Be Aware of the Hot Zone: A Warning System of Hazard Area Prediction to Intervene Novel Coronavirus COVID-19 Outbreak

Zhenxin Fu*
Wangxuan Institute of Computer Technology, Peking University
Beijing 100871, China
fuzhenxin@pku.edu.cn

Yu Wu*
Peking University Health Science Center
Beijing 100083, China
yuwu@pku.edu.cn

Hailei Zhang
Laiye Technology Inc.
Beijing 100080, China
hailei@laiye.com

Yichuan Hu
Laiye Technology Inc.
Beijing 100080, China
will@laiye.com

Dongyan Zhao
Wangxuan Institute of Computer Technology, Peking University
Beijing 100871, China
zhaody@pku.edu.cn

Rui Yan†
Wangxuan Institute of Computer Technology, Peking University
Beijing 100871, China
ruiyan@pku.edu.cn

ABSTRACT

Dating back from late December 2019, the Chinese city of Wuhan has reported an outbreak of atypical pneumonia, now known as lung inflammation caused by novel coronavirus (COVID-19). Cases have spread to other cities in China and more than 180 countries and regions internationally. World Health Organization (WHO) officially declares the coronavirus outbreak a pandemic and the public health emergency is perhaps one of the top concerns in the year of 2020 for governments all over the world. Till today, the coronavirus outbreak is still raging and has no sign of being under control in many countries. In this paper, we aim at drawing lessons from the COVID-19 outbreak process in China and using the experiences to help the interventions against the coronavirus wherever in need. To this end, we have built a system predicting hazard areas on the basis of confirmed infection cases with location information. The purpose is to warn people to avoid of such hot zones and reduce risks of disease transmission through droplets or contacts. We analyze the data from the daily official information release which are publicly accessible. Based on standard classification frameworks with reinforcements incrementally learned day after day, we manage to conduct thorough feature engineering from empirical studies, including geographical, demographic, temporal, statistical, and epidemiological features. Compared with heuristics baselines, our method has achieved promising overall performance in terms of *precision*, *recall*, *accuracy*, *F1 score*, and *AUC*. We expect that our efforts could be of help in the battle against the virus, the common opponent of human kind.

CCS CONCEPTS

• **Information systems** → **Web applications**; • **Applied computing**; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Hazard area prediction, classification with reinforcement, epidemic feature engineering

ACM Reference Format:
Zhenxin Fu, Yu Wu, Hailei Zhang, Yichuan Hu, Dongyan Zhao, and Rui Yan. 2020. Be Aware of the Hot Zone: A Warning System of Hazard Area Prediction to Intervene Novel Coronavirus COVID-19 Outbreak. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401429

1 INTRODUCTION

The outbreak of novel coronavirus COVID-19 epidemic is perhaps the most serious public health crisis in recent years. Starting from late December 2019 in Wuhan, a city of China, the disease now spreads all over the world, causing global emergency right now. By March 31, 2020, the disease has been confirmed in more than 180 countries and regions, with infections of 750,890 confirmed cases, causing 36,405 deaths [1]. More importantly, the disease has not been fully under control, which makes the situation even worse. More and more countries (and regions) are facing the danger of possible escalation of infected patients and death victims.

People are now taking actions to fight against the epidemic outbreak. Doctors, nurses, and caretakers are striving to save life. On the other hand, virologists are racing against the time, analyzing the genetic map and characteristics of the virus while researchers of the pharmaceutical industry are working days and nights, seeking for cure and vaccine. For computer scientists, people are trying every effort to provide more information about the situation updates [2], news summarization and epidemic trend prediction [3].

*The first two authors contributed equally to this research.
†Corresponding author: Rui Yan (ruiyan@pku.edu.cn)

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401429

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/
[2] https://coronavirus.jhu.edu/map.html
[3] https://2019-ncov.aminer.cn/data

Now we know the novel coronavirus is airborne [15] and is extremely contagious through contacts (direct or indirect) by coughing, sneezing, and droplet transmission within a particular area [11]. To intervene the epidemic outbreak as early as possible, it plays a key role to detect—and then to block—the disease transmission routes. Thanks to the information transparency on the Web, we are able to acquire information reporting the infection cases from the government bulletins and other media. We are able to further analyze the data by extracting location information, activity trajectory, diagnosis time, and live updates of the disease. It is quite straightforward to come up with an approach to utilize such information to predict hazard areas and warn people to pay attention to the risk of contagion in these zones. People are suggested to avoid unnecessary contact with the warned locations or to take extra cautions with better protection in these areas. Perhaps, we are able to reduce community-level outbreaks and to some extent, to intervene the outbreak of the COVID-19 epidemic.

To this end, we conduct an empirical study to build a system to predict hazard areas as the precaution advice for both the government and the public. In particular, based on the information of infection cases, we predict the *hazard level* (from 1 to 5) of potential risks in different locations. We formulate the prediction task as a standard classification problem. To the best of our knowledge, we are the first to take efforts to predict hazardous areas in order to intervene the transmission and outbreak of COVID-19 virus.

Although a classification problem is well-defined, there still exist multiple challenges for the area prediction task:

• First of all, we need to distill data such as locations, time, and statistics, etc. by information extraction from unstructured texts. It is non-trivial to extract the data associated with infection cases. One of our contributions is to build such a dataset and then release the data to facilitate further studies.

• Secondly, we are still unclear about the mechanism and key schema for the COVID-19 disease. Without proper prior knowledge understanding, it is impractical to use machine learning regime to automatically extract features: we count on human experiences and expertise instead. Thus, a second contribution is that we investigate thorough feature engineering with various possible factors and characteristics to feed the classification models.

• Moreover, since the epidemic situation is changing from time to time, some model hyperparameters should be adjusted accordingly. The target is to train a model in the online scenario with data incrementally updated, and the third contribution is that we adapt the classification formulation with a reinforcement learning framework. In this way, the learning and prediction model can maintain to be up-to-date.

To sum up, our contribution is manifold by solving the above challenges. We build the warning system to predict hazard areas. The system consists of a streaming pipeline: extracting information from webpages, pre-processing data, computing features, training the model, and then predicting results. We hope the experience in China can be timely and helpful to other countries and regions in the world facing the COVID-19 epidemic outbreak.

2 RELATED WORK

2.1 Virology Studies

By the end of year 2019, the first atypical pneumonia case with unknown cause was reported in Wuhan, China. The patient works at the wholesale Hua'nan seafood market in Wuhan, where is generally assumed to be the starting spot of the new epidemic outbreak in China [23, 31]. Wu *et al.* identified the causative pathogen, a novel RNA virus strain, which belongs to the coronavirus family [27].

Phylogenetic analysis of the complete viral genome suggested that the novel coronavirus is highly similar (more than 80%) to a bat-derived Severe Acute Respiratory Syndrome (SARS)-like coronavirus (i.e., SARS-Cov) [16, 27]. Therefore, the coronavirus is then named as SARS-Cov-2. Among the several human-infecting coronaviruses, the SARS-Cov-2 is genetically related to but distinct from the original SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV) [16]. To infect cells, SARS-CoV-2 uses angiotensin converting enzyme2 (ACE2) as cell entry receptor, just like SARS-CoV do [19, 33]. Compared with the SARS outbreak in 2002 and MERS in 2012, COVID-19 disease outbreak by SARS-CoV-2 results in relatively milder symptoms with lower death rate, but has significantly stronger infectious capability [7, 31].

2.2 Transmission Studies

In the beginning, the novel coronavirus is believed to be limited human-to-human transmissible, until prominent evidence reveals clear human-to-human transmission [10, 31]: doctors and nurses are being infected. The initial basic reproductive number (R_0) was estimated as 2.2-3.8 [14, 22], which indicates that on average 2-3 cases are expected to generate directly by a particular case when absence of "any deliberate intervention" in the disease transmission. Unfortunately, the outbreak of the novel coronavirus disease COVID-19 has rapidly spread from China to other countries.

According to the number of cases exported from Wuhan internationally, Wu *et al.* proposed the estimation of the size of the epidemic in Wuhan [28]. Jung *et al.* statistically estimated the cumulative incidence and confirmed case fatality risk (cCFR) in mainland China also by analyzing the exported cases [12]. By fitting the statistic numbers into the epidemiology models such as the Susceptible-Exposed-Infectious-Recovered (SEIR) model [13], Peng *et al.* estimated the parameters in the SEIR model and forecasted the infection point and probable ending time [19].

Using a global meta-population disease transmission model, Chinazzi *et al.* indicated that travel restriction of Wuhan can modestly delay the epidemic trajectory in China, but can slow down the international epidemic progression more remarkably [3]. Studies also revealed that the world-wide lockdown policies could have more prominent impacts if combined with a reduction of city-level travels and commuting [5, 24, 28].

To this end, the idea inspires us that in combination with travel restrictions in global and local communities, a warning system indicating hazard areas can possibly make such policies more effective and more efficient: residents within high risk areas should conform to stay-at-home orders; people shall be advised to avoid unnecessary commuting to these locations; if possible, these areas shall be carefully monitored, and sanitized by sterilizing droplets in the environment.

3 HAZARD AREA PREDICTION SYSTEM

To introduce the warning system of hazard area prediction, we first briefly describe the system pipeline with processing component modules. Then we formulate the model and elaborate the features as well as the learning framework with reinforcements.

3.1 System Pipeline

The first step of the system is *Data Collection*. Then we crawled the webpages reporting the confirmed infection cases. In the component module of *Information Extraction*, we are able to obtain the timestamps, motion trajectories, residential location and other relevant information. The data of confirmed cases are totally anonymous, with no personal privacy leaks.

With the extracted data, we conduct the *Pre-processing and Filtering* module to keep the valid samples and rule out incomplete data which cannot be trained for location prediction. Then we train the classification module as the *Prediction Model*. Note that since new data stream in every day, the prediction model learning is actually an online setup and incrementally reinforced day after day. In this way, we update parameters and learn the model iteratively. Given the prediction results, we have a *Visualization Interface* component to show the results via online map [4]. Users are able to check the predicted areas with different levels of warning on the map directly.

3.2 Model Formulation

Since we target at predicting the hazard area to warn people to pay extra precaution to the possibly contagious zone, we formulate the task as labeling the specific location l with associated features $\mathbf{x}$ as "in hazard" (y=1) or "not in hazard" (y=0), which is a standard binary classification formulation learned as $f(y, \mathbf{x})$. It is intuitive to establish a pre-defined threshold (for example, 0.5) to determine a label when $f(\cdot)$ above the threshold and otherwise when $f(\cdot)$ below the threshold. In this paper, we characterize the hazard in a finer-granularity: given the labels, we train the classification function. When predicting the labels, we categorize the hazard in 5-level scale function Label$(\cdot)$:

$$\text{Label}(l) = \begin{cases} \text{Level-5} & \text{if} \quad 0.8 < f(y, \mathbf{x}) \leq 1; \\ \text{Level-4} & \text{if} \quad 0.6 < f(y, \mathbf{x}) \leq 0.8; \\ \text{Level-3} & \text{if} \quad 0.4 < f(y, \mathbf{x}) \leq 0.6; \\ \text{Level-2} & \text{if} \quad 0.2 < f(y, \mathbf{x}) \leq 0.4; \\ \text{Level-1} & \text{if} \quad 0.0 \leq f(y, \mathbf{x}) \leq 0.2; \end{cases}$$

To predict the label of a location l, it is critical to extract the associated features $\mathbf{x}$ which are expected to be relevant to the disease transmission and epidemic outbreak. In the following section, we elaborate the feature engineering for COVID-19 disease.

3.3 Features

We categorize the features into a couple of major categories, and extract the features from different groups of characteristics.

3.3.1 Geographical Features. The outbreak of epidemic disease is highly relevant to geolocation information. Big cities such as transportation hubs are generally facing with greater risk due to

[4]Baidu Map API: https://lbsyun.baidu.com/

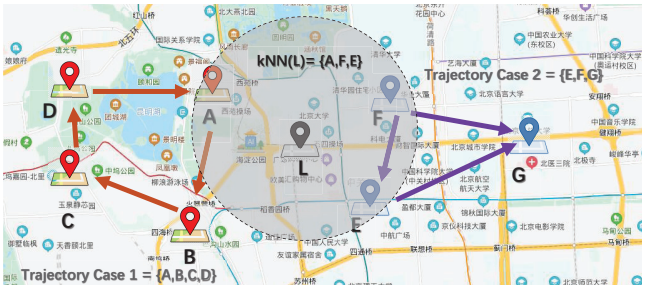

Figure 1: Illustration of nearest neighbors and individual motion trajectories. A confirmed case revealed the travel path. In this figure, we illustrate two trajectories on the map. Besides the nearest neighbors illustrated, we also show the activity range as the grey circle.

massive population mobility. It is believed that coronavirus quickly spreads from Wuhan—the COVID-19 starting spot of its outbreak in China—because Wuhan efficiently and conveniently connects to other parts of the whole country. Therefore, we first introduce two features related to the geography as *longitude* and *latitude*, where the calculation is defined as the corresponding coordinates:

$$x_1 = \text{longitude}(l) \tag{1}$$

$$x_2 = \text{latitude}(l) \tag{2}$$

Since the coronavirus is airborne through close contacts, it matters how close it is for an area to the location(s) of confirmed case(s). Here we introduce a distance function $dist(\cdot)$ to calculate the distance between two locations. Here, we introduce a hyperparameter of k to identify the k-nearest cases with locations (denoted as a cluster of L_{kNN}), which are calculated using the distance function. We formulate the following distance-based features:

$$x_3 = \min_{l_j \in L_{\text{kNN}}} \text{dist}(l, l_j) \tag{3}$$

$$x_4 = \text{avg}_{l_j \in L_{\text{kNN}}} \text{dist}(l, l_j) \tag{4}$$

which is to calculate the shortest distance to the k-nearest confirmed locations and the average distance to the k-nearest cases.

Usually, the confirmed locations do not simply indicate discrete location points due to human mobility. For some cases, we are able to identify the motion trajectory from the information release, and we believe the areas within the trajectory might be hazardous. We denote all the locations within the scope of a particular trajectory and denote these locations as a cluster $L_{\text{trajectory}}$. Then we have:

$$x_5 = \min_{l_j \in L_{\text{trajectory}}} \text{dist}(l, l_j) \tag{5}$$

$$x_6 = \text{avg}_{l_j \in L_{\text{trajectory}}} \text{dist}(l, l_j) \tag{6}$$

In this paper, we set an upper limit of the longest travel distance d within the trajectory area for an individual in Figure 1.

Finally, we assume that the location of a confirmed case is unlikely to be a single data point: people generally have a range of daily activities. We set up a hyperparameter of radius range r, and define a simple boolean function which indicates l is located within the range of confirmed cases. Within the range, people still have

the risk to be infected by the confirmed cases via occasional and unexpected contacts.

$$x_7 = \begin{cases} 0 & \text{if} \quad \min_{l_j \in L_{\text{kNN}}} dist(l, l_j) > r \\ 1 & \text{if} \quad \text{otherwise} \end{cases} \tag{7}$$

Note that x_5, x_6 and x_7 are different given the hyperparameters, which is determined by motion trajectory and daily activities.

3.3.2 Demographic Features. Epidemic outbreak is highly relevant to population distribution: it is intuitive that higher population density possibly results in higher danger of facing disease contagion. The demographic feature is associated with function facilities within the area, and we have the statistics of function facilities (i.e., *supermarkets*, *shopping malls*, *hospitals*, *metro stations*, and *apartment complex*) within the activity range r of the area to predict:

$$x_8 = \text{num}(l_{\text{market}}) \quad \text{where} \quad \text{dist}(l, l_{\text{market}}) \leq r \tag{8}$$

$$x_9 = \text{num}(l_{\text{mall}}) \quad \text{where} \quad \text{dist}(l, l_{\text{mall}}) \leq r \tag{9}$$

$$x_{10} = \text{num}(l_{\text{hospital}}) \quad \text{where} \quad \text{dist}(l, l_{\text{hospital}}) \leq r \tag{10}$$

$$x_{11} = \text{num}(l_{\text{apartment}}) \quad \text{where} \quad \text{dist}(l, l_{\text{apartment}}) \leq r \tag{11}$$

$$x_{12} = \text{num}(l_{\text{metro}}) \quad \text{where} \quad \text{dist}(l, l_{\text{metro}}) \leq r \tag{12}$$

The function $num(\cdot)$ is to count the number of typical function facilities within the area range.

The numbers of facilities cannot always accurately characterize demographic density. Hence we introduce two more index to show population of the area. We roughly estimate the population density based on the functions in orders of magnitude (0.1k people for a market, 0.5k for a mall and a hospital, 1k for a metro station, and 10k residents in the apartments):

$$\begin{aligned} x_{13} &= \frac{0.1x_8 + 0.5x_9 + 0.5x_{10} + 10x_{11} + 1x_{12}}{\text{Unit Land Area}} \\ &\propto 0.1x_8 + 0.5x_9 + 0.5x_{10} + 10x_{11} + 1x_{12} \end{aligned} \tag{13}$$

When calculating the density for each particular location, we use the same unit land area. Thus, we omit the denominator in Equation (13). Note that the naive number estimation is a highly simplified model based on city statistics in China. The population numbers actually vary from city to city. As a pilot study, we empirically choose the same setup for all cities in China.

Besides the population density of the neighborhood, we use a macro feature of total city population for the location l which we need to make a prediction:

$$x_{14} = population(\text{City}(l)) \tag{14}$$

The population of a city in China is publicly available.

3.3.3 Temporal Features. A unique character of an infectious disease to spread is that it takes time to incubate after exposure and infection, which is known as incubation time. Thus, we associate the relationship between temporal information and disease transmission. $t(\cdot)$ is to record the timestamp of the case with location.

$$x_{15} = |t(l) - t(l_i)| \text{ where } l_i = argmin_{l_i \in L_{\text{kNN}}} dist(l, l_i) \tag{15}$$

$$x_{16} = min_{l_i \in L_{\text{kNN}}} |t(l) - t(l_i)| \tag{16}$$

$$x_{17} = avg_{l_i \in L_{\text{kNN}}} |t(l) - t(l_i)| \tag{17}$$

$$x_{18} = min_{l_i \in L_{\text{trajectory}}} |t(l) - t(l_i)| \tag{18}$$

We keep the record of the time difference between the current time and the case time. We expect the larger the time gap is, the less influence is supposed to exist within the neighborhood.

3.3.4 Temperature. Clinical studies have revealed that temperature is highly relevant to the transmission of coronavirus of respiratory diseases such as SARS [2]. Since SARS transmission is sensitive to temperature and the novel coronavirus is known to be quite similar to SARS in the genome sequence. Given such an assumption, we introduce the temperature information for our prediction model:

$$x_{19} = min(temperature(City(l))) \tag{19}$$

$$x_{20} = max(temperature(City(l))) \tag{20}$$

$$x_{21} = avg(temperature(City(l))) \tag{21}$$

The features indicate the minimum, maximum and average temperature for the location l respectively.

3.3.5 Live Update Features. The daily disease situation reports are publicly accessible by official press. We believe the statistic and fact numbers are highly relevant to the warning system: intuitively, once there are still a large number of active cases, we will be facing with severe situation to massive areas in the battle against the pandemic. We include (almost) all important facts as features.

$$x_{22} = confirmed(City(l)) \tag{22}$$

$$x_{23} = confirmed(National) \tag{23}$$

$$x_{24} = suspected(City(l)) \tag{24}$$

$$x_{25} = suspected(National) \tag{25}$$

$$x_{26} = recovery(City(l)) \tag{26}$$

$$x_{27} = recovery(National) \tag{27}$$

$$x_{28} = death(City(l)) \tag{28}$$

$$x_{29} = death(National) \tag{29}$$

Since the cities within the whole country are not absolutely isolated from each other and people are traveling between cities, we introduce the facts from both the nation-wide level and the city-level results. Besides the cumulative numbers of the features, we also incorporate the new case numbers for each day.

$$x_{30} = new\_confirmed(City(l)) \tag{30}$$

$$x_{31} = new\_confirmed(National) \tag{31}$$

$$x_{32} = new\_suspected(City(l)) \tag{32}$$

$$x_{33} = new\_suspected(National) \tag{33}$$

$$x_{34} = new\_recovery(City(l)) \tag{34}$$

$$x_{35} = new\_recovery(National) \tag{35}$$

$$x_{36} = new\_death(City(l)) \tag{36}$$

$$x_{37} = new\_death(National) \tag{37}$$

3.3.6 Epidemiological Dynamics. Furthermore, we take a look at the transmission dynamics of the novel coronavirus COVID-19 to measure how it diffuses and disperses. First, we introduce the index of diffusion and dispersion.

$$x_{38} = \frac{x_{30}}{x_{22} - x_{26} - x_{28}} \tag{38}$$

$$x_{39} = \frac{x_{31}}{x_{23} - x_{27} - x_{29}} \tag{39}$$

$$x_{40} = \frac{x_{26} + x_{28}}{x_{22} - x_{26} - x_{28}} \tag{40}$$

$$x_{41} = \frac{x_{27} + x_{29}}{x_{23} - x_{27} - x_{29}} \tag{41}$$

The higher index of diffusion indicates higher risks for new cases to be confirmed from both the city-level (x_{38}) and the nation-level (x_{39}). The index of dispersion indicates how likely the disease is going to retreat by removal of inactive cases (i.e., 'recovery'+'death') based on the city-level (x_{40}) and the nation-level (x_{41}).

Growth Rate and *Doubling Time* are also important concepts to characterize the severity of a pandemic outbreak [25]. The doubling time is time it takes for the confirmed cases to double in size [13]. In general, we have the growth rate defined as follows:

$$x_{42} = \frac{\ln(x_{22}) - \ln(x_{22} - x_{30})}{\Delta t = 1} \tag{42}$$

$$x_{43} = \frac{\ln(x_{23}) - \ln(x_{23} - x_{31})}{\Delta t = 1} \tag{43}$$

We set the time unit as 1-day, and suppose that the growth rate within a day is a constant. The growth rate characterize the ratio of confirmed cases between the current day and the day before in the natural logarithm. x_{42} and x_{43} correspond to the city-level and nation-level features.

Given the growth rate, it is straightforward to obtain the *doubling time* for the local cases x_{44} and national cases x_{45} by:

$$x_{44} = \frac{\ln(2)}{\ln(1 + x_{42})} \tag{44}$$

$$x_{45} = \frac{\ln(2)}{\ln(1 + x_{43})} \tag{45}$$

There are approaches such as the SEIR model to simulate the epidemic outbreak process. However, we cannot obtain the accurate number of exposed cases, we degenerate the SEIR model to an *susceptible-infectious-recovered* (SIR) model with exact analytical solutions [8]. There are three groups of people: those that are healthy but susceptible to the disease (S), the infected (I) and the people who have recovered (R). To model the dynamics of the outbreak we need three differential equations, one for the change in each group, where β is the parameter that controls the transition between S and I and γ which controls the transition between I and R. A healthy individual can be infected and then can be recovered:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$
$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

where $N = S+I+R$. Since we have the numbers for the S, I, and R, we can use the known data to infer the parameters of γ and β. The infection probability of β is very critical to characterize; the recovery rate of γ depends on the medical level, public health measurements, and government policies. More importantly, we calculate the *basic reproductive number* (R_0) which denotes the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection using β and γ.

$$x_{46} = \beta(City(l)) \tag{46}$$

$$x_{47} = \beta(National) \tag{47}$$

$$x_{48} = \gamma(City(l)) \tag{48}$$

$$x_{49} = \gamma(National) \tag{49}$$

As to the basic reproductive numbers:

$$x_{50} = R_0(City(l)) = \frac{\beta(City(l))}{\gamma(City(l))} = \frac{x_{46}}{x_{48}} \tag{50}$$

$$x_{51} = R_0(National) = \frac{\beta(National)}{\gamma(National)} = \frac{x_{47}}{x_{49}} \tag{51}$$

Note that x_{46}-x_{51} are not constants: they are subject to change as the situations have changed on the city-level and nation-level.

3.4 Learning Model

Given the extracted features $\mathbf{x}$=$\{x_1, x_2, \ldots, x_{51}\}$, we predict the label using $f(y, \mathbf{x})$. The function $f(\cdot)$ can be standard machine learning models such as Support Vector Machine (SVM) [4], Decision Tree (DT) [20], Naive Bayes (NB) [17], Random Forests (RF) [9], Multi-Layer Perceptron (MLP) [1], and Gradient Boosting Decision Tree (GBDT) [18]. These standard classification models are all compatible with our proposed framework.

Note that in our proposed features, there are a set of hyperparameters, i.e., the number of nearest neighbors k, the range of daily activity r, and the upper bound of distance for a trajectory d. To tune these hyperparameters, we introduce a reinforcement learning framework to adapt with the updating data streams.

The reinforcement learning framework consists of a generator $g(\cdot)$ to generate the hyperparameters and a policy gradient method to optimize the hyperparameter generator. In this paper, we instantiate the generator as a vanilla recurrent neural network (RNN) [1, 29, 30]. We use h to denote the hyperparameter set and each hyperparameter h_i is regarded as a token. We have an assumption here that the hyperparameters are not completely independent on each other. The RNN generator is able to generate the hyperparameter tokens one by one. We maintain a set of predefined setups for these hyperparameters. The generator transforms them into token representations and selects the token through a softmax layer, which is similar to word generation by the look-up table through text vocabularies [6]. At each timestep, the generator $g(\cdot)$ generates a hyperparameter.

One limitation is that we do not have any supervised signal to train the hyperparameter generator. A natural solution is that we utilize the reinforcement learning based on policy gradient to give feedback to the hyperparameter generator, which is inspired from Neural Architecture Search (NAS) [32]. In practice, we train the

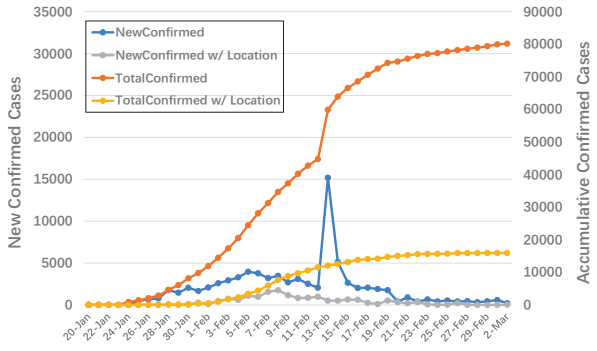

Figure 2: Data statistics of China from Jan 20 to Mar 2. We keep the data samples with location information for prediction. The vertical axis on the left side indicates new cases while the vertical axis on the right side indicates accumulative cases (both axis indicating raw data and pre-processed data with location information).

prediction model, calculate the results and then denote the reward as q to optimize the hyperparameters of the prediction model during the reinforcement process. The training objective is to minimize the negative expected reward following the REINFORCE algorithm [26]. θ denote the parameters of the generator.

$$J(\theta) = \mathbb{E}[q|(h,\theta)] \tag{52}$$

The gradient is estimated as:

$$\nabla J(\theta) \approx (q-b)\nabla\pi_\theta(h,\theta) \tag{53}$$

$\pi_\theta(h,\theta)$ represents the probability of generating the current hyperparameters h. b denotes the baseline value to reduce the variance of the gradient estimate while keeping it unbiased [21]. To be more specific, b is an exponential moving average of the previous rewards, and is updated with a coefficient empirically set as 0.8:

$$b \leftarrow 0.8 \times b + (1-0.8) \times q \tag{54}$$

We set the model learning in an online mode, which is to be updated every previous day (when $t_i < t$) as new data are incorporated. The function yields the following gradient to update the generator:

$$\nabla J(\theta) \approx \sum_{i=1}^{t-1}(q_i - b)\nabla\pi_\theta(h,\theta) \tag{55}$$

The parameters are updated with a learning rate η ($\eta = 0.01$):

$$\theta \leftarrow \theta + \eta\nabla J(\theta) \tag{56}$$

4 EXPERIMENTS AND EVALUATIONS

In this section, we will introduce the experimental results based on the epidemic outbreak of COVID-19 coronavirus in China. In the meanwhile, some case studies are demonstrated for revealing more insights from our proposed model, features and data.

4.1 Dataset

All the information released from the government officials can be publicly accessible through government webpages, including the statistics of all administrative cities and provinces in China. We release the data via our project page [5] for everyone wishing to contribute efforts for the fight against COVID-19 novel coronavirus.

The dataset consists of 81,604 reported cases in total, among which 16,166 cases are associated with location information. Our work aims at predicting the hazard areas under the pandemic scenario. Thus, incomplete data points without location information will be filtered and removed as pre-processing. Unfortunately, most of the reported cases in the city of *Wuhan* are incomplete without location information and have to be removed. The timestamps of the reported cases are associated with the publish dates.

It is important to note that reported cases are not supposed to have permanent impacts on model learning. Generally, it is believed that an area without new cases for a period of incubation (i.e., 7-14 days) would be safe from local community COVID-19 spreading: the area is clear. To this end, we remove data points reported 14 days ago when learning the predictive model for a particular date. We summarize the statistics of the dataset in Table 1 and Figure 2.

Table 1: Statistics about sample size and details about the COVID-19 epidemic outbreak data. The statistics indicate samples after pre-processing and incomplete data removal. The data are collected till March, 2020.

Statistics	Values
total # of reported cases (raw data)	81,604
total # of reported cases (w/ location)	16,166
total # of reported cities	297
avg # of reported case per city	54
max # of reported case within a city	1,104
min # of reported case within a city	1
max peak # of new cases within a city	414
min peak # of new cases within a city	1
avg peak # of new cases within a city	23
earliest starting date (first new case)	Jan 21 (Meizhou, etc.)
earliest clearance date (no more new case)	Jan 28 (Yingkou)
ending date	Mar 2 (Shanghai, etc.)
longest duration (in terms of days)	28 (Anyang)
shortest duration (in terms of days)	1 (Chengde, etc.)
avg duration (in terms of days)	12
total duration (in terms of days)	42

Table 2: Experimental setups for model training/learning.

Phase	Statistics	Training	Validation	Testing
I	Duration	Jan 20-Feb 5	Feb 6-Feb 7	Feb 8-Mar 2
	# of Cases	8,926	5,871	24,542
	Pos(+)/Neg(-)	1,756/7,170	1,961/3,910	6,132/18,410
II	Duration	Jan 20-Feb 12	Feb 13-Feb 14	Feb 15-Mar 2
	# of Cases	32,743	6,596	21,658
	Pos(+)/Neg(-)	8,853/23,890	996/5,600	2,668/18,990
III	Duration	Jan 20-Feb 19	Feb 20-Feb 21	Feb 22-Mar 2
	# of Cases	55,235	5,762	15,892
	Pos(+)/Neg(-)	11,975/43,260	542/5,220	672/15,220

[5] https://github.com/fuzhenxin/COVID19Warnings

Table 3: Model performance for Phase I, II, and III w.r.t. Level-5 and Level-4 predictions.

	Predicting Level-5				Predicting Level-4 and Above				Overall
Phase I	p	r	F1	acc.	p	r	F1	acc.	AUC
Baseline1	0.4478	0.6021	0.5136	0.7151	0.3565	0.7118	0.4751	0.6070	0.6884
Baseline2	0.4473	0.6029	0.5135	0.7146	0.3565	0.7118	0.4751	0.6069	0.6883
NB	0.6365	0.7265	0.6785	0.8280	0.5911	0.7634	0.6663	0.8089	0.8922
NB+RL	0.6352	**0.7319**	**0.6802**	**0.8280**	0.5861	**0.7811**	**0.6697**	0.8075	**0.8935**
GBDT	**0.8624**	0.1670	0.2798	0.7854	**0.8351**	0.3579	0.5011	**0.8218**	0.8615
GBDT+RL	0.8579	0.1249	0.2180	0.7762	0.8109	0.3017	0.4398	0.8078	0.8622
MLP	0.8188	0.3034	0.4428	0.8087	0.7711	0.3973	0.5244	0.8190	0.8618
MLP+RL	0.8235	0.2722	0.4091	0.8029	0.7809	0.3626	0.4952	0.8145	0.8624
	Predicting Level-5				Predicting Level-4 and Above				Overall
Phase II	p	r	F1	acc.	p	r	F1	acc.	AUC
Baseline1	0.2974	0.7440	0.4250	0.7520	0.2139	**0.8253**	0.3397	0.6048	0.7676
Baseline2	0.2967	0.7448	0.4243	0.7510	0.2139	0.8253	0.3397	0.6048	0.7672
NB	0.4749	0.7496	0.5815	0.8671	0.4349	0.8025	0.5641	0.8472	0.9105
NB+RL	0.4797	**0.7526**	**0.5859**	0.8690	0.4194	0.7965	0.5495	0.8391	0.9111
GBDT	**0.9557**	0.3480	0.5102	**0.9177**	0.8712	0.5449	0.6704	0.9340	0.9404
GBDT+RL	0.9493	0.3238	0.4829	0.9146	**0.8871**	0.5394	**0.6708**	**0.9347**	**0.9423**
MLP	0.7714	0.3123	0.4446	0.9037	0.6402	0.3927	0.4868	0.8974	0.7828
MLP+RL	0.8210	0.3071	0.4470	0.9063	0.7163	0.3799	0.4965	0.9046	0.8106
	Predicting Level-5				Predicting Level-4 and Above				Overall
Phase III	p	r	F1	acc.	p	r	F1	acc.	AUC
Baseline1	0.0776	0.5327	0.1355	0.7126	0.0685	0.7530	0.1255	0.5563	0.6748
Baseline2	0.0780	0.5372	0.1361	0.7117	0.0685	**0.7545**	0.1257	0.5561	0.6757
NB	0.1943	0.5863	0.2919	0.8797	0.1626	0.6890	0.2631	0.8368	0.8652
NB+RL	0.1971	**0.5967**	0.2963	0.8801	0.1634	0.6652	0.2623	0.8418	0.8639
GBDT	0.6645	0.2560	0.3696	0.9621	**0.4716**	0.4246	**0.4469**	**0.9551**	**0.9194**
GBDT+RL	**0.6674**	0.2688	**0.3833**	**0.9623**	0.4498	0.4315	0.4405	0.9532	0.9185
MLP	0.1965	0.3021	0.2381	0.9180	0.1616	0.4335	0.2354	0.8805	0.7662
MLP+RL	0.2380	0.3224	0.2738	0.9278	0.1794	0.4638	0.2587	0.8867	0.7488

4.2 Evaluation Metrics

We include the classic evaluation metrics for the classification task using *accuracy*, *precision*, *recall* and *F1* scores [17]. Since we have different levels of prediction, we list the results for different levels as well. In particular, we pay extra attention to the areas with higher alerts (Level-4 and above). We adopt the same evaluation standard for all methods in our experiments.

In the real-world process of decision making, people usually decide a latent threshold to identify risky areas. Empirically, the threshold is highly relevant to personal choices. Thus, we comprehensively validate the performance using the AUC index to measure the performance under different levels [17]. AUC can reflect model performance within different boundary values between classes and thus is widely used in two-class classification problems.

4.3 Experimental Setups

In general, we have a dataset with a duration of 42 days nationwide. We investigate the prediction capability of our proposed model for 3 distinctive phases in China: **Phase I**) beginning stage (i.e., late January or early February), **Phase II**) outbreaking stage (i.e., mid-February), and **Phase III**) ending stage (i.e., late February or early March). We roughly divide the phases according to the growth rate. To be more specific, the details of training/validation/testing split for each phase are illustrated in Table 2.

Considering the incubation time, we predict all the future cases within 7-14 days to make the system timely effective. As mentioned in Section 3.4, various standard classification models are compatible with our framework. we use Naive Bayes (NB), GBDT and MLP for model learning. We use the model parameter setups by default. As to the hyperparameters, we set r and d from a range of 1 km to 20 km while select k from 5 to 20, all trained by reinforcement learning. In our experiments, we set the AUC scores as the reward of reinforcement learning.

It is critical to choose eligible negative samples to pair up with positive samples for model learning. We have investigated random negative sampling, which would make the data points too sparse to make accurate predictions. To this end, we use the strategy of

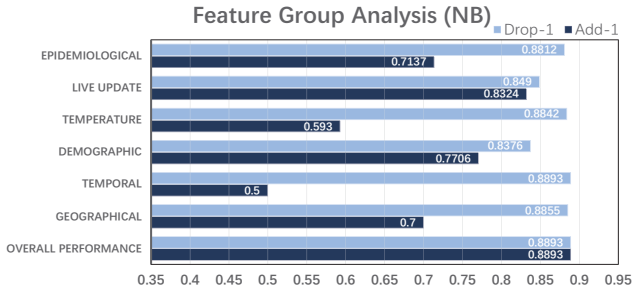

Figure 3: Feature analysis (NB). The performance is the average result of Phase I, II, and III, measured by AUC.

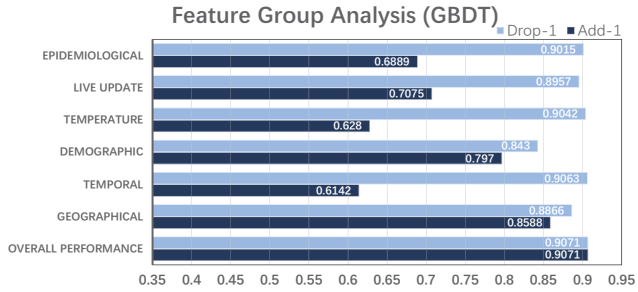

Figure 4: Feature analysis (GBDT). The performance is the average result of Phase I, II, and III, measured by AUC.

restricted negative sampling, which is to sample negative data points within the same cities as the positive samples.

4.4 Baseline Approaches

Our proposed model is learned with a series of features extracted from the dataset based on the reinforced machine learning framework. To demonstrate the effectiveness of our proposed method with empirical features, we establish two heuristic baselines based on the most straightforward features: distance to the near-by confirmed cases and/or trajectories.

• *Baseline 1.* We measure the distance of the candidate area to the nearest confirmed case, and predict the levels of hazard accordingly.

$$\text{Label}(l) = \begin{cases} \text{Level-5} & \text{if} \quad 0 \le \min_{l_j \in L_{\text{kNN}}} dist(l, l_j) < r; \\ \text{Level-4} & \text{if} \quad r \le \min_{l_j \in L_{\text{kNN}}} dist(l, l_j) < 2r; \\ \text{Level-3} & \text{if} \quad 2r \le \min_{l_j \in L_{\text{kNN}}} dist(l, l_j) < 3r; \\ \text{Level-2} & \text{if} \quad 3r \le \min_{l_j \in L_{\text{kNN}}} dist(l, l_j) < 4r; \\ \text{Level-1} & \text{if} \quad 4r \le \min_{l_j \in L_{\text{kNN}}} dist(l, l_j); \end{cases}$$

• *Baseline 2.* We measure the distance of the candidate area to the nearest trajectory of a confirmed case, and similarly, predict the levels of hazard in the same way as Baseline 1.

$$\text{Label}(l) = \begin{cases} \text{Level-5} & \text{if} \quad 0 \le \min_{l_j \in L_{\text{trajectory}}} dist(l, l_j) < r; \\ \text{Level-4} & \text{if} \quad r \le \min_{l_j \in L_{\text{trajectory}}} dist(l, l_j) < 2r; \\ \text{Level-3} & \text{if} \quad 2r \le \min_{l_j \in L_{\text{trajectory}}} dist(l, l_j) < 3r; \\ \text{Level-2} & \text{if} \quad 3r \le \min_{l_j \in L_{\text{trajectory}}} dist(l, l_j) < 4r; \\ \text{Level-1} & \text{if} \quad 4r \le \min_{l_j \in L_{\text{trajectory}}} dist(l, l_j); \end{cases}$$

Here, we use the best tuned range $r \in$ [1km, 20km] for the baselines, which is the same empirical setup for our methods in the reinforcement learning framework.

4.5 Results

4.5.1 Overall Performance. We compare the performance of all methods including baselines and our proposed prediction model with various features, measured in terms of all evaluation metrics. In Table 3 we list the overall results for these methods.

Surprisingly, the heuristics-inspired baselines are rather competitive. Yet, it is not difficult to understand that distance-oriented information is critical for the hazard area prediction task (considering that COVID-19 coronavirus is transmitted within a short distance). Such an assumption may explain the performance of the two baselines which are designed by distance measurements.

The learning based models (NB and GBDT) generally perform better than the heuristic baselines. Distance features are important, while we ascribe the improvement of the results from other various features. We will further ablate the feature analysis in the following sections. The performance of MLP is not as expected. As is well-understood that neural networks are data hungry, the insufficient data samples perhaps lead to the unsatisfying results of MLP.

We take a closer look at the details of the scores achieved by different methods. All models, including two baselines, have better performance in *recall* scores, except the GBDT model. Generally, GBDT tends to have a relatively higher *precision* score. We conclude that different models have different preference towards precision and recall. For the city residents, the *recall* metric matters more: it is important to identify as more hazard areas as possible. People should be warned to avoid those areas.

There are two warning strategies: 1) Level-5 and 2) Level-4 and above. There is trade-off between the two strategies: the Level-5 strategy is more conservative with much higher *precision* but lower *recall*. In contrast, the latter strategy is more aggressive, trying to retrieve more possible hazard areas while the precision score naturally drops. The AUC score indicates the overall performance of predictions on all levels.

As to the performance of different phases, we can see the models predict better during the starting phase and the outbreaking phase (i.e., Phase I and Phase II). The prediction performance for Phase III demonstrates a clear drop in terms of almost all metrics for all methods. We ascribe this phenomenon to the efforts of reaction forces: the government is taking actions to prevent disease transmission while people are wearing face masks and other protections. Therefore, there are much fewer infection cases than the model originally expected!

It is interesting to see that the reinforcement learning component generally contributes to the overall performance of the learning based models. Yet there are two concerns of the reinforcement component: 1) the improvement is not consistently stable for all models, and 2) the improvement is to some extent marginal. As to the marginal effect, since we use reinforcement learning to train only hyperparameters (k, r, and d), which may be a limitation to the reinforcement component. The benefit of utilizing reinforced learning is to tune hyperparameters incrementally as the epidemic

evolves, which leads to more accurate model prediction. Thus, when the conditions are intervened (such as government efforts), the effects may be subject to change as well, which leads to less stable performance in our experiments (e.g., Phase III).

4.5.2 Feature Analysis. We further analyze the contribution of all features. We conduct an ablation study on the features and visualize the result in Figures 3-4. With 51 features in total, we group them into 6 different feature groups: 1) *geographical features*, 2) *demographic features*, 3) *temperature*, 4) *temporal*, 5) *live updates*, and 6) *epidemiological dynamics*. We also list the overall performance of the full model which employs all factors for comparison. Here we examine the contributions of the different feature groups defined in our method. To be more specific, in the ablation experiments, we show the performance of all the factors in isolation (namely '*Add-1*') and then leave-one-out (namely '*Drop-1*') from the full combination of all features, one feature group at a time.

From Figures 3 and 4, we see that all of the feature groups are generally positive in our evaluation tasks, although have different performance in different prediction models. The NB model prefers to use demographic features and live update facts for prediction while the model decides that temperature features and temporal features have little effect for classification. The GBDT model concurs with the results of the NB model: temperature and temporal features contribute to the overall performance but not too much. For both the NB and the GBDT models, demographic features are quite important for the prediction performance. Geographical features, live update facts, and epidemiological features are also demonstrated to be the top factors in the Add-1 tests. Such features are expected to be useful for prediction because they are closely associated with the disease transmission: dense population to infect, short distance of transmission and the current overall situation of infectious group.

We have examined the variance of temperature during the epidemic outbreak. The temperature does not change too much during the entire February for most of the cities, which could be the reason why temperature features fail to be effectively discriminative. As to the temporal features, since we have pre-processed the data samples by removing cases reported more than 14 days ago, the effectiveness of temporal features may be compromised.

4.5.3 A More Proactive Strategy. For the common people, the best way to use the warning system is to avoid unnecessary contact with the predicted hazard areas with high alert levels. However, we do have a potential choice to be more proactive in the combat against COVID-19. For instance, the administrative forces can closely monitor areas of high alerts, and sanitize such areas regularly to prevent possible disease transmission with a good chance.

Different models have different capability of prediction: some models tend to warn more locations with higher recall while others may warn fewer locations for better precision. Without doubt, it is impractical to take proactive actions for all predicted locations. Therefore, we investigate the top prediction results. We *rank* the predicted areas by the score of the classification models. In particular, we test the results for p@500 and r@500.

From Table 4, we observe that the top predictions have extremely high precision (0.9+) in both Phase I and Phase II! The phenomenon indicates that if any efforts can be used to successfully take care of these places, we might have a great chance to stop the local transmission of COVID-19 coronavirus as soon as the confirmed cases are spotted. During Phase III, the precision performance decreases. We have analyzed the reason for the performance drop in Phase III: government actions may have effective results and the expected cases are therefore intervened.

Table 4: Top-500 predictions of all methods. We examine the result of predicted hazard areas to see if they should be taken good care of in priority.

	Phase I		Phase II		Phase III	
Method	p@500	r@500	p@500	r@500	p@500	r@500
Baseline1	0.1140	0.0093	0.0320	0.0060	0.0100	0.0074
Baseline2	0.1140	0.0093	0.0320	0.0060	0.0100	0.0074
NB	0.9200	0.0750	0.9700	0.1818	0.3720	0.2768
NB+RL	**0.9220**	**0.0752**	0.9760	0.1829	0.3840	0.2857
GBDT	0.8593	0.0701	0.9847	0.1845	0.5087	0.3785
GBDT+RL	0.8833	0.0720	**0.9860**	**0.1848**	**0.5100**	**0.3795**
MLP	0.8993	0.0733	0.9673	0.1813	0.2233	0.1662
MLP+RL	0.9020	0.0735	0.9660	0.1810	0.2760	0.2054

Table 5: Case studies on different cities: Beijing, Shanghai and Chongqing are out of Hubei Province while Suizhou, Huanggang and Jingmen are cities in Hubei Province.

	Best Baseline			Best Our Method		
City	p	r	F1	p	r	F1
Beijing	0.0801	**0.9285**	0.1467	**0.4909**	0.8334	**0.6178**
Shanghai	0.0935	0.6666	0.1623	**0.2358**	**0.8586**	**0.3699**
Chongqing	0.6511	0.7257	0.5936	**0.7131**	**0.7833**	**0.7286**
Suizhou	0.2977	0.688	0.4156	**0.5697**	**0.748**	**0.6468**
Huanggang	0.9147	**0.8655**	**0.8894**	**0.9294**	0.6365	0.7556
Jingmen	0.4415	0.4716	0.4561	**0.7919**	**0.6618**	**0.721**

4.5.4 Case Studies. We investigate how the model performs in different cities with different situations. Hubei Province is the center of COVID-19 outbreak in China with Wuhan as the capital. Since most data in Wuhan are absent without locations, we select several other representative cities inside and outside Hubei. We have a general observation that the model will be better at predicting local transmission cases than imported cases from other places.

One group of cities—*Beijing*, *Shanghai* and *Chongqing*—are located outside Hubei. In cities outside Hubei, cases are mixed with hybrid sources: both local transmission and imported cases. The early cases from Hubei gradually cause massive local transmission. The situation is more severe in *Suizhou* and *Jingmen*, both of which are cities located in Hubei Province. Another city in Hubei, *Huanggang*, shows very high alert predictions and we note that the baselines perform pretty well. We look into the data samples in Huanggang and identify highly centralized community-level outbreak, which explains why simple heuristics of distance-based methods have excellent performance.

5 CONCLUSIONS AND FUTURE WORK

We build a warning system to predict hazard areas in order to intervene the novel coronavirus COVID-19 epidemic transmission.

We crawl the data from public information release, extract relevant features based on empirical studies and conduct model learning. We also incorporate a reinforcement learning module to facilitate hyperparameter learning. The experiments have demonstrated that the system is able to predict hazard areas of future cases, and have better performance than heuristic baselines by various metrics.

We conduct additional experiments on ablation studies, which indicate the feature groups have positive impacts on model performance. In general, the demographic features and geographical features are demonstrated to have stronger contribution because they are closely related to disease transmission. Temporal and temperature features are the least effective. Through the case studies from different cities, we observe that our proposed method is better at predicting local spread on the community-level. For the cities with only imported cases, the prediction model is less effective.

Now the epidemic outbreak is still raging and people are trying all efforts to intervene coronavirus pandemic. As our future work, we will scale up the data size and adapt our model for more countries. It is the human fight against the virus, and hopefully we will win the war against the COVID-19 in the near future.

ACKNOWLEDGMENTS

We feel so grateful for the enormous efforts that people have devoted into the fight against COVID-19 and to save the lives.

We thank the anonymous reviewers for their constructive comments and suggestions. This work was partially supported by the National Natural Science Foundation of China (NSFC Grant No. 61876196). Rui Yan was sponsored as a Young Fellow of Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] Yoshua Bengio. 2009. Learning deep architectures for AI. *Foundations and trends® in Machine Learning* 2, 1 (2009), 1–127.

[2] KH Chan, JS Peiris, SY Lam, LLM Poon, KY Yuen, and WH Seto. 2011. The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Advances in virology* 2011 (2011).

[3] Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).

[4] Corinna Cortes and Vladimir Vapnik. 1995. Support vector machine. *Machine learning* 20, 3 (1995), 273–297.

[5] Z Du, L Wang, S Cauchemez, X Xu, X Wang, BJ Cowling, and LA Meyers. 2020. Risk for transportation of 2019 novel coronavirus disease from Wuhan to other cities in China. *Emerging infectious diseases* 26, 5 (2020).

[6] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*. 663–670.

[7] Yan-Rong Guo, Qing-Dong Cao, Zhong-Si Hong, Yuan-Yang Tan, Shou-Deng Chen, Hong-Jun Jin, Kai-Sen Tan, De-Yun Wang, and Yan Yan. 2020. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. *Military Medical Research* 7, 1 (2020), 1–10.

[8] Tiberiu Harko, Francisco SN Lobo, and MK Mak. 2014. Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Appl. Math. Comput.* 236 (2014), 184–194.

[9] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

[10] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 395, 10223 (2020), 497–506.

[11] David S Hui, Esam I Azhar, Tariq A Madani, Francine Ntoumi, Richard Kock, Osman Dar, Giuseppe Ippolito, Timothy D Mchugh, Ziad A Memish, Christian Drosten, et al. 2020. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases* 91 (2020), 264–266.

[12] Sung-mok Jung, Andrei R Akhmetzhanov, Katsuma Hayashi, Natalie M Linton, Yichi Yang, Baoyin Yuan, Tetsuro Kobayashi, Ryo Kinoshita, and Hiroshi Nishiura. 2020. Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. *Journal of clinical medicine* 9, 2 (2020), 523.

[13] William Ogilvy Kermack and Anderson G McKendrick. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115, 772 (1927), 700–721.

[14] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* (2020).

[15] Cheng-wei Lu, Xiu-fen Liu, and Zhi-fang Jia. 2020. 2019-nCoV transmission through the ocular surface must not be ignored. *The Lancet* 395, 10224 (2020), e39.

[16] Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* 395, 10224 (2020), 565–574.

[17] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

[18] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. 2000. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*. 512–518.

[19] Liangrong Peng, Wuyue Yang, Dongyan Zhang, Changjing Zhuge, and Liu Hong. 2020. Epidemic analysis of COVID-19 in China by dynamical modeling. *arXiv preprint arXiv:2002.06563* (2020).

[20] J Ross Quinlan. 1986. Induction of decision trees. machine learning 1. *1 (1)* (1986), 81–106.

[21] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732* (2015).

[22] Jonathan M Read, Jessica RE Bridgen, Derek AT Cummings, Antonia Ho, and Chris P Jewell. 2020. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *MedRxiv* (2020).

[23] Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, and Riaz Agha. 2020. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *International Journal of Surgery* (2020).

[24] Huaiyu Tian, Yonghong Liu, Yidan Li, Moritz UG Kraemer, Bin Chen, Chieh-Hsi Wu, Jun Cai, Bingying Li, Bo Xu, Qiqi Yang, et al. 2020. Early evaluation of transmission control measures in response to the 2019 novel coronavirus outbreak in China. *medRxiv* (2020).

[25] Byung Mook Weon. 2020. Doubling time tells how effective Covid-19 prevention works. *medRxiv* (2020).

[26] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[27] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579, 7798 (2020), 265–269.

[28] Joseph T Wu, Kathy Leung, and Gabriel M Leung. 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* 395, 10225 (2020), 689–697.

[29] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 55–64.

[30] Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 685–694.

[31] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* (2020).

[32] Barret Zoph and Quoc V Le. 2016. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578* (2016).

[33] Xin Zou, Ke Chen, Jiawei Zou, Peiyi Han, Jie Hao, and Zeguang Han. 2020. Single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to 2019-nCoV infection. *Frontiers of Medicine* (2020), 1–8.