Artificial Intelligence Evolution
http://ojs.wiserpub.com/index.php/AIE/

UNIVERSAL WISER PUBLISHER

Research Article

Data Analytics for COVID-19 Pandemic

Zhen-Zhen Chen[1], Rong-Jie Li[1], Xin-Yi He[1], Zhen-Xin Lian[1], Zne-Jung Lee[2*]

[1]School of Big Data, Fuzhou University of International Studies and Trade, China
[2]School of Intelligent Construction, Fuzhou University of International Studies and Trade, China
Email: lrz@fzfu.edu.cn

Received: 15 November 2021; **Revised:** 5 January 2022; **Accepted:** 6 January 2022

Abstract: Since the outbreak of the coronavirus disease 2019 (COVID-19) pandemic, the pandemic situation has begun to undergo positive changes with the joint efforts of various countries and world organizations. However, pressures such as the COVID-19 mutations and the sharp rise in confirmed cases have brought uncertainties to the prevention and control of the pandemic. The overall situation is still severe and complex. Based on the multi-dimensional spatial-temporal COVID-19 data collected by the open-source NetEase News (NEN) website and a real-time dynamic website, it is to explore the characteristics of the pandemic data, visualize the development trend, and analyze the spread of the pandemic in this paper. Moreover, it is to provide a rule basis for the prevention and control of the COVID-19 pandemic by constructing the decision tree model. From the results, some suggestions are provided for decision-makers.

***Keywords*:** COVID-19, pandemic, data analytics, decision rules, decision tree

1. Introduction

At the end of December 2019, Wuhan City, Hubei Province in China, was the first to report new coronavirus pneumonia called Corona Virus Disease 2019 (COVID-19). Thereafter, Wuhan city was locked down on January 23, 2020. The nationwide COVID-19 prevention and control measures have been rapidly upgraded, and the Chinese government has organized more than 40,000 medical staff across the country to support Wuhan city, Hubei. Thanks to the joint efforts of the people across the country, the pandemic has finally been effectively controlled. Since late February 2020, the foreign pandemic situation has developed rapidly. The cumulative number of confirmed new cases worldwide has exceeded 200 million after August 5, 2021, and more than 4.25 million people have died [1-3]. The global pandemic prevention and control situation is unprecedentedly severe.

This paper comprehensively studies the multi-dimensional space-time characteristics of COVID-19 pandemic data to analyze the future trends hidden in the data and then data visualization [4-5]. It first uses Python powerful module to actively capture NEN-related data through web crawlers, complete the acquisition of new coronavirus-related pandemic data, use Tableau for data analysis and visualization, and intuitively express the characteristics of the COVID-19 pandemic [6]. Moreover, the construction of the decision tree model focuses on mining the characteristics of the COVID-19 pandemic, the developing situation, and analyzing its trend.

This paper is organized as follows. Section 2 describes the data collection and preprocessing. Data visualization is presented in Section 3. The introduction of the decision tree, data analytics, and simulation results are shown in Section 4.

Copyright ©2022 Zne-Jung Lee, et al.
DOI: https://doi.org/10.37256/aie.3120221247
This is an open-access article distributed under a CC BY license
(Creative Commons Attribution 4.0 International License)
https://creativecommons.org/licenses/by/4.0/

Finally, Section 5 is the conclusion.

2. Data collection and preprocess

The dataset in this paper was collected using python language to process a crawler based on the dynamics of the COVID-19 pandemic in NEN and a real-time dynamic website [7-8]. In python language, a third-party module is used to process the content of the webpage, and then the format of JSON data is analyzed. At the same time, the Timer class is used to periodically update the captured webpage data. The Python crawler crawls the dynamic data of the COVID-19 pandemic updated on its web page. The paper selects the cumulative data of the worldwide pandemic and that of data in China.

Table 1. The field name and annotation of the used data in China

Field Name	Annotation
date	Date
province_code	Chinese province code
province_name	Province (city/autonomous region/special administrative region) name
today_newly	Cases-newly reported in last 24 hours
today_healed	Cases-newly healed in last 24 hours
today_deaths	Cases-newly deaths in last 24 hours
month_newly	Cases-newly reported in one month
total_confirm	Cases-a cumulative total
total_healed	Healed-a cumulative total
total_deaths	Deaths-a cumulative total
total_newly	Newly Reported Cases-a cumulative total

Table 2. The field name and annotation of the used data in the worldwide

Field Name	Annotation
date	Date
month_newly	Cases-newly reported in one month
total_deaths	Deaths-a cumulative total
total_healed	Healed-a cumulative total
total_confirm	Cases-a cumulative total

In the data preprocessing, the missing value is deleted and a series of pandemic-related data such as confirm, heal, dead, and severe is selected. To better observe the distribution of the pandemic status, the article uses the province

code of each province to classify the pandemic status of each region (province, city, autonomous region, and special administrative region) and province name in China. For the crawled pandemic data, this article selects the data of today_newly, today_healed, today_deaths, total_confirm, total_healed, total_deaths, total_newly, and month_newly. The field name and annotation of the used data in China are listed in Table 1. In order to study the overall dynamics of the COVID-19 pandemic, this article also crawled the worldwide pandemic data. After deleting the missing value, it is to analyze and predict the worldwide pandemic data. Using a python crawler to crawl the data, it facilitates data collection to a large extent. These collected data are date, month_newly, total_deaths, total_healed, and total_confirm. The field name and annotation of the used data worldwide are shown in Table 2.

3. Data visualization

The visual analysis of major event data can achieve the effects of intuitive expression and the exchange of ideas. Since the outbreak of the COVID-19 pandemic, a lot of work has been done on the visualization of COVID-19 pandemic data. It is to enhance persuasiveness, which is more helpful to discover hidden events and trends [9]. Based on the collected pandemic data, with the help of visualization tools, the data of worldwide and China are presented in a graphical form. From the beginning of the COVID-19 pandemic on January 1, 2020, to October 3, 2021, it is to explore the characteristics, development, and evolution of pandemic data. Since the outbreak of the pandemic, the number of cumulative global cases has continued to grow rapidly. The number of new cases per day has increased in stages, and the overall situation is not optimistic as shown in Figure 1. As of August 29, 2021, the global healed rate of COVID-19 cases is about 76.39%, and the deaths rate is about 1.85%. The current healed rate is maintained at a good level, and the deaths rate is low, but there are still about 21.76% of confirmed cases, and the burden of the pandemic is heavy.

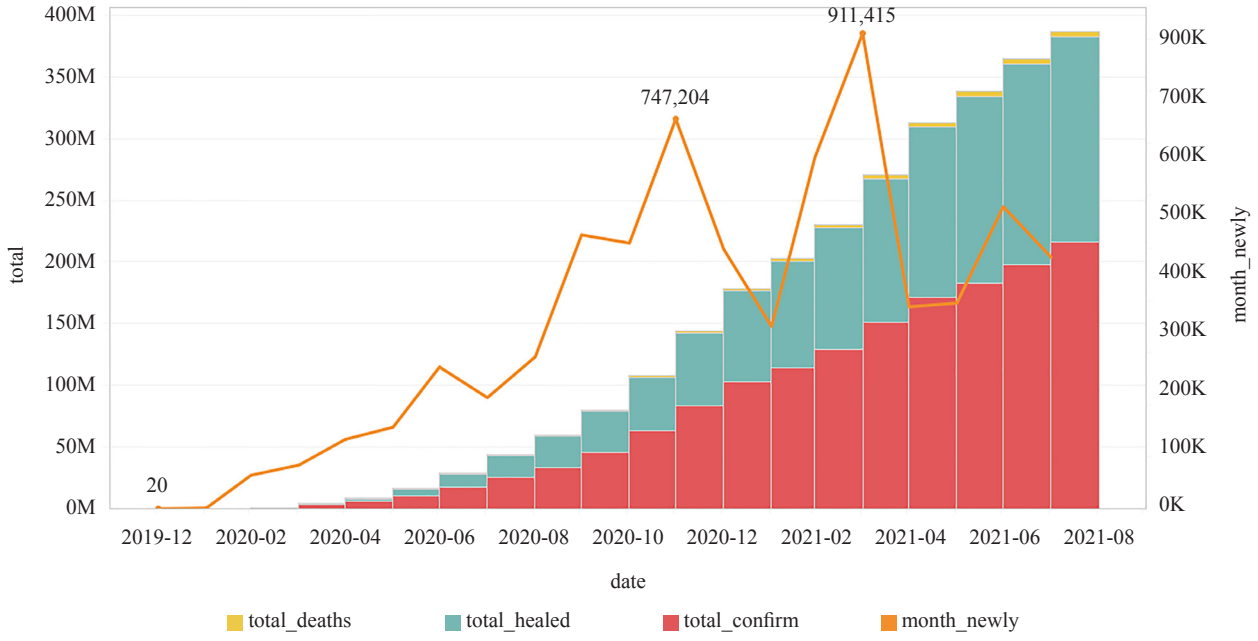

Figure 1. Global development trend of COVID-19

Among the cumulative global cases, the number of confirmed cases in the United States and India topped the list, accounting for 33.28% of the total number of confirmed cases in the world. China's pandemic prevention and control have achieved remarkable results, and the overall situation is relatively optimistic. Therefore, the follow-up study focuses on the comparison of the development situation of the pandemic situation in the United States, India, and China.

As shown in Figure 2, the early stage of the pandemic was dominated by the pandemic in China. During this period, the number of confirmed cases in China increased significantly, reaching the highest number of newly confirmed cases in February. In the face of the sudden pandemic, the Chinese government quickly introduced a series of prevention and control policies such as restrictions on travel, implementation of lockdowns, and consolidation of the medical system to effectively block the spread of the pandemic, normalize the prevention and control of the national pandemic, and then gradually stabilize. The whole domestic area is in a sporadic outbreak state, and imported cases have been controlled basically. The cumulative cases in China have been basically at a level, and the development trend of the pandemic is developing in a good direction. Obvious results have been achieved in prevention and control. In contrast, the pandemic in the United States and India spread rapidly and the situation gradually deteriorated. After surpassing China for the first time in late March and May respectively, the cumulative number of confirmed cases in the two countries has continued to increase substantially, and the cumulative number of confirmed cases far exceeds that of China. The pandemic situation is gradually deteriorating [10]. Among them, the number of newly diagnosed cases in the United States in a single month can reach up to 228 thousand, and the peak of newly confirmed cases in India in a single month can even reach 386 thousand, which is about a hundred times the peak of new cases in China in a single month. With the advancement of pandemic prevention and control work in the later period, the severe situation has eased. However, due to the pandemic of COVID-19 mutations and the imbalance of vaccine supply, the COVID-19 pandemic in the United States and India still faces complex and severe challenges.

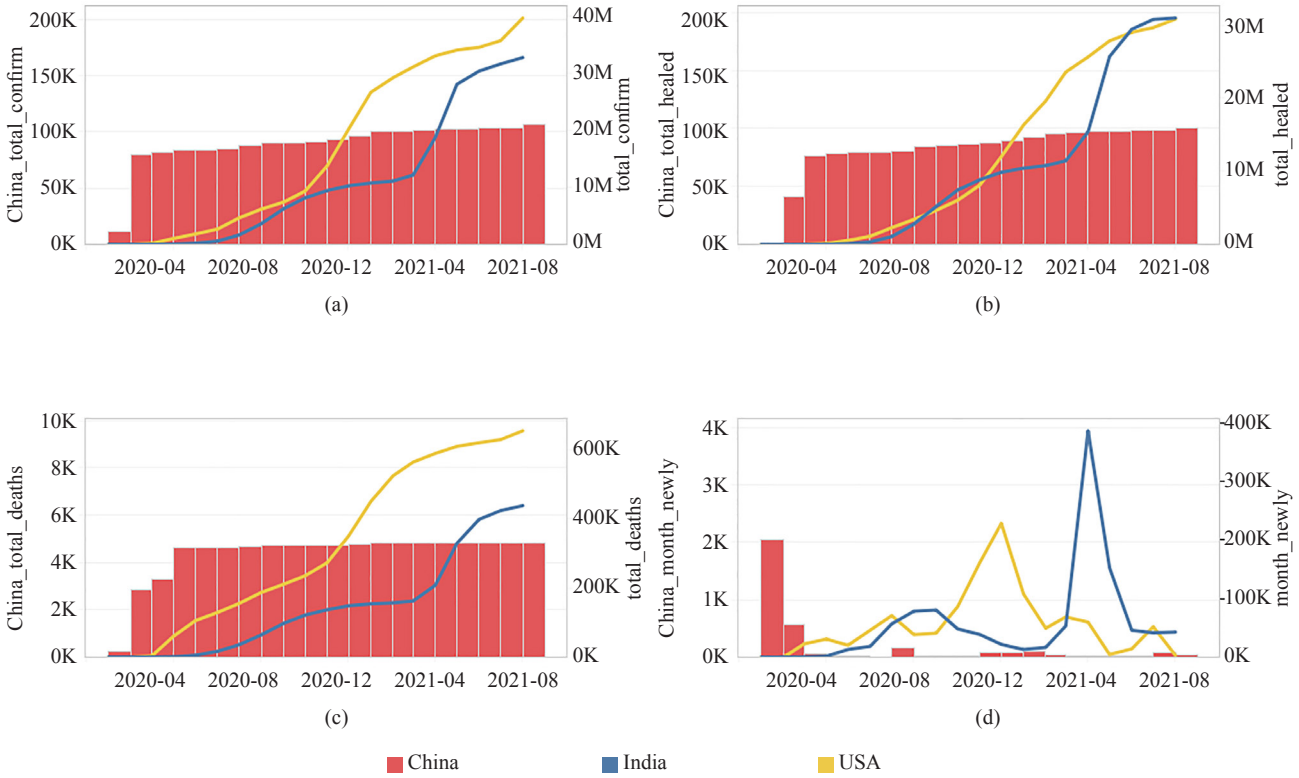

Figure 2. The COVID-19 development trend of China, India, and the USA

Noted: (a) is the cumulative total cases;
(b) is the cumulative total healed cases;
(c) is the cumulative total deaths;
(d) is the monthly new reported cases.

In the map of confirmed pandemic cases in all provinces (except Taiwan Province) in China as shown in Figure 3. The cumulative total confirmed cases in Hubei Province is quite different from that of other provinces. To better observe

the trend of data changes in other provinces, Hubei Province is extracted separately. It can be seen from Figure 3(a) that Hubei Province is the most serious pandemic in China. The cumulative total confirmed cases is 68,298, far exceeding other provinces, followed by Guangdong, Heilongjiang, and Shanghai Provinces. Observing from Figure 3(b), it can be concluded that the current pandemic prevention and control trends in various provinces are relatively good, and the current confirmed cases are within the three-digit range. Due to the pressure of overseas imports and the impact of population movements, there have been sporadic outbreaks in some areas. Fujian is the most serious, followed by Yunnan and Heilongjiang Provinces.

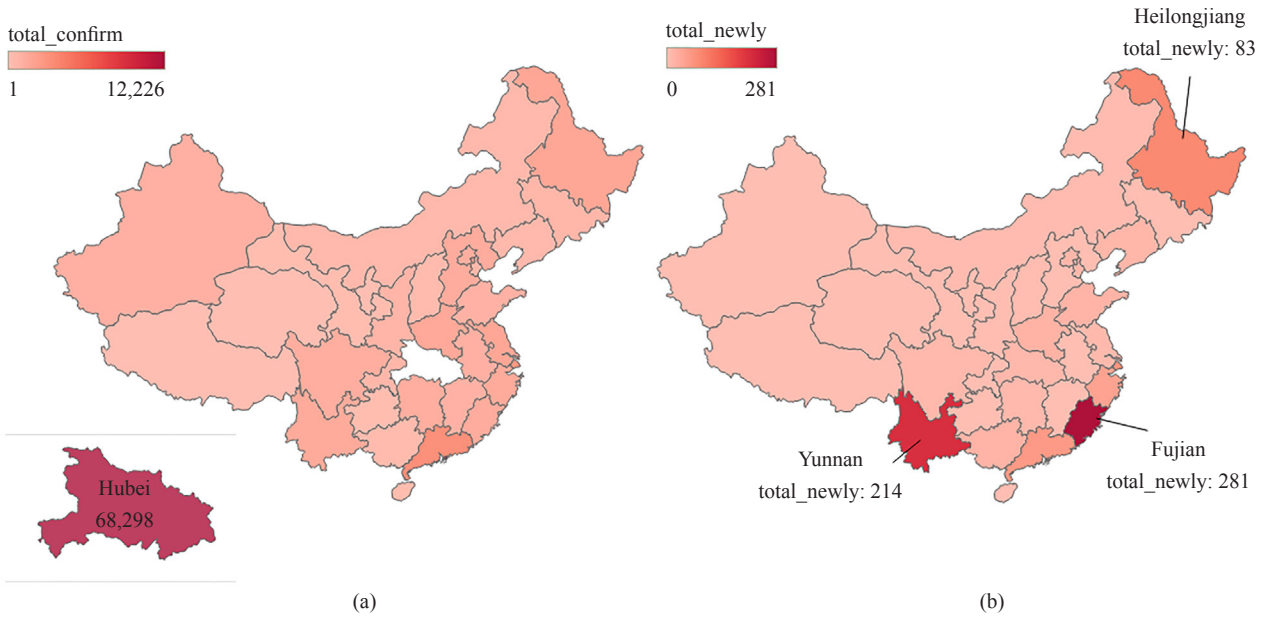

Figure 3. The confirmed cases of COVID-19 in China

Note: (a) is the cumulative total confirmed cases
(b) is the cumulative total newly reported cases

Observing the statistical data of new cases in China as shown in Figure 4, it can be roughly divided into three stages, which are the initial outbreak stage, the control stage, and the closing stage. At the initial stage of the outbreak, the cases were mainly from Hubei Province, with Wuhan city being the most serious. No similar cases or low-risk cases have been found in other areas, but there is a trend of spreading at this time. Affected by the large-scale population movement during the Chinese Lunar New Year, the pandemic has spread across the country. The number of newly reported cases nationwide continues to rise, reaching a peak on February 12, 2020, with 15,159 newly reported cases, the medical burden is too heavy and the pandemic situation is not optimistic. During the pandemic control stage, as the pandemic prevention and control measures continued to deepen, the number of newly reported cases per day gradually decreased, and it had gradually stabilized by March. At the sametime, the number of newly healed cases on a single day continues to increase up to 3,605 cases. Obvious results have been achieved in pandemic prevention and control. At the end of the pandemic, pandemic prevention and control across the country have become normalized with sporadic outbreaks in some areas. Due to the less optimistic situation of the pandemic situation overseas, the mutation of COVID-19 is complicated and the pressure on the imported cases is gradually increasing. At this stage, the main focus is on the prevention and control of overseas imports. Although the pandemic has broken out in some parts of the country, the situation can be quickly controlled.

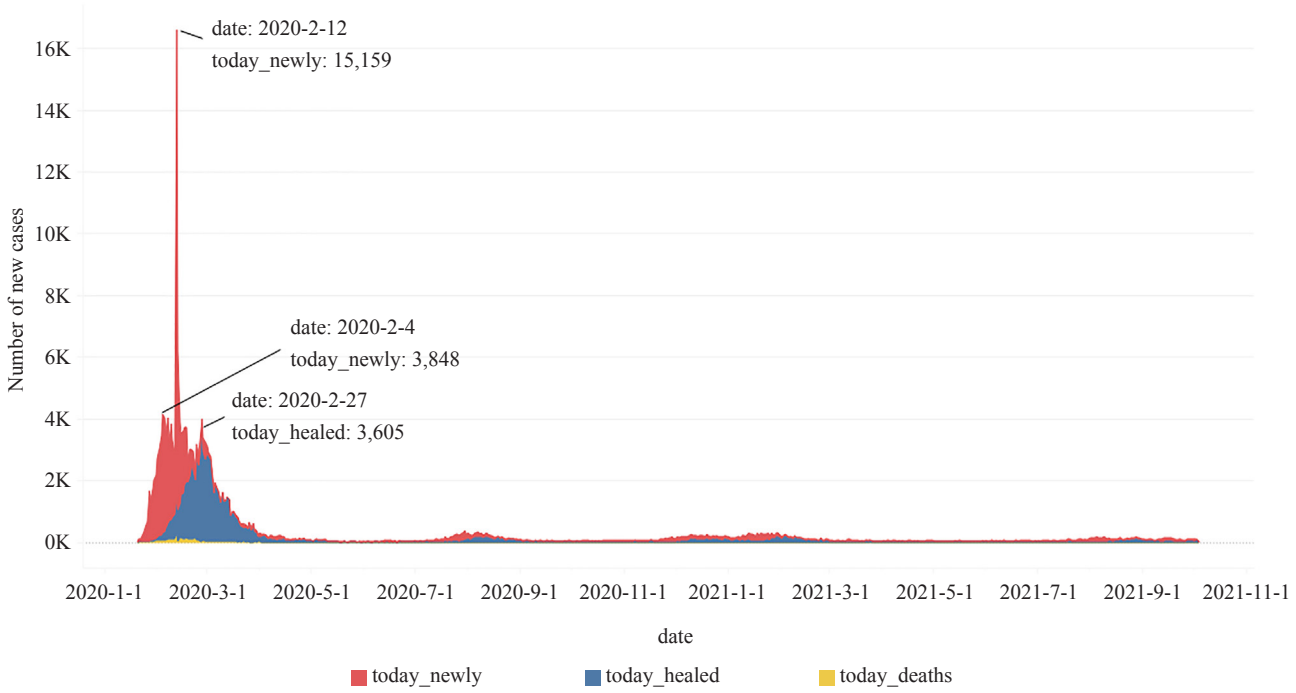

Figure 4. China's COVID-19 development trend

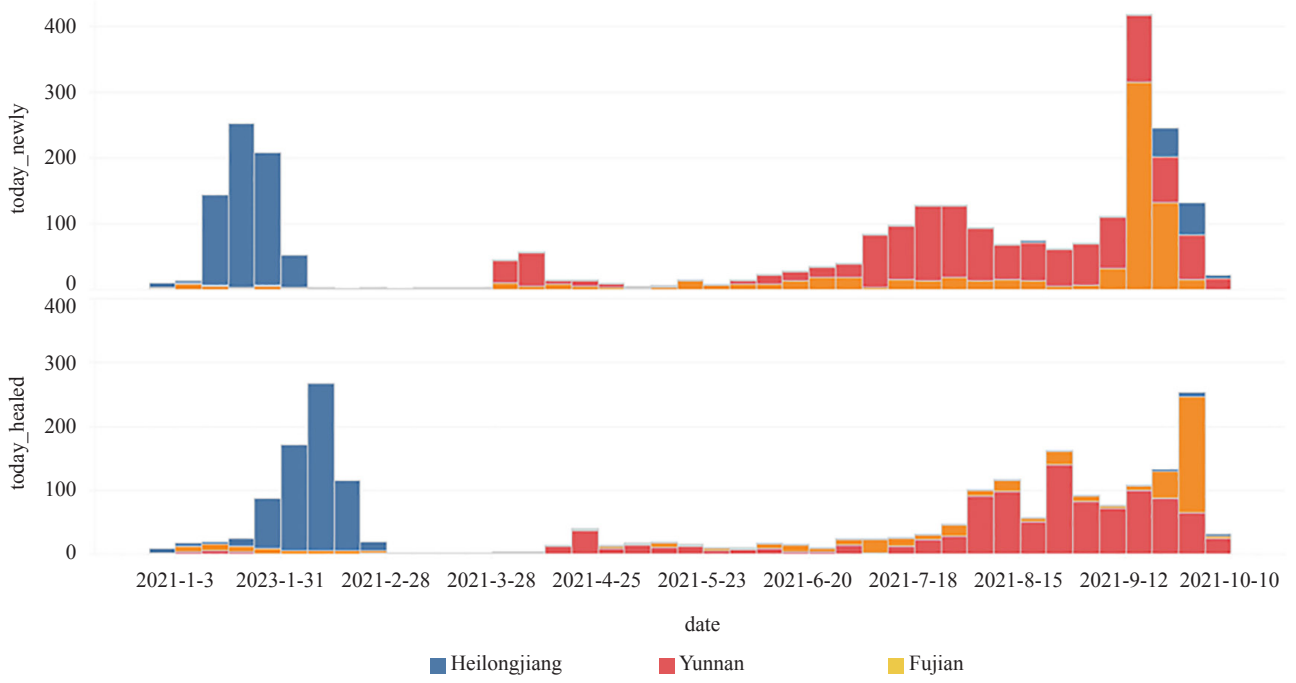

Figure 5. The COVID-19 development trend of Heilongjiang, Yunnan, and Fujian

In this article, the authors also focus on the pandemics in Fujian, Yunnan, and Heilongjiang Provinces in the past 10 months in China. Observing the changes in the pandemic case data in these three provinces, as shown in Figure 5,

the overall pandemic prevention and control situation is good. Judging from the number of new cases, in early 2021, the outbreak in Heilongjiang Province rebounded strongly, while only sporadic outbreaks occurred in the other two provinces. After about two consecutive months, the pandemic situation in the three provinces was in a state of easing. In the later period, affected by the sudden change of COVID-19 and the pressure of overseas import, certain areas of Yunnan and Fujian began to experience a certain scale of pandemic rebound. At this time, the risk of the pandemic in Heilongjiang Province is relatively small. In terms of the number of newly healed cases, the number of healed cases in the three provinces is almost the same as the number of reported cases, indicating that the confirmed patients have received effective treatment and the overall medical burden is relatively small.

4. Data analytics and simulation results

Some useful methods have been proposed to deal with data classification and prediction [11-14]. The decision tree is one of the most useful methods and has been widely used in the field of medical disease prediction [15]. It is a tree structure composed of internal nodes, branches, and leaves. The decision tree algorithm can be used for regression, in which the decision tree uses the squared difference minimization criterion to perform feature selection to generate a binary tree. It. The decision tree is used when the target variable is continuous. Assuming that X and Y are input and output variables respectively, and Y is continuous, the given data is $T = \{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)\}$. A decision tree is generated according to the training data set T. The generation criterion according to the decision tree is based on the square difference minimization criterion, that is, the prediction error is minimized. Find the dividing point, use this point as the dividing line to divide the training set T into two subsets T_1 and T_2, and minimize the square difference between the data sets T_1 and T_2. Then find similar demarcation points in T_1 and T_2, and continue to process until the termination condition is met [16-18]. Letting a leaf is labeled as class for the decision tree T. The information gain (IG) is shown as follows [18-20]:

$$IG(X) = Info(T) - \ln fo_X(T) \tag{1}$$

$$Info(T) = -\sum_{i=1}^{K} \frac{freq(T_i, T)}{|T|} \times \log_2 \left\{ \frac{freq(T_i, T)}{|T|} \right\} \tag{2}$$

$$Info_X(T) = \sum_{i=1}^{k} \frac{|T_i|}{|T|} \times Info(T_i) \tag{3}$$

where $Info(T)$ is the average amount of information needed to identify the class of instances in T, $Info(T)$ is the expected information value for attribute X to the partition T, $|T|$ is the number of cases in T, $freq(T_i, T)$ is the number of cases included in T, n is the number of outputs for attribute X, T_i is a subset of T corresponding to the i^{th} output, and $|T_j|$ is the number of instances in T_j.

Based on the pandemic data collected by the crawler, it is to build a decision tree to obtain the development trend of the COVID-19 pandemic. Based on the current severe development of the pandemic situation in the worldwide and Fujian Province of China, the analysis of worldwide and Fujian Province is mainly carried out.

4.1 *The data analytics of worldwide pandemic*

Taking the number of newly reported cases per day (today_newly) as the output variable, and the rest as input variables for a worldwide pandemic. Up to August 29, 2021, the cases of total_deaths is 4,504,961 and total_confirm is 216,512,035. The rate of total_deaths is defined as = cases of total_deaths/4504961 × 100%, and the rate of total_newly is defined as = cases of today_newly/216512035 × 100%. The rules of the decision tree are obtained as follows:

Rule 1: If 83% ≤ rate of total_deaths < 89%,

rate of today_newly = 0.17%;

Rule 2: If 89% ≤ rate of total_deaths < 90%,
rate of today_newly = 0.203%;

Rule 3: If 93% ≤ rate of total_deaths < 96%,
rate of today_newly = 0.26%;

Rule 4: If 96% ≤ rate of total_deaths < 97%,
rate today_newly = 0.339%.

From the rules of the decision tree, the variable of cumulative total deaths (total_deaths) has the greatest impact on the newly reported cases (today_newly) worldwide. When the rate of total_deaths is between 83% and 97%, the more of the rate of total_deaths and the more of the rate of today_newly.

4.2 *The data analytics of Fujian pandemic*

Taking the number of newly reported cases per day (today_newly) as the output variable, and the rest as input variables for the Fujian pandemic. Up to October 4, 2021, the cases of total_deaths is 1 and total_confirm is 1,287. The rate of today_newly is defined as = cases of today_newly/1287 × 100%. The rules of the decision tree are obtained as follows:

Rule 1: If rate of total_confirm < 63.2%,
rate of today_newly = 0.12%;

Rule 2: If 63.2% ≤ rate of total_confirm < 67.2%,
rate of today_newly = 1.79%;

Rule 3: If 68.2% ≤ rate of total_confirm < 84%,
rate of today_newly = 4.27%.

From the rules of the decision tree, the total_confirm has the greatest impact on today_newly in Fujian Province, China. When the rate of total_confirm is between 63.2% and 84%, the more of the rate of total_confirm, the more of the rate of today_newly.

5. Conclusions

Based on the pandemic-related data collected by the web crawler technology, this paper conducts a data visualization and analytics of pandemic situations in worldwide and China. The decision tree for the current development situation of Fujian and the worldwide are constructed for decision-makers. From the rules of the decision tree worldwide, the variable of cumulative total deaths has the greatest impact on the newly reported cases. When the rate of total_deaths is between 83% and 97%, the more of the rate of total_deaths, the more of the rate of today_newly. For Fujian Province, the case of total_deaths is only one person, and the total_confirm has the greatest impact on the today_newly. However, it is similar to worldwide situation. When the rate of total_confirm is between 63.2% and 84%, the more of the rate of total_confirm, the more of the rate of today_newly.

The risk of a rebound in some areas due to uncertain factors such as the imported cases and the mutation of COVID-19 virus still exists. It is recommended to maintain the necessary pandemic prevention and control measures and strengthen the control of imported personnel. With the implementation of pandemic prevention and control measures and vaccination work in various countries around the world, the COVID-19 pandemic in most countries has been alleviated. However, due to the prevalence of mutation COVID-19 and the uneven supply of vaccines, the global

COVID-19 pandemic still faces complex and severe challenges. It needs to adjust prevention and control measures in a timely and dynamic manner in response to the new characteristics of the mutation of COVID-19.

Acknowledgements

This paper was supported by Fuzhou University of International Studies and Trade, Grant No. 202113762014.

Conflict of interest

The authors declare that there is no personal or organizational conflict of interest with this work.

References

[1] Whitelaw S, Mamas MA, Topol E, Van Spall HG. Applications of digital technology in COVID-19 pandemic planning and response. *The Lancet Digital Health*. 2020; 2(8): 435-440. Available from: https://doi.org/10.1016/S2589-7500(20)30142-4.

[2] Dong L, Bouey J. Public mental health crisis during COVID-19 pandemic, China. *Emerging Infectious Diseases*. 2020; 26(7): 1616-1618. Available from: https://doi.org/10.3201/eid2607.200407.

[3] Xu WW, Wu J, Cao LD. COVID-19 pandemic in China: Context, experience and lessons. *Health Policy and Technology*. 2020; 9(4): 639-648. Available from: https://doi.org/10.1016/j.hlpt.2020.08.006.

[4] Wu HJ, Liu F, Zhao L, Shao YB, Cui R. Application research of crawler and data analysis based on Python. *International Journal of Advanced Network, Monitoring and Controls*. 2020; 5(2): 64-70. Available from: https://doi.org/10.21307/ijanmc-2020-018.

[5] Zhou ZH, Zhang HR, Xie J. Data crawler for Sina Weibo based on Python. *Journal of Computer Applications*. 2014; 34(11): 3131-3134.

[6] Akhtar N, Tabassum N, Perwej A, Perwej Y. Data analytics and visualization using Tableau utilitarian for COVID-19 (Coronavirus). *Global Journal of Engineering and Technology Advances*. 2020; 3(2): 28-50. Available from: https://doi.org/10.30574/gjeta.2020.3.2.0029.

[7] Netease News. Real-time update: COVID-19 pandemic dynamic map [EB/OL]. Available from: https://wp.m.163.com/163/page/news/virus_report/index.html.

[8] The real-time COVID-19 dynamic data. Available from: http://jjh.cngold.org/apps/tg_yq.html.

[9] Comba JLD. Data visualization for the understanding of COVID-19. *Computing in Science & Engineering*. 2020; 22(6): 81-86. Available from: https://doi.org/10.1109/MCSE.2020.3019834.

[10] Chen L, Wang KH. Analysis of the development trend of the global COVID-19 pandemic based on cluster analysis. *Journal of Yuncheng University*. 2021; 39(3): 12-16.

[11] Singh PK. Fourth dimension data representation and its analysis using Turiyam Context. *Journal of Computer and Communications*. 2021; 9(6): 222-229.

[12] Singh PK, Gani A. Fuzzy concept lattice reduction using Shannon entropy and Huffman coding. *Journal of Applied Non-Classical Logics*. 2015; 25(2): 101-119.

[13] Yan HH, Wan JF, Zhang CH, Tang SL, Hua QS, Wang ZR. Industrial big data analytics for prediction of remaining useful life based on deep learning. *IEEE Access*. 2018; 6: 17190-17197. Available from: https://doi.org/10.1109/ACCESS.2018.2809681.

[14] Liang X, Hong TZ, Shen GQP. Occupancy data analytics and prediction: A case study. *Building and Environment*. 2016; 102: 179-192. Available from: https://doi.org/10.1016/j.buildenv.2016.03.027.

[15] Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*. 2004; 18(6): 275-285. Available from: https://doi.org/10.1002/cem.873.

[16] Karnon J. A simple decision analysis of a mandatory lockdown response to the COVID-19 pandemic. *Applied Health Economics and Health Policy*. 2020; 18(3): 329-331. Available from: https://doi.org/10.1007/s40258-020-00581-w.

[17] Li Z, Wang L, Huang LS, Zhang M, Cai XH, Xu F, et al. Efficient management strategy of COVID-19 patients

based on cluster analysis and clinical decision tree classification. *Scientific reports*. 2021; 11(1): 1-13. Available from: https://doi.org/10.1038/s41598-021-89187-3.
[18] Pandya R, Pandya J. C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*. 2015; 117(16): 18-21.
[19] Saeed MS, Mustafa M, Bin W, Sheikh UU, Salisu S, Mohammed OO. Fraud detection for metered costumers in power distribution companies using C5.0 decision tree algorithm. *Journal of Computational and Theoretical Nanoscience*. 2020; 17(2-3): 1318-1325. Available from: https://doi.org/10.1166/jctn.2020.8807.
[20] Marzuki M, Iqbal M, Nivada A, Sofyan H, Usman T, Nazaruddin N, et al. Implementation of decision tree using C5.0 algorithm in preference and electability survey results on regional head election in Aceh. *Journal of Physics: Conference Series*. 2021; 1882(1): 12132. Available from: https://doi.org/10.1088/1742-6596/1882/1/012132.