ICIC Express Letters
Volume **15**, Number **3**, March **2021**

ICIC International ©2021 ISSN 1881-803X
pp. **241–248**

VISUALIZATION OF DATA MINING DISTRIBUTION OF COVID-19 IN INDONESIA USING SELF-ORGANIZING MAPS ALGORITHM

Iskandar Fitri[1], Agung Triayudi[1,*], Iksal[2], Zaenal Muttaqin[3] and Sumiati[3]

[1]Informatic Department
Universitas Nasional
Jalan Sawo Manila, Pejaten Barat, Pasar Minggu, Jakarta 12520, Indonesia
iskandar.fitri@civitas.unas.ac.id; *Corresponding author: agungtriayudi@civitas.unas.ac.id

[2]Electrical Engineering
Universitas Faletehan
Jl. Raya Cilegon KM. 06 Pelamunan Kramatwatu Serang, Banten 42161, Indonesia
iksal_r@yahoo.com

[3]Information and Communication Technology
Universitas Serang Raya
Jl. Raya Cilegon No. Km. 5, Taman, Drangong, Serang 42162, Indonesia
d.zaey.vu@gmail.com; sumiati82@yahoo.com

Received July 2020; accepted September 2020

Abstract. *Coronavirus Disease or COVID-19 is a virus that in late 2019 and early 2020 shook the world. Initially, this COVID-19 came from one of the cities in China, namely the city of Wuhan. In Wuhan City, the total number of cases up to May 2020 was around 82,992 cases. While for COVID-19 cases in the world, it touched 5,495,061 cases. Since the end of 2019 until May 2020, the total number of COVID-19 cases has increased. Because of this case, all countries in the world declared a state of emergency. When viewed from a health viewpoint, socioeconomic conditions in a country belong to the state of health of the community, especially if there is an infectious disease outbreak in the community. The main focus of this research is that the amount of COVID-19 cases is a nonlinear role of socioeconomic impressions that are generally not divided between districts. The aim is to combine multivariate datasets describing social and economic factors in evaluating the system that districts with the same socioeconomic symptoms show a similar distribution of COVID-19. We use the Self-Organizing Maps (SOM) algorithm and the required distribution of 8 social and economic variables to classify 34 provinces in Indonesia into five clusters. Clusters determined by SOM are then compared to the distribution of COVID-19 cases. Our research shows a positive association within socioeconomic circumstances and COVID-19 cases in existing provinces using SOM methods to overwhelm data and methodological difficulties traditionally faced in public health research.*

Keywords: COVID-19, Socioeconomic, Indonesia, Self-Organizing Maps (SOM), Cluster

1. **Introduction.** In this day and age, every day, much new data is collected. The collected data are then managed in a database along with several other correlated terabytes of data [1,4]. Visual Data Mining (VDM) is a process of interaction and analytic reasoning with one or more visual representations of abstract data. This VDM process leads to a visual discovery of a pattern that is in the data or also provides guidance for applying data mining and other analysis techniques [3,6]. VDM can implement in various fields of human endeavour such as business, technology development, healthcare, medicine, and other fields that produce many data that is difficult to understand [8]. To be able to use

DOI: 10.24507/icicel.15.03.241

VDM, visualization technology is needed to transform extensive data in datasets into images or graphics that are easier to read. In short, VDM is a combination of visualization and data mining [8].

Self-Organizing Map (SOM) algorithms or often called topology-preserving map is a technique that was begun by Teuvo Kohonen in 1996 [13]. SOM is one of the techniques in Neural Network (NN) that can be used to visualize data by decreasing data dimensions within self-organizing neural networks so that ordinary people can understand about high-dimensional data mapped in the construction of low-dimensional data. SOM is one of the critical methods in VDM. The learning method used by SOM is without direction from data from unsupervised learning which assumes a topology that is structured into several clusters or classes. Therefore, an observation position can be compared with each unit in the second layer. The algorithm can then find which set best represents the domain of observation. The set is determined on a two-dimensional grid, so the sets that have high resemblance will be close together [5].

Socioeconomic is a social science that investigates how economic movement can influence and develop social manners. In general, this science analyzes how the economic situation of people in the local and global area, whether the economy of the community is progressing, backward, or stable [4,15]. In the early 2020s, when the outbreak of COVID-19 in all countries, the economy in all countries experienced significant changes. According to the UN Framework for the Immediate Socioeconomic, the COVID-19 pandemic is not just a health crisis but has an impact on other cores such as social and economic [16-18].

Research that discusses COVID-19 is very prevalent recently, such as research conducted on social media [19] with a total of 115,299 Weibo posts consisting of an average of 2,956 posts per day. This study used a quantitative analysis by finding a positive correlation between the number of Weibo posts and the number of cases reported from Wuhan, using the four classifications in the Weibo discussion of disease causes, epidemiological characteristics of the outbreak, public reaction to outbreak control and response measures. The results of this study provide initial insight into the outbreak of the corona disease caused by the COVID-19 virus, based on quantitative data and qualitative analysis of Chinese social media data in Wuhan City [22]. This research can carefully conduct an initial analysis through social media Weibo which can be used for early action to prevent COVID-19.

In another study [20], Geographic Information System (GIS) has played an important role in many aspects, such as advantages in the field of accurate and accurate aggregation and visualization for COVID-19 spatial tracking, provision of good spatial information support is very important for decision making, regarding the prevention and control of COVID-19. The use of GIS experiences difficulties when dealing with heterogeneous data acquisition and integration of data; here we need cooperation between government, business, and academic institutions to jointly formulate the right policies. At the technical level, spatial analysis methods are developing very rapidly.

In another study [21], research on geographic tweets posted about COVID-19 found consistent data according to the official WHO report on the incidence of COVID-19 cases during the study period. This reflects the recommended methods of monitoring and tracing these infections very precisely. The limitation of this method is that it cannot be used to monitor and track infectious diseases in poor areas or areas where there is no access to social media. The language of the tweets in this study is English. The results of this study, due to the very rapid development of social media, and using the mining of this web news used by each community, the geographic and demographic of users can be identified accurately. This is due to the fact that statistical data reports can easily be found in comments, photos, videos, etc. regarding COVID-19.

The main contribution in this paper is to visualize the correlation between the socioeconomic and Coronavirus Disease (COVID-19) outbreaks using the Self-Organizing Map

(SOM) algorithm by searching for areas that are more similar in the socio-economic field that have more similar attributes in the transmission of COVID-19. Hidden information can be found from large amounts of data, and provides an intuitive visualization. In this paper, we use the SOM algorithm to analyze the correlation between eight socioeconomic variables and the COVID-19 flash [9].

2. **Research Method.** One example of this application is competitive learning, where output neurons compete with one another to get good results. In the algorithm training process, the first unit layer output SOM or node is assigned a set of random vectors or applied to as a codebook [10,11]. After BMU found a demanding set of input vectors then allocated to the same thing, then the output unit vector value is fixed to be closer to the input set value. The neighbouring BMU unit is also adjusted close to the latest values. Likewise, the entire set of input data is committed to their BMU at the output layer, mapping the same input data vector unitedly on a two-dimensional arrangement with most of the original properties retained. Therefore, an SOM display that has been previously trained can enable analysts to see useful knowledge previously not known implicitly in raw data in the form of patterns, constructions and associations [9].

The standard architecture of a self-organizing map network is shown in Figure 1. The quintessential constituents of the feature map are as ensues [13].

- An array of neurons that calculates a simple output function from an input that enters from input entered arbitrarily by dimensionality.
- The mechanism for selecting neurons with the most significant output.
- The adaptive mechanism that updates the weight of the selected neuron and its neighbours.

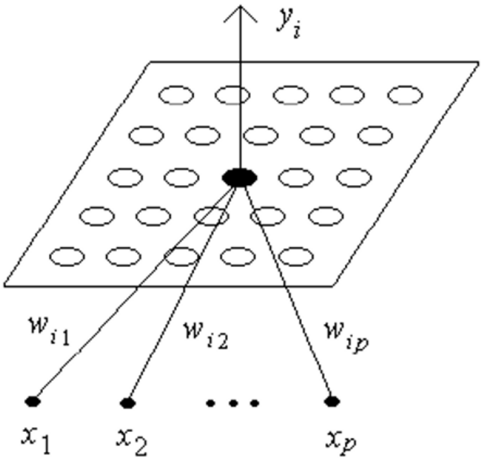

FIGURE 1. The standard architecture of a self-organizing map network

The training algorithm suggested by Kohonen to form the map features is paraphrase as ensues.

Step 1: Initialization: Select a blind value for the initial weight $w_j(0)$.

Step 2: Winning Findings: Find the neuron j^* at time k, adopting the minimum distance Euclidean criteria:

$$j^* = \arg \left\| \underline{x}(k) - \underline{w}_j \right\|, \quad j = 1, \ldots, N^2 \tag{1}$$

where $\underline{x}(k)$ represents the kth input pattern, N^2 is the total number of neurons, and $\|\ \|$ showing Euclidean norms.

Step 3: Weight Update: Adjust the weight of the winner and its neighbors, using the following rules

$$\underline{w}_j(k+1) = \underline{w}_j(k) + \eta(k)\Lambda_{j^*}(k)\left[\underline{x}(k) - \underline{w}_j(k)\right] \tag{2}$$

where $\eta(k)$ is a constant positive and $\Lambda_{j^*}(k)$ are the topological environmental functions of the winning neuron j^* at time k. The typical choice of $\Lambda_{j^*}(k)$ is a Gaussian type function

$$\Lambda_{j^*}(k) = \exp\left(-\frac{d^2_{j^*,j}}{2\sigma^2}\right) \tag{3}$$

where the parameter σ is the "width effect" of the topological environment and $d^2_{j^*,j}$ is the lateral distance between neurons j^* and j in discrete output space. It must be emphasized that the success of map formation is highly dependent on how the values of the main parameters (i.e., $\eta(k)$ and $\Lambda_{j^*}(k)$) the initial value of the weight vector, and the number of iterations are predetermined.

3. **Results and Discussion.** Coronavirus Disease (COVID-19) was first recognized as a global threat in early 2020. The first known case occurred in the city of Wuhan, China, at the end of 2019. The international spread of this outbreak resulted in 5,495,061 cases in 178 countries, with 350,958 deaths. As an example of VDM using SOM, we chose socioeconomic data to train in maps. Then we look for the relationship between the COVID-19 outbreak and socioeconomic factors in 34 provinces in Indonesia.

COVID-19 data covering the number of COVID-19 patients in each region in Indonesia were obtained from the report of the Task Force for the Acceleration of Handling COVID-19, then socioeconomic and population data from the report of the Indonesian Central Statistics Agency (BPS). Table 1 presents an example of the data contained in the dataset.

In the next step, correlation analysis is used to determine the social factors related to economics. The results are as follows in Table 2.

Table 1. Socialeconomic data in Indonesia 2020

Region	**Jakarta**	**East Java**	**Aceh**	**Bangka Belitung Island**
Number of cases	6,400	3,129	19	36
Death rate	500	256	1	1
Population	10,012,271	38,610,202	4,906,835	1,343,900
Retail rate	−34.9326	−19.41573034	−16.0674	−23.4494382
Grocery and pharmacy rate	−15.4382	−8.674157303	−6.69663	−7.898876404
Parks	−37.9438	−26.06741573	−16.3933	−18.96629213
Transit station rate	−40.6742	−32.07865169	−27	−38.25842697
Workplaces	−26.6067	−16.65168539	−11.573	−13.56179775
Residential	14.49438	9.921348315	7.101124	8.505617978

Table 2. Relationship analysis

Attribute 1	**Attribute 2**	**Correlation**
Number of cases	Death rate	0.999684313
Number of cases	Population	0.345605873
Number of cases	Retail rate	−0.807169241
Number of cases	Grocery and pharmacy rate	−0.93769357
Number of cases	Parks	−0.991428192
Number of cases	Transit station rate	−0.565945582
Number of cases	Workplaces	−0.971818333
Number of cases	Residential	0.968553807

Correlation analysis used to determine the social factors related to economics and the retail rate and transit station rate has a low correlation rate in the presence of the COVID-19 outbreak. While the death rate, population, transit station rate, retail rate, and residential have a high level of correlation with COVID-19 outbreaks. From the table above it also shows that the mortality rate rose quite high, and the grocery and pharmacy rate decreased significantly related to the COVID-19 outbreak.

The results of the data clusters are significantly related to input the clusters obtained in Figure 2. In this figure, all data is classified into five parts, where West Java province is separated by itself because it has the most population, then Java province Central and East Java are classified together because they have almost the same data rate. After that, we put all the related factors into the net, which then obtained the following results.

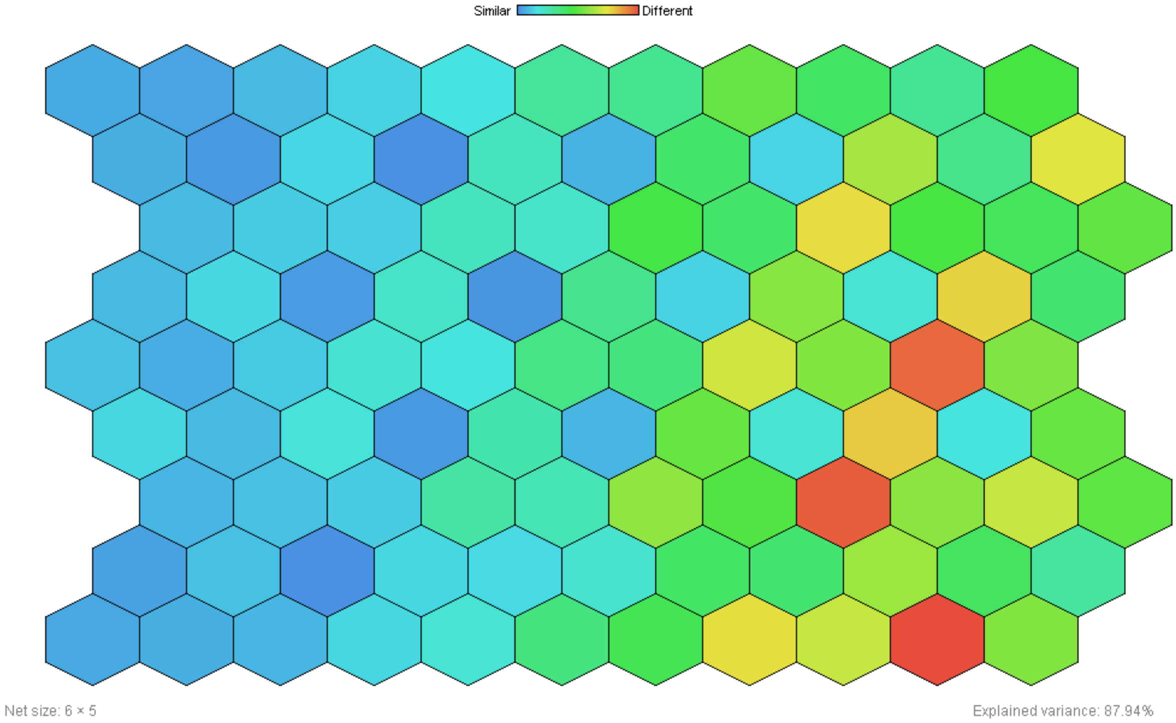

Figure 2. (color online) Clusters as a whole

In Figure 3, there are hexagons, each of which has its values and colours. If the colour of the hexagon is red, it means the hexagon has a value that is far different from other neighbouring data. Conversely, if the colour of the hexagon is blue, then the hexagon has a value that is quite similar to other neighbouring data. From the picture above, it can see that cluster 1 has the most members compared to other clusters. Then in cluster 2, it only has one member.

Table 3 shows the result of data processing using the Self-Organizing Map (SOM) algorithm on RapidMiner. In this research, we use RapidMiner [14] because it can process SOM algorithms without making coding scripts like in Python or R. In Table 3, it can see that cluster 1 has the most number of provinces, namely 19 provinces (Bali, Bengkulu, Special Region of Yogyakarta, Gorontalo, Jambi, South Kalimantan, Central Kalimantan, East Kalimantan, North Kalimantan, Bangka Belitung Island, Riau Island, Maluku, North Maluku, Papua, West Papua, West Sulawesi, Central Sulawesi, Southeast Sulawesi and North Sulawesi). Cluster 1 had a total of 194,947 cases with 7,632 deaths. Cluster 1 also has the lowest workplace rate, which is −15,353. Then in cluster 2, which only has one province, West Java, has the lowest death rate, which is 125 cases. Then in cluster 3 in this result, the provinces of Jakarta together with North Sumatra and Banten were collected into the same group in one cluster. Cluster 3 has the highest mortality rate among

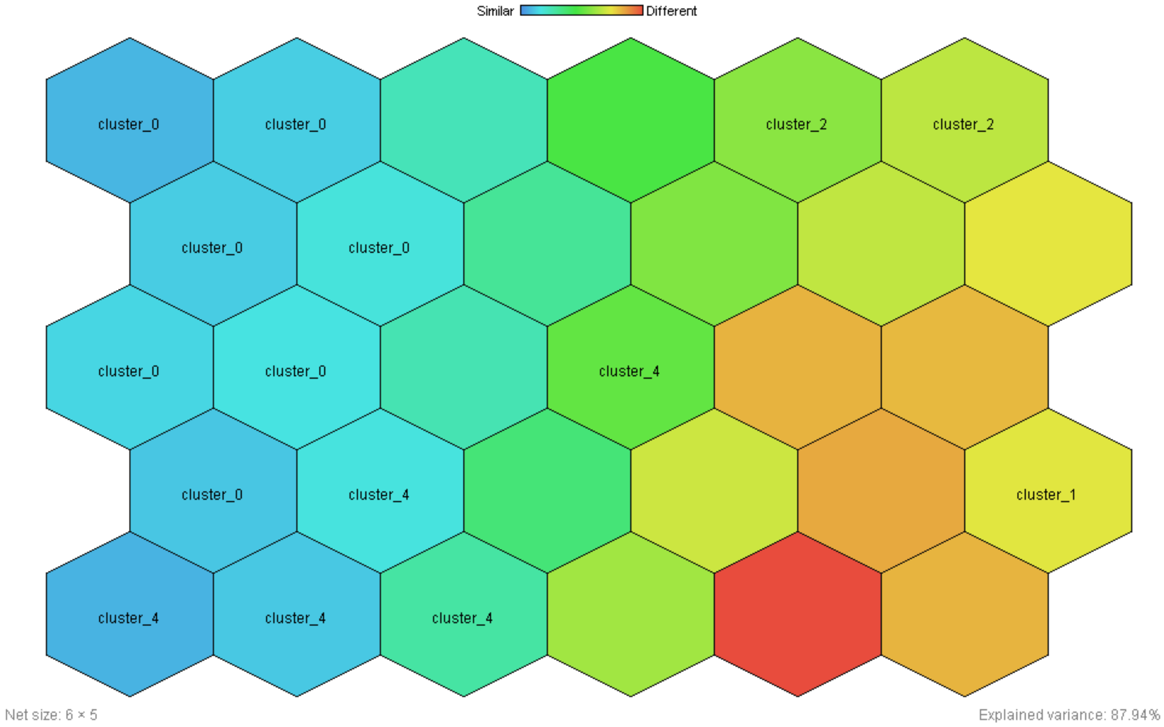

Figure 3. (color online) Clusters use all datasets

Table 3. Cluster result using all related data

Cluster	Province
C1	Bali, Bengkulu, Special Region of Yogyakarta, Gorontalo, Jambi, South Kalimantan, Central Kalimantan, East Kalimantan, North Kalimantan, Bangka Belitung Island, Riau Island, Maluku, North Maluku, Papua, West Papua, West Sulawesi, Central Sulawesi, Southeast Sulawesi, North Sulawesi
C2	West Java
C3	Banten, Jakarta, North Sumatra
C4	Central Java, East Java
C5	Aceh, West Kalimantan, Lampung, West Nusa Tenggara, East Nusa Tenggara, Riau, South Sulawesi, West Sumatra, South Sumatra

other clusters. In addition, cluster 3 also experienced the highest economic decline in all sectors used in this study (retail, grocery and pharmacy, parks, transit stations, workplaces, and residential). Then in cluster 4, namely Central Java and East Java provinces, the highest total cases were 2,181,500, and the death rate was 163. Cluster 4 had quite high rates in the retail, grocery and pharmacy sectors. Furthermore, in cluster 5, it has 9 cluster members, namely, Aceh, West Kalimantan, Lampung, West Nusa Tenggara, East Nusa Tenggara, Riau, South Sulawesi, West Sumatra, and South Sumatra. Cluster 5 had a total of 361,444 cases, with a mortality rate of 14,667. Cluster 5 also has a high rate in the parks sector and low in the residential sector.

4. **Conclusions.** In this paper, the Self-Organizing Map (SOM) algorithm used to visualize the correlation between socioeconomic and Coronavirus Disease outbreaks (COVID-19). The results show that socioeconomic factors correlate with COVID-19 transmission. More similar regions in the socioeconomic area have more similar attributes in COVID-19 transmission. To get further information, data must be collected from several related

sources such as the Task Force for the Acceleration of Handling COVID-19 to retrieve outbreak case data, and the Central Statistics Agency (BPS) to retrieve economic data in all provinces in Indonesia. By using the SOM algorithm, hidden information can be found from a large amount of data, and provides intuitionistic visualization. With this, we can have a more straightforward web structure, strong automatic learning abilities, and calculate quickly. SOM algorithms can be useful in visualizing data mining in the public health sector. The suggestion for further research is to widen the coverage of the data, divided according to the zones established by the Indonesian government at this time, such as the black zone, the red zone, the yellow zone and the green zone, and combined with other algorithms to be able to predict the spread of COVID-19.

Acknowledgment. This research is the result of the basic research scheme of the Indonesian Dikti Grant B/87/E3/RA.00/2020.

REFERENCES

[1] A. Triayudi and I. Fitri, ALG clustering to analyze the behavioural patterns of online learning students, *Journal of Theoretical & Applied Information Technology*, vol.96, no.16, pp.5327-5337, 2018.

[2] A. Triayudi and I. Fitri, A new agglomerative hierarchical clustering to model student activity in online learning, *Telkomnika*, vol.17, no.3, pp.1226-1235, 2019.

[3] J. Chen, W. Wei, C. Gui, L. Tang and L. Sun, Textual analysis and visualization of research trends in data mining for electronic health records, *Health Policy and Technology*, pp.389-400, 2017.

[4] A. Triayudi and I. Fitri, Comparison of parameter-free agglomerative hierarchical clustering methods, *ICIC Express Letters*, vol.12, no.10, pp.973-980, 2018.

[5] G. B. Gebremeskel, Y. Chai, Z. He and H. Dawit, Combined data mining techniques based patient data outlier detection for healthcare safety, *International Journal of Intelligent Computing and Cybernetics*, vol.9, no.1, pp.42-68, 2016.

[6] D. Gu, J. Li, X. Li and C. Liang, Visualizing the knowledge structure and evolution of big data research in healthcare informatics, *International Journal of Medical Informatics*, pp.22-32, 2017.

[7] P. Joao and O. Postolache, Healthcare outlier detection with hierarchical self-organizing map, *International Conference on Sensing and Instrumentation in IoT Era*, Lisbon, Portugal, 2019.

[8] A. Kunjir, H. Sawant and N. Shaikh, Data mining and visualization for prediction of multiple diseases in healthcare, *International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, Chirala, India, pp.329-334, 2017.

[9] H. Kurdia and N. AlMansour, Identifying accurate classifier models for a text-based MERS-CoV dataset, *Intelligent Systems Conference (IntelliSys)*, London, UK, pp.430-435, 2018.

[10] N. Nijiru and E. Opiyo, Clustering and visualizing the status of child health in kenya: A data mining approach, *International Journal of Social Science and Technology*, pp.128-156, 2018.

[11] E. Oliver, I. Valles-Perez, R.-M. Banos, A. Cebolla, C. Botella and E. Soria-Olivas, Visual data mining with self-organizing maps for "self-monitoring" data analysis, *Sociological Methods & Research*, pp.1-15, 2016.

[12] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich and J. Lessler, The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application, *Annals of Internal Medicine*, vol.172, no.9, pp.577-582, 2020.

[13] T. Kohonen, E. Oja, O. Simula, A. Visa and J. Kangas, Engineering applications of the self-organizing map, *Proc. of the IEEE*, vol.84, no.10, pp.1358-1384, 1996.

[14] R. Klinkenberg, I. Mierswa and S. Fischer, *RapidMiner*, 2006.

[15] H. A. Rothan and S. N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak, *Journal of Autoimmunity*, 2020.

[16] F. Stephany, N. Stoehr, P. Darius, L. Neuhäuser, O. Teutloff and F. Braesemann, The CoRisk-Index: A data-mining approach to identify industry-specific risk assessments related to COVID-19 in real-time, *arXiv Preprint*, arXiv:2003.12432, 2020.

[17] S. M. Ayyoubzadeh, H. Zahedi and M. Ahmadi, Predicting COVID-19 incidence using Google Trends and data mining techniques: A pilot study in Iran, *JMIR Public Health and Surveillance*, 2020.

[18] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe and X. Cheng, Artificial intelligence and machine learning to fight COVID-19, *Physiological Genomics*, vol.52, no.4, pp.200-202, 2020.

[19] J. Li, Q. Xu, R. Cuomo, V. Purushothaman and T. Mackey, Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: Retrospective observational infoveillance study, *JMIR Public Health and Surveillance*, vol.6, no.2, 2020.

[20] C. Zhou, F. Su, T. Pei, A. Zhang, Y. Du, B. Luo and C. Song, COVID-19: Challenges to GIS with big data, *Geography and Sustainability*, 2020.

[21] K. Jahanbin and V. Rahmanian, Using Twitter and web news mining to predict COVID-19 outbreak, *Asian Pacific Journal of Tropical Medicine*, vol.13, 2020.

[22] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T. L. Gao, W. Duan, K. F. Tsoi and F. Y. Wang, Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo, *IEEE Trans. Computational Social Systems*, vol.7, no.2, pp.556-562, 2020.