a**) The daily effective contact frequency per capita. The Spring Festival holiday in 2020 was originally from January 24 to January 30, then extended to February 2 due to COVID-19. On January 23, Wuhan announced the lockdown of the city and other local governments called on people to reduce unnecessary outdoor activities and maintain social distance. (**b**) Map of high-risk group's average effective contact frequency. (**c**) Contact strength before and after the Spring Festival. We divided the fourteen days into three slots: before the Spring Festival (January 22–January 23), the Spring Festival (January 24–February 2) and after the Spring Festival (February 3–February 4).

## Results

### Crowd contact based on mobile phone tracking.

According to the contact network, we can obtain the crowd contact characteristics on time trend, regional distribution and different groups. Generally, the spatiotemporal contact characteristics are consistent with the situation in Shanghai at that time, which also confirm the rationality of the contact network. In these 14 days, each high-risk individual was exposed to effective contact 51.81 times a day on average, while that of individuals in the general group was only 27.33 times. As shown in Fig. 3a, the effective contact frequencies of all isolators, non-isolated high-risk group and general group showed L-shaped as a whole. These curves declined at the beginning, and January 24 was a turning point. Since then, the curves have tended to be stable. On January 22, the effective contact frequency of each non-isolated high-risk individual was 213.36 times, which was higher than that of isolated group. However, the situation reversed since the Spring Festival. The daily effective contact frequency of non-isolated high-risk group has been lower than that of isolators, and closed to that of general group, maintained at about 30 times per capita. After February 3, people returned to work, and there was no apparent rebound in the effective contact frequency in Shanghai. This indicated that the policies of tighten travel restriction and keeping social distance called for by the government were well implemented. The average effective contact frequency of high-risk group (Fig. 3b) in Pudong was the highest, accounting for 176.36 times a day, which was partly due to the huge inter-cities population flow of Pudong Airport. Similarly, the effective contact frequency in Minhang, which has another airport and Hongqiao Railway Station with the largest passenger traffic volume in Shanghai, was also very high. As an important industrial area in Shanghai and a vital highway transportation hub connecting Jiangsu Province, Jiading had an effective daily contact frequency of 147.10 times a day. In contrast, the effective contact frequencies of high-risk group in suburbs such as Chongming and Fengxian, and urban centers such as Yangpu and Hongkou were much lower.

Figure 3c illustrates the differences in contact strength among isolators, non-isolated high-risk group and general group in three slots. Before the Spring Festival, the contact strength of the three groups was relatively close. However, the government had taken the quarantine measures, adopted the travel restriction policy and appealed for keeping social distance successively since January 23. These actions brought that the non-isolated high-risk group and the general group have reduced the contact strength by more than half during the Spring Festival. Although there was a slight increase after returning to work on February 3, the contact strength still remained low. Owing to gathering for medical observation, the isolated cases' mobile phone signaling would be received by the same base station frequently, which caused incessant effective contacts. In addition, most isolators have a larger virus carrying risk coefficient. Therefore, even if the overall population mobility in Shanghai declined, the contact strength of isolators still increased during the observation period continuously.

| | *l* | Weight* | | | | Baseline |
|---|---|---|---|---|---|---|
| | | 20% | 40% | 60% | 80% | |
| Isolator with epidemic hot zone sojourn | 0.01 | 70.85% | 64.57% | 63.82% | 65.33% | 1.89% |
| | 0.02 | 57.94% | 57.67% | 60.85% | 57.41% | |
| | 0.03 | 54.32% | 54.59% | 60.27% | 65.14% | |
| Isolator without epidemic hot zone sojourn | 0.01 | 18.35% | 27.52% | 24.77% | 31.19% | 0.00% |
| | 0.02 | 16.36% | 16.36% | 18.18% | 18.18% | |
| | 0.03 | 18.70% | 21.95% | 20.33% | 15.45% | |

**Table 1.** Recall of COVID-19 exposure risk in test set. *in order to study the influence of the loss function's weight on the exposure risk prediction better, the weights of rare categories are not set as a constant value, but set as the proportion of training set sample size.

**Neural network classification.** After constructing the Shanghai contact networks, we trained the neural network under different hyper-parameter settings. We took the neural network with general cross entropy loss function as the baseline and compared the classification results of the neural network with weighted cross entropy loss function with it (Table 1). Except for the loss function, the neural network structure of baseline is the same as that adopted in our extreme events model. Our classification goal is to accurately find true positives, but people who are really likely to be infected account for a small part of the population. A large proportion of negative cases will make many indicators such as accuracy fail. For example, even if all positive cases are classified as negative, accuracy will equal the proportion of negative cases in the samples and the result will show very well. In our data set, the accuracy will never be lower than 99%, which makes no sense to measure the quality of the model. Conversely, recall can evaluate whether all actual positive examples have been predicted and can support our study objectives better.

The baseline recalls of two isolated group are only 1.89% and 0.00% respectively. However, by using the advanced extreme event neural network model, 70.85% of the isolators with a sojourn to epidemic hot zone can be identified successfully. Even if the highest recall of isolators without sojourn to epidemic hot zone is only 31.19%, it is still significantly higher than that of baseline. The results show that our improved model is superior to the general neural network in extreme event prediction and can effectively identify the individuals who are included in the Shanghai CDC isolation list due to the different ways of contacts. According to the ablation experiment results, the model with Leaky ReLU slope of 0.01 and rare category weight proportion of 20% was selected to predict the exposure risk of all individuals.

**COVID-19 transmission risks in shanghai.** Since we have labeled the isolators into two categories, the regional transmission risk can be divided into the following two kinds correspondingly: one is the risk caused by the inflow of people from epidemic hot zones, and the other is the risk caused by close contacts within Shanghai. Figure 4a shows two types of COVID-19 transmission risks in Shanghai. As a whole, Shanghai transmission risk due to the epidemic hot zones' people inflow was 30.76%, among which Pudong, Fengxian, Jiading, Jinshan and Chongming exceed the city average risk. In contrast, the transmission risk due to epidemic hot zones in Songjiang was the lowest, only 12.07%. That's because Songjiang has a college town and large-scale industrial areas, a large number of students and migrant workers returned home as early as before the Spring Festival, and did not return until the observing period. Besides, Hongkou and Jing'an, which are located in the center of the city, have relatively low transmission risk due to the inflow from epidemic hot zones, which was about 15%. Meanwhile, the COVID-19 transmission risk due to close contacts in Shanghai was 7.9%, among which Qingpu, Putuo, Fengxian, Changning and Jinshan exceed the city average level. It is worth noting that the Shanghai Public Health Clinical Center is located in Jinshan, where all isolators get the medical care. The centralized medical isolation may be one of the critical reasons for the high transmission risk caused by close contacts in Jinshan.

It can be seen from Fig. 4b that the areas with high total transmission risk of COVID-19 were mainly concentrated at the border of Shanghai. Pudong's total risk was particularly high, reaching 66.55%. On the contrary, the total transmission risk in the center of Shanghai was relatively low. In terms of risk due to inflow from epidemic hot zones, the transmission risks in suburban streets were much higher than that in urban (Fig. 4c), especially the eastern, southern and northwestern borders of Shanghai. The streets of Pudong in particular deserve mention — the transmission risks from epidemic hot zones of most streets were all greater than 60% except for Lujiazui and other minority areas. The streets with high risk due to close contact (Fig. 4d) were mainly concentrated in the west of Shanghai, and Xianghuaqiao street of Qingpu had the highest risk, with a risk of 36.51%. In addition, some streets located in urban, such as Caoyangxincun street and Ganquanlu street, also have high risk, reaching 13.51% and 13.40% respectively.

## Discussion and conclusion

In this paper, a regional epidemic transmission risk precaution based on machine learning is proposed. Firstly, we distinguish whether individuals appear at the same time through the trajectories recorded by their mobile phones and construct the contact networks according to the way they contact. Then, the contact network is transformed into an individual-centered contact feature matrix, and the extreme event neural network is used to classify the isolated people. Finally, according to the classification results, we select the optimal model to predict the probability of each individual becoming a high-risk infected person and estimate the regional transmission risks.

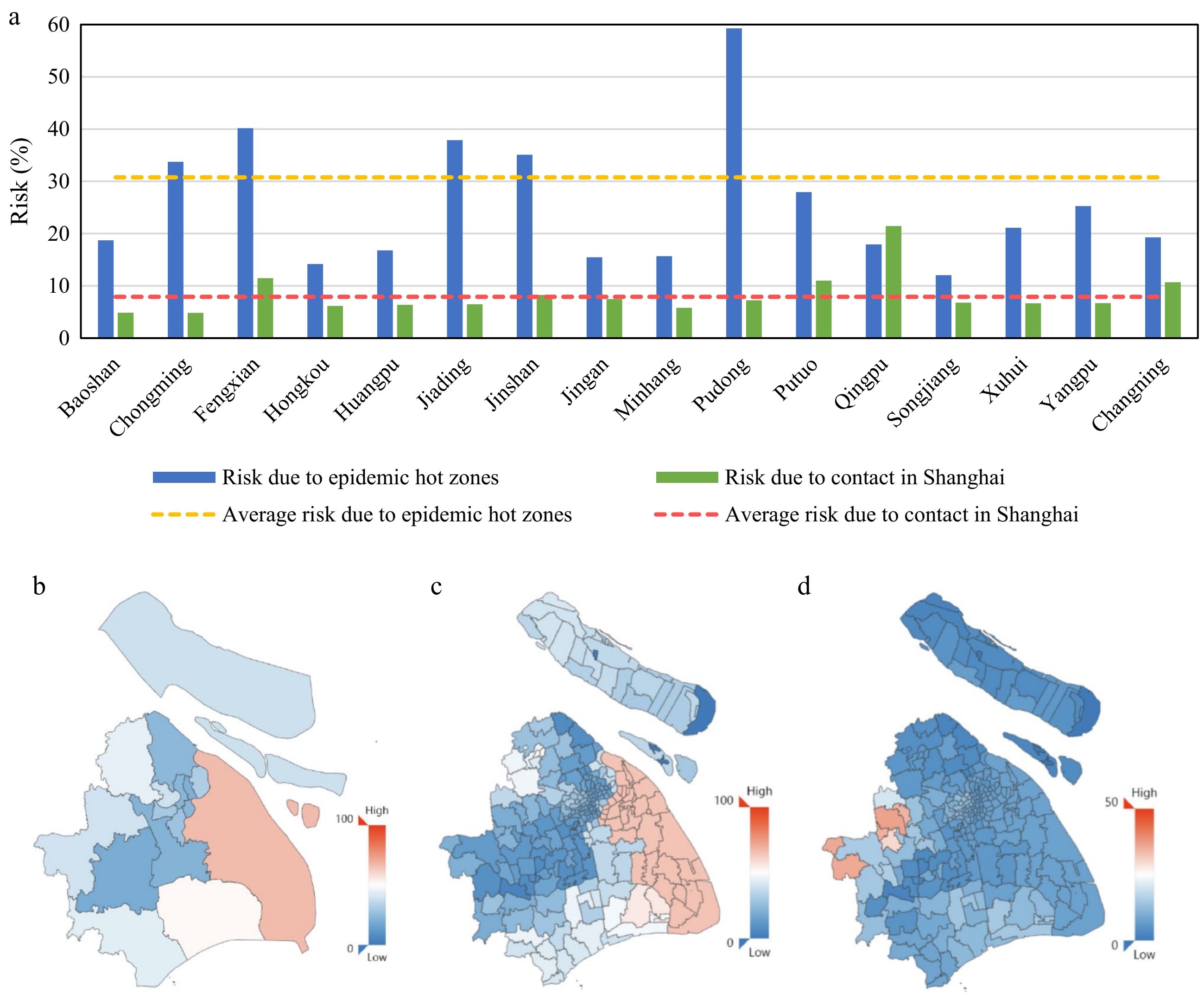

**Figure 4.** COVID-19 transmission risks in Shanghai. (**a**) Two kinds of COVID-19 transmission risks in 16 districts of Shanghai. (**b**) Map of Shanghai total transmission risk in district level. Baoshan, Jiading and Qingpu locate in the west of Shanghai, bordering Jiangsu Province; while Qingpu and Jinshan border on Zhejiang Province. According to Baidu Migration Index, Jiangsu and Zhejiang are the two major provinces of immigration to Shanghai from January 22 to February 4, 2020 (see more details in Supplementary Information). (**c**) Map of street-level transmission risk due to epidemic hot zones people inflow. (**d**) Map of street-level transmission risk due to close contacts within Shanghai.

We conducted a large-scale experiment with about 7.5 million people in Shanghai at the beginning of the COVID-19 outbreak in 2020. In the case of extremely imbalanced samples, the model can predict the rare categories effectively, and the recall can reach more than 70% among the isolators with epidemic hot zone sojourn. However, the recall of the isolators without high-risk areas sojourn history is only 31.19%, but it is still higher than that predicted by general neural network. On the one hand, this kind of isolators only accounts for 0.0975% of the samples. The scarcity of such isolators not only makes it difficult to capture their contact features, but also the proportions of various groups in the data set will be seriously unbalanced, which also can easily lead to model misjudgment. On the other hand, the coverage of the sample is insufficient. Considering that there were 40.92 million mobile phone users in Shanghai in 2020, the sample of China Telecom's mobile phone users is even less than one-fifth of Shanghai's mobile phone market. Nevertheless, the COVID-19 cases used in this study only cover half of Shanghai and the cases in the other eight districts are not taken into account. Due to the limitation of experimental data, the whole population's contact situation in Shanghai was not fully described when constructing contact network, which will have a negative impact on the prediction results of the model. However, as a whole, the precaution framework is of great significance for the regional transmission risk estimation of COVID-19 and other similar epidemics.

Artificial intelligence has been widely adopted in many fields in our real life[64–66], including the prevention and control of infectious diseases. Different from the previous studies on epidemic transmission through wearable devices or mobile phones, this machine learning based regional epidemic transmission risk precaution is completely bottom-up and can be used for early warning of regional epidemic on the premise of anonymity.

When facing the changes of regional isolation and flow restriction policies[67,68], which are very common in reality, this method has better flexibility and can make self-adaptive adjustment. In addition, using mobile phone signaling to estimate the risk of regional epidemic spread can provide effective auxiliary information support for government policy making and epidemic prevention work with high efficiency and low cost. Especially for low-income and middle-income countries, it can alleviate the financial difficulties caused by epidemic prevention and control. In order to implement effective intervention measures, it requires close interaction between policy makers and model prediction during the outbreak of epidemic[69]. But, remarkably, digital governance has raised the global concern on the citizens' privacy protection when using public data[70–72]. Therefore, all countries need to strictly abide by the data privacy law when using trajectory data and the scope of data usage should be limited in accordance with the minimization principle, including obtaining the explicit consent of users, collecting as little information as possible and ensuring data security. At the same time, the data and information holders should guarantee the data privacy through emerging technology, such as desensitizing data, and reduce the possibility of data abuse from the source[73].

As mentioned above, this method was proposed for regional epidemic transmission risk precaution. Its main purpose is to provide early warning before a large-scale epidemic outbreak and provide auxiliary information to decision makers. Labor loss, production suspension, trade obstruction, and rising market uncertainty may all become the consequences of national epidemic prevention policies. If the governors cannot balance the control measures and economic pressures well, an economic crisis may follow the pandemic[74]. By controlling the risk before the virus spreads widely, governors can moderate the enormous economic and social disruption caused by control measures for infectious diseases. Thus, the research design mainly focuses on the contact network and extreme events classification. On the one hand, we pay attention to the inflow risk from the external epic hot zone when calculating the contact strength; on the other hand, the transmission risk has a relatively long window period (14 days in the experiment), which has an impact on the contact strength and the individual centered contact feature. As a proactive prevention and control method, the best time for it to work is when there are only a few infected people because of the aforementioned mechanism design. Generally, the limitation of this risk precaution is that when a large-scale and mass outbreak occurs in the city, such as the Omicron virus pandemic in Shanghai in the spring of 2022, its early warning effect will be greatly reduced. The outbreak of the Omicron virus pandemic this time is so sudden that social resources such as the CDC, public health departments, communication operators and so on are fully occupied. Therefore, it is worthy to retrospectively analyze the differences of these two outbreaks in the future.

## Data availability

The data that support the findings of this study are available from Shanghai Ideal Information Industry (Group) Co., LTD but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Shanghai Ideal Information Industry (Group) Co., LTD.

Received: 29 March 2022; Accepted: 18 November 2022
Published online: 28 November 2022

## References

1. Peters, B. G. Governing in a time of global crises: the good the bad and the merely normal. *Glob. Public Policy Gov.* **1**(1), 4–19. https://doi.org/10.1007/s43508-021-00006-x (2021).
2. Haug, N. *et al.* Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* **4**, 1303–1312 (2020).
3. World Bank. GDP growth (annual %). https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?
4. ILO. *COVID-19 and the world of work*. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/briefingnote/wcms_740877.pdf (2020).
5. Wesolowski, A. *et al.* Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proc. Natl. Acad. Sci.* **112**, 11114–11119 (2015).
6. Bi, Q. *et al.* Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919 (2020).
7. Nishiura, H. *et al.* The rate of underascertainment of novel coronavirus (2019-nCoV) infection: Estimation using Japanese passengers data on evacuation flights. *J. Clin. Med.* **9**, 419 (2020).
8. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
9. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the COVID-19 epidemic?. *The Lancet* **395**, 931–934 (2020).
10. Jia, J. S. *et al.* Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* https://doi.org/10.1038/s41586-020-2284-y (2020).
11. Meyers, L. A., Pourbohloul, B., Newman, M. E. J., Skowronski, D. M. & Brunham, R. C. Network theory and SARS: Predicting outbreak diversity. *J. Theor. Biol.* **232**, 71–81 (2005).
12. Yang, C. H. & Jung, H. Topological dynamics of the 2015 South Korea MERS-CoV spread-on-contact networks. *Sci. Rep.* **10**, 1–11 (2020).
13. Wesolowski, A. *et al.* (2014) Commentary: Containing the Ebola outbreak-the potential and challenge of mobile network data. PLoS Curr. **6**.
14. Liu, Q.-H. *et al.* Measurability of the epidemic reproduction number in data-driven contact networks. *Proc. Natl. Acad. Sci.* **115**, 12680–12685 (2018).
15. Yang, S., Santillana, M. & Kou, S. C. Accurate estimation of influenza epidemics using google search data via ARGO. *Proc. Natl. Acad. Sci.* **112**, 14473–14478 (2015).
16. Dukic, V., Lopes, H. F. & Polson, N. G. Tracking epidemics with google flu trends data and a state-space SEIR model. *J. Am. Stat. Assoc.* **107**, 1410–1426 (2012).

17. Samaras, L., García-Barriocanal, E. & Sicilia, M.-A. Comparing social media and google to detect and predict severe epidemics. *Sci. Rep.* **10**, 1–11 (2020).
18. Radin, J. M., Wineinger, N. E., Topol, E. J. & Steinhubl, S. R. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: A population-based study. *Lancet Digit. Health* **2**, e85–e93 (2020).
19. Salathé, M. *et al.* A high-resolution human contact network for infectious disease transmission. *Proc. Natl. Acad. Sci.* **107**, 22020–22025 (2010).
20. Huppert, A. & Katriel, G. Mathematical modelling and prediction in infectious disease epidemiology. *Clin. Microbiol. Infect.* **19**, 999–1005 (2013).
21. Wang, H. *et al.* Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan China. *Cell Discov.* **6**, 1–8 (2020).
22. Kucharski, A. J. *et al.* Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *Lancet Infect. Dis.* https://doi.org/10.1016/S1473-3099(20)30144-4 (2020).
23. Kuniya, T. Prediction of the epidemic peak of coronavirus disease in Japan, 2020. *J. Clin. Med.* **9**, 789 (2020).
24. Ghostine, R., Gharamti, M., Hassrouny, S. & Hoteit, I. An extended SEIR model with vaccination for forecasting the COVID-19 pandemic in Saudi Arabia using an ensemble kalman filter. *Mathematics* **9**, 636 (2021).
25. Qian, X. & Ukkusuri, S. V. Connecting urban transportation systems with the spread of infectious diseases: A Trans-SEIR modeling approach. *Transp. Res. Part B Methodol.* **145**, 185–211 (2021).
26. Piccirillo, V. Nonlinear control of infection spread based on a deterministic SEIR model. *Chaos Solitons Fract.* **149**, 111051 (2021).
27. Scabini, L. F. S. *et al.* Social interaction layers in complex networks for the dynamical epidemic modeling of COVID-19 in Brazil. *Phys. Stat. Mech. Appl.* **564**, 125498 (2021).
28. Lyu, Z. & Takikawa, H. The disparity and dynamics of social distancing behaviors in Japan: Investigation of mobile phone mobility data. *JMIR Med. Inform.* **10**, e31557 (2022).
29. Diez Roux, V. A. The study of group-level factors in epidemiology: Rethinking variables, study designs, and analytical approaches. *Epidemiol. Rev.* **26**, 104–111 (2004).
30. Saunders, M. J. *et al.* A household-level score to predict the risk of tuberculosis among contacts of patients with tuberculosis: A derivation and external validation prospective cohort study. *Lancet Infect. Dis.* **20**, 110–122 (2020).
31. Litvinova, M., Liu, Q.-H., Kulikov, E. S. & Ajelli, M. Reactive school closure weakens the network of social interactions and reduces the spread of influenza. *Proc. Natl. Acad. Sci.* **116**, 13174–13181 (2019).
32. Anderson, R. M. & May, R. M. *Infectious diseases of humans: Dynamics and control* (OUP Oxford, 1992).
33. Hu, N., Tian, Z., Lu, H., Du, X. & Guizani, M. A multiple-kernel clustering based intrusion detection scheme for 5G and IoT networks. *Int. J. Mach. Learn. Cybern.* **12**, 3129–3144 (2021).
34. Lu, H. *et al.* DeepAutoD: Research on distributed machine learning oriented scalable mobile communication security unpacking system. *IEEE Trans. Netw. Sci. Eng.* **9**, 2052–2065 (2022).
35. Vigfusson, Y. *et al.* Cell-phone traces reveal infection-associated behavioral change. *Proc. Natl. Acad. Sci.* **118**, e2005241118 (2021).
36. Hijazi, H. *et al.* Wearable devices, smartphones, and interpretable artificial intelligence in combating COVID-19. *Sensors* **21**, 8424 (2021).
37. Lu, F. S., Hattab, M. W., Clemente, C. L., Biggerstaff, M. & Santillana, M. Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. *Nat. Commun.* **10**, 147 (2019).
38. Wesolowski, A., Buckee, C. O., Engø-Monsen, K. & Metcalf, C. J. E. Connecting mobility to infectious diseases: The promise and limits of mobile phone data. *J. Infect. Dis.* **214**, S414–S420 (2016).
39. Bengtsson, L. *et al.* Using mobile phone data to predict the spatial spread of cholera. *Sci. Rep.* **5**, 1–5 (2015).
40. Tokey, A. I. Spatial association of mobility and COVID-19 infection rate in the USA: A county-level study using mobile phone location data. *J. Transp. Health* **22**, 101135 (2021).
41. Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* https://doi.org/10.1126/science.abb6936 (2020).
42. Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: Challenges and promises of big data in healthcare. *Nat. Med.* **26**, 29–38 (2020).
43. Broad, J. D. & Luthans, F. Positive resources for psychiatry in the fourth industrial revolution: Building patient and family focused psychological capital (PsyCap). *Int. Rev. Psychiatry* **32**, 542–554 (2020).
44. Javaid, M. *et al.* Industry 4.0 technologies and their applications in fighting COVID-19 pandemic diabetes. *Metab. Syndr. Clin. Res. Rev.* **14**, 419–422 (2020).
45. Alimadadi, A. *et al.* Artificial intelligence and machine learning to fight COVID-19. *Physiol. Genomics* **52**, 200–202 (2020).
46. Tracking and tracing COVID: Protecting privacy and data while using apps and biometrics. OECD http://www.oecd.org/coronavirus/policy-responses/tracking-and-tracing-covid-protecting-privacy-and-data-while-using-apps-and-biometrics-8f394636/.
47. Jalabneh, R. *et al.* Use of mobile phone apps for contact tracing to control the COVID-19 pandemic: A literature review. In *Applications of Artificial Intelligence in COVID-19* Vol. 1 (eds Nandan Mohanty, S. *et al.*) 389–404 (Springer, Singapore, 2021). https://doi.org/10.1007/978-981-15-7317-0_19.
48. Google and Apple Reveal How Covid-19 Alert Apps Might Look. Wired.
49. Mata, A. S. An overview of epidemic models with phase transitions to absorbing states running on top of complex networks. *Chaos Interdiscip. J. Nonlinear Sci.* **31**, 012101 (2021).
50. Silva, C. J. *et al.* Complex network model for COVID-19: Human behavior, pseudo-periodic solutions and multiple epidemic waves. *J. Math. Anal. Appl.* **514**, 125171 (2022).
51. Wesolowski, A. *et al.* Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270 (2012).
52. Danon, L. *et al.* Networks and the epidemiology of infectious disease. *Interdiscip. Perspect. Infect. Diseases* **2011**, 284909 (2011).
53. Heesterbeek, H. *et al.* Modeling infectious disease dynamics in the complex landscape of global health. *Science* **347**(6227), aaa4339 (2015).
54. Wang, K. *et al.* Current trends and future prediction of novel coronavirus disease (COVID-19) epidemic in China: A dynamical modeling analysis. *Math. Biosci. Eng.* **17**, 3052–3061 (2020).
55. Scherer, C. *et al.* Moving infections: Individual movement decisions drive disease persistence in spatially structured landscapes. *Oikos* **129**(5), 651–667 (2020).
56. Bansal, S., Chowell, G., Simonsen, L., Vespignani, A. & Viboud, C. Big data for infectious disease surveillance and modeling. *J. Infect. Dis.* **214**, S375–S379 (2016).
57. Khoury, M. J. & Ioannidis, J. P. A. Big data meets public health. *Science* **346**, 1054–1055 (2014).
58. Herland, M., Bauder, R. A. & Khoshgoftaar, T. M. The effects of class rarity on the evaluation of supervised healthcare fraud detection models. *J. Big Data* https://doi.org/10.1186/s40537-019-0181-8 (2019).
59. Artetxe, A., Graña, M., Beristain, A. & Ríos, S. Balanced training of a hybrid ensemble method for imbalanced datasets: A case of emergency department readmission prediction. *Neural Comput. Appl.* **32**, 5735–5744 (2020).
60. Rath, A., Mishra, D., Panda, G. & Satapathy, S. C. Heart disease detection using deep learning methods from imbalanced ECG samples. *Biomed. Signal Process. Control* **68**, 102820 (2021).
61. Haixiang, G. *et al.* Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **73**, 220–239 (2017).

62. Nejatian, S., Parvin, H. & Faraji, E. Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification. *Neurocomputing* **276**, 55–66 (2018).
63. Razzaghi, T., Roderick, O., Safro, I. & Marko, N. Multilevel weighted support vector machine for classification on healthcare data with missing values. *PLoS ONE* **11**(5), e0155119 (2016).
64. Pan, M. *et al.* DHPA: Dynamic human preference analytics framework: A case study on taxi drivers' learning curve analysis. *ACM Trans. Intell. Syst. Technol.* **11**, 1–19 (2020).
65. Lu, H. *et al.* AutoD: Intelligent blockchain application unpacking based on JNI layer deception call. *IEEE Netw.* **35**, 215–221 (2021).
66. Lu, H. *et al.* Research on intelligent detection of command level stack pollution for binary program analysis. *Mob. Netw. Appl.* **26**, 1723–1732 (2021).
67. Bauch, C. T. & Galvani, A. P. Social factors in epidemiology. *Science* **342**, 47–49 (2013).
68. Funk, S., Salathé, M. & Jansen, V. A. A. Modelling the influence of human behaviour on the spread of infectious diseases: A review. *J. R. Soc. Interface* **7**, 1247–1256 (2010).
69. Pandey, A. *et al.* Strategies for containing Ebola in West Africa. *Science* **346**, 991–995 (2014).
70. Kubler, K. State of urgency: Surveillance, power, and algorithms in France's state of emergency. *Big Data Soc.* **4**, 2053951717736338 (2017).
71. Nay, O. Can a virus undermine human rights?. *Lancet Public Health* **5**, e238–e239 (2020).
72. Overton Sarah, M., Larson Lisa, J. & Carlson, S. Kleinschmit Public data primacy: the changing landscape of public service delivery as big data gets bigger. *Glob. Public Policy Gov.* https://doi.org/10.1007/s43508-022-00052-z.
73. Jia, K. & Chen, S. Global digital governance: paradigm shift and an analytical framework. *Glob. Public Policy Gov.* **2**(3), 283–305. https://doi.org/10.1007/s43508-022-00047-w (2022).
74. McKee, M. & Stuckler, D. If the world fails to protect the economy, COVID-19 will damage health not just now but also in the future. *Nat. Med.* **26**, 640–642 (2020).

## Acknowledgements

This work was supported by the National Science Foundation for Distinguished Young Scholars of China (71925010), the Science and Technology Commission of Shanghai Municipality Grant (20DZ1200600), the Three-year Action Program of Shanghai Municipality for Strengthening the Construction of Public Health System (GWV-1.1) and the Shanghai Talent Development Fund (2021098). We thank Shanghai Ideal Information Industry (Group) Co., LTD for providing access to the data used in this study and Fudan Association for Computational Social Science for providing technical supports.

## Author contributions

Z.S. Develop model and data processing, write the main content; H.Q. Develop model and data processing, design the paper structure, revise the main content; Y.L. Data processing; F.W. Design the study, collect the data; L.W. Design the study and the paper structure, revise the main content **.**

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-24670-z.

**Correspondence** and requests for materials should be addressed to H.Q. or L.W.

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

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