Spatial Epidemiological Analysis of Early COVID-19 in the Municipality of Los Baños, Laguna, Philippines using K-means Clustering

Jonardo R. Asor

Abstract—**This paper aims to analyze the spread of COVID-19 in the municipality of Los Baños, Laguna in the Philippines through the use of clustering algorithms. The record of the COVID-19 cases in Los Baños from March 2020 up-to March 2021 was used as dataset which includes susceptible, probable, confirmed, recovered and death cases. Following the clustering technique in data mining, a model was created to further analyzed the patterns of COVID-19. Three famous clustering algorithms were used in this study namely; K-means, K-medoids and mean shift. Furthermore, GeoPandas was used in this study for spatial analysis using cluster data while evaluation metrics for clustering such as Dunn index and Euclidean distance dendrogram were used to inspect clustering capability. Through the use of Dunn index, the study had identified K-Means as an efficient clustering method for COVID-19 cases. Hence, shown in this paper that barangay Tuntungin Putho, Mayondon, San Antonio, and Batong Malake formed a relationship.**

Index Terms—**COVID-19, contagious disease, clustering, machine learning, pattern recognition**

I. Introduction

As of April 2021, the total number cases of COVID-19 worldwide were reach up to 130, 459, 184 [1] while in Philippines, the totality reach up-to 1,006428 total COVID-19 cases [2]. It is a viral disease that held the awareness of the public worldwide for longer a year and will negatively be recognized as one of the unusual pandemics which had molded the decade and present time [3].

Due to the worldwide pandemic of COVID-19, a few rules and regulations to avoid the transmission were taken. A monstrous lockdown, restrictions and quarantine protocols have been executed by governments in the Philippines [4] and this response depicted as being one of the longest hence, The most severe lockdown in the world. The whole world was placed into lockdown, versatility was restricted, and the wearing of face mask and face shield and social distancing were strictly implemented [5, 6]. Violation was met with correctional activity. The public authority depended strongly on manpower like police and the military to guarantee that protocols and restrictions were kept up and that all well-being agreements were followed [5]. Moreover, the "StayAtHome" policy was elevated by the Philippine government to avoid the spread of COVID-19. World health organizations are right now developing a vaccination; at this point, there is no compelling and proper medication that has been developed for the treatment of Coronavirus contaminations [4, 5]. Until this point in time, the Philippines is encountering one of the most pessimistic scenarios of the COVID-19 outbreaks as the second-higher affirmed cases and passing in the ASEAN, close to Indonesia [2].

The quick spread of COVID-19 has forced researchers to make serious counter-measures to end this epidemic [5]. Different research works have proposed and carried out different advances to diminish the undesirable results of the pandemic also, to quicken the recuperation stage. Some of these advancements incorporate artificial intelligence (AI), Machine learning, deep learning and data mining among others. There is incredible potential for these advancements to achieve an insurgency in the medical services industry [7]. AI and related advancements are progressively predominant in business and society, and also starting to be applied to medical [8]. These innovations can possibly change numerous parts of patient consideration, just as regulatory cycles inside supplier, payer and pharmaceutical industries [6]. There are now various studies proposing that AI can proceed just as or better than people at key medical care undertakings, like diagnosing and predicting diseases [9].

Clustering algorithms have turned into the main method utilized in each space of the computational work [10]. Gaining from uneven data collections has turned into a vital issue in machine learning lately and is frequently utilized in different applications like computer security, designing, biomedicine, and medical [11]. It is an interaction where a gathering of unlabeled examples was divided into number of sets with the goal that comparative examples are appointed to a similar group, and disparate designs are relegated to various clusters. There are two objectives for clustering algorithm: deciding great clusters and doing as such proficiently [12]. Clustering has turned into a generally concentrated issue in an assortment of use areas, for example, in data mining and information revelation, factual information investigation, data clustering for pattern identification and analysis [13]. Understanding of clustering algorithms to COVID-19 status in barangays can be used to have crisis management, medical services and pattern identification and analysis [14].

In this paper, clustering algorithms such as K-means, K-medoids and mean shift was employed in clustering COVID-19 cases in barangays within the local municipality of Los Baños using the March 2020 to April 2021 records. By using

Manuscript received January 23, 2023; revised March 23, 2023; accepted May 25, 2023.

J. R. Asor is with Laguna State Polytechnic University, Los Baños, Laguna, Philippines. E-mail: asor.jonardo@lspu.edu.ph (J.RA.)

doi: 10.18178/ijfcc.2023.12.3.604

these methods, it is possible to find the patterns from each potential Barangays which can be used in an effort for analyzing COVID-19 cases based on the spatial data similarity [15].

II. Materials and Methods

A. Materials

In executing this project, different software and libraries were used which are well-known in data mining and machine learning implementation. Google colaboratory is used as the main platform in executing the codes for clustering implementation. Hence, libraries like Pandas, Numpy, Matplotlib and Sklearn were used to implement machine learning algorithms in clustering. GeoPandas is another library in python used in this project to conduct spatial analysis using geographical representation and to show the political boundaries of the municipality of Los Baños.

B. Data Acquisition and Preparation

The COVID-19 cases recorded by the Information and Communication Systems Office of Los Baños, Laguna, Philippines from March 2020 to March 2021 was gathered and used as dataset in this study. It contains the following variables, date, address, and status.

The dataset attained was scaled using standard scaling to normalize the data and then used as an input vector for the feature extraction phase. Filling in missing qualities, smoothing uproarious information, eliminating anomalies and settling irregularities was done first as preprocessing method. Then reducing the volume of data (but protecting the pattern) by eliminating rehashed perceptions and applying occurrence choice just as component choice strategies. Discretization of nonstop ascribes is additionally a method of information decrease.

Standard scaler for the feature scaling was also conducted in this study. Standard scaler assumes the data is normally distributed within each feature and will scale them such that the distribution is now centered around 0, with a standard deviation of 1**.**

$$\boldsymbol{z = \frac{x-\mu}{\sigma}} \tag{1}$$

whrer μ is the mean.

σ is the standard deviation.

For feature extraction, Principal Component Analysis (PCA) is implemented in this paper. PCA is commonly used for reducing dimensions of datasets by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data.

C. Clustering and Evaluation

In this phase, three centroid-based clustering algorithms, K-means, K-medoids, and Mean shift. K-means and K-medoids requires a parameter k or the number of clusters which could be determined using the elbow method and silhouette analysis.

Elbow method is utilized by selecting the value of k at the "elbow" or the point after which the distortion/inertia start decreasing in a linear fashion. In Fig. 1, it can conclude that the optimal number of clusters for the data is 3.

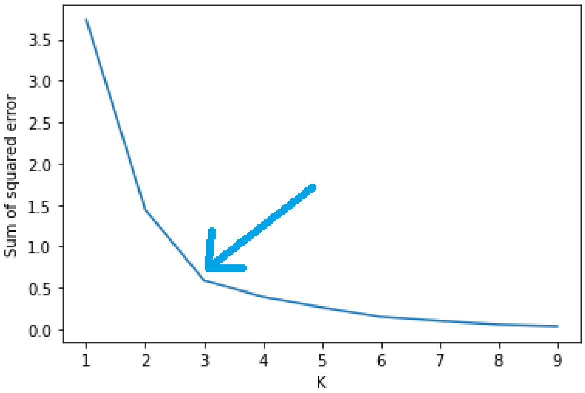


Fig. 1. Sum of Square Error for Each *k* Iteration.

It is worth to note that the nature of elbow method makes it ambiguous to determine the optimal number of k. An alternative for elbow method is the silhouette analysis. Silhouette analysis measures how close each point in a cluster is to the points in its neighboring clusters. Each data point is given a silhouette value with a range of -1 to 1. When a data point has a silhouette value closer to -1 it means that that point is closer to its neighboring cluster than its own and vice versa. The average silhouette of observations for different values of *k* is computed and whichever *k* maximizes this average silhouette is the optimal number of clusters. Fig. 2 is example of silhouette analysis of K-means where k = 4, which shows the thickness of the silhouette plot and displaying that cluster 0 clustering most data points.

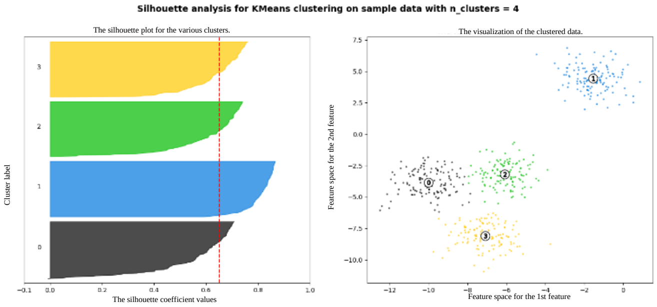


Fig. 2. Silhouette analysis for K-Means clustering with 4 clusters sample.

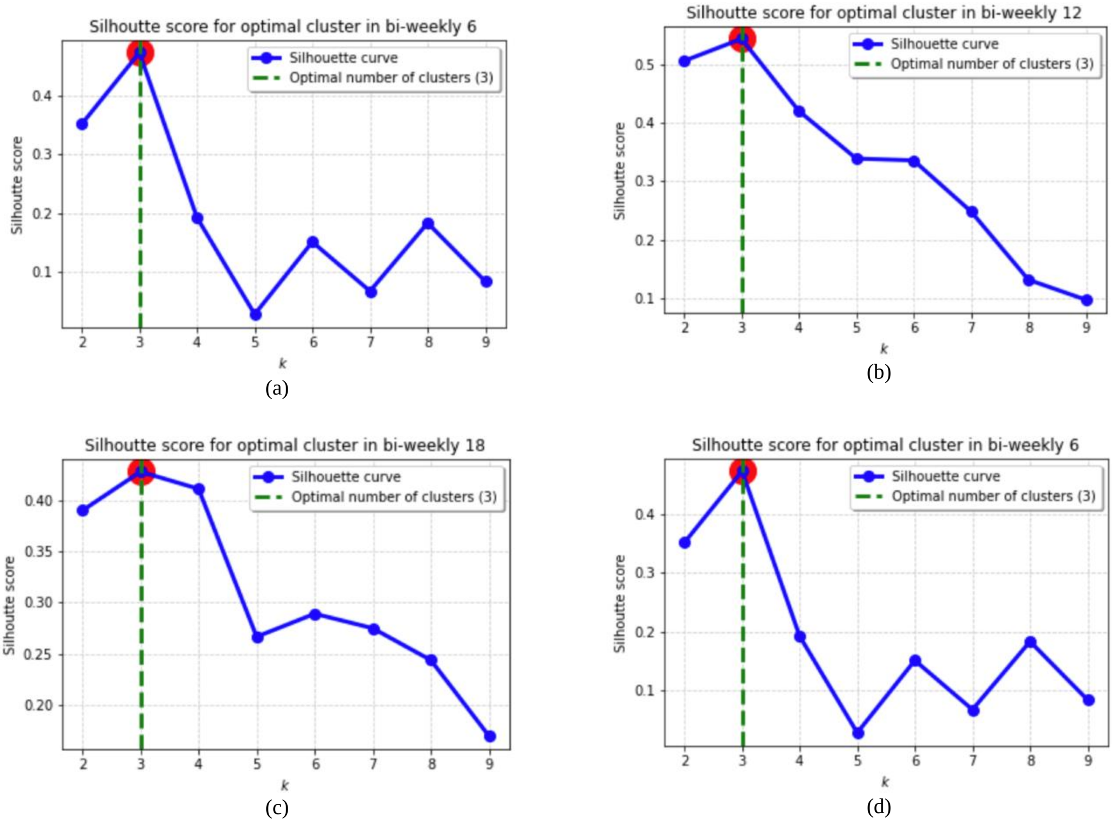
Like silhouette analysis, Dunn index is an internal cluster validation technique used to measure the compactness of clusters by calculating the distance of data points inside the cluster and measuring how well separated clusters are by measuring the distance of different clusters. Dunn index should be maximized therefore the algorithm which has the highest Dunn index performed the best clustering of the data points. Below is how Dunn index is computed:

$$\boldsymbol{dunn\ index = \frac{min_{1\le i<j\le n} d(i,j)}{max_{1<k<n} d'(k)}} \tag{2}$$

where *i*, *j*, and *k* are each index for clusters, d measures the inter-cluster distance, and d' measures the intra-cluster difference.

III. Result and Discussion

In finding the optimal number of clusters, silhouette scoring has done. The clustering was done in bi-weekly to make it congruent in the policy of the municipality of Los Baños in declaring confirmed cases of COVID-19 which is also found in the [16, 17]. Hence, the clusters were presented in quarterly manner to show the whole year in most efficient way. Fig. 3 shows that the most optimal number of clusters in quarterly analysis are 3 and 2. Therefore, it is necessary to take a look at every 2nd and 3rd clusters in spatial analysis.



(a)

(b)

(c)

(d)

Fig. 3. Optimal clusters based on silhouette analysis.

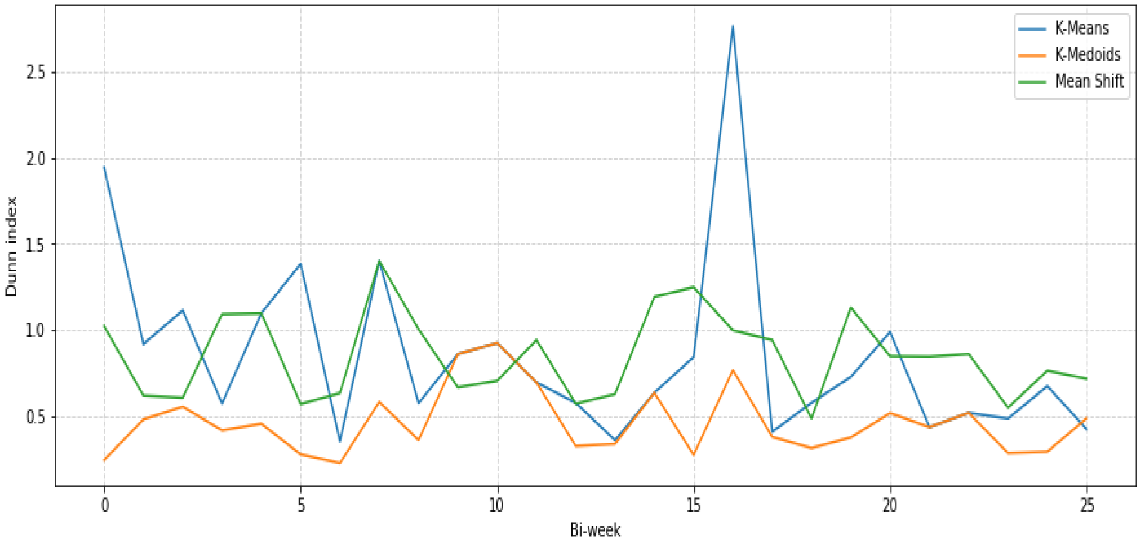
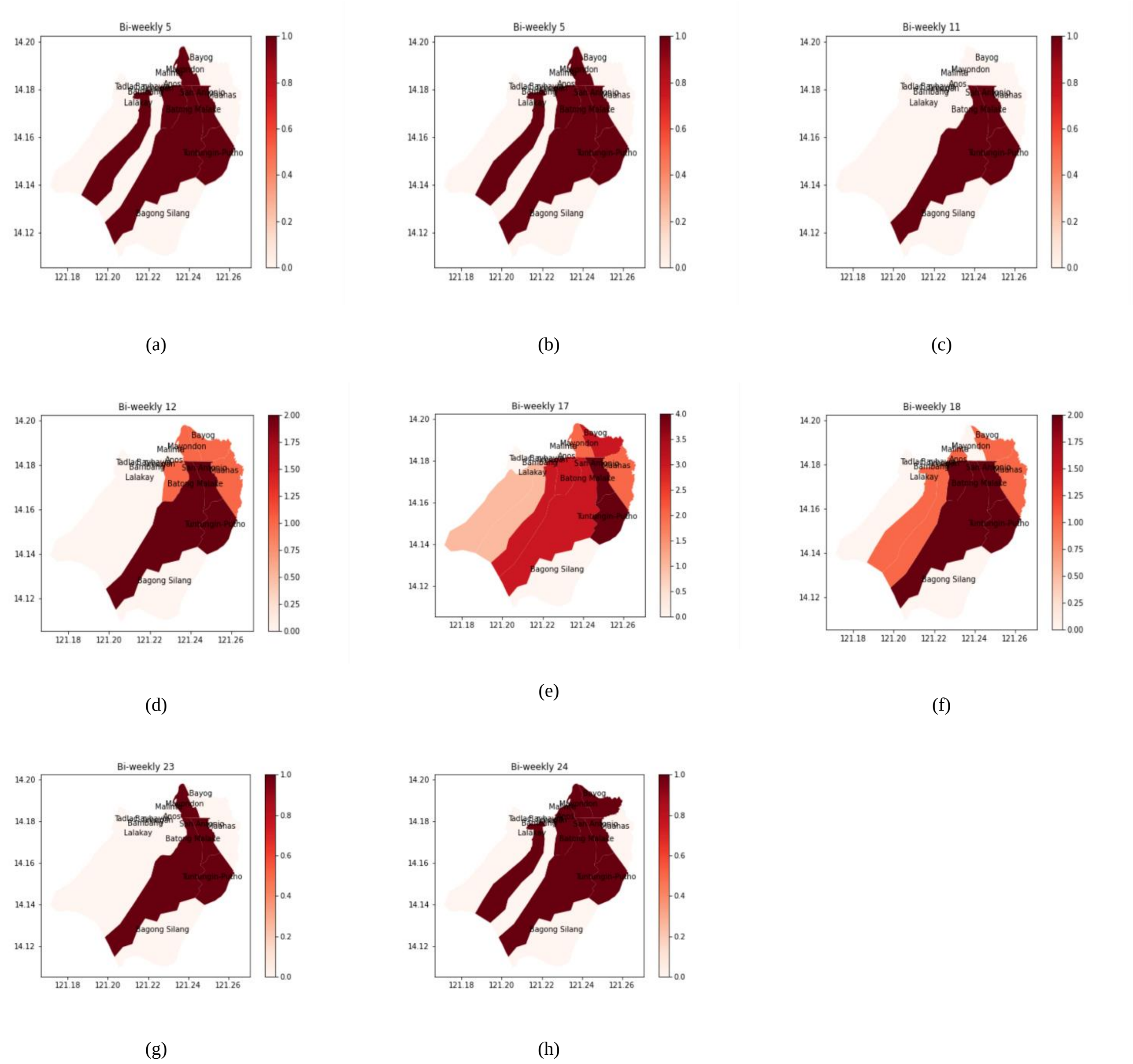


Fig. 4. Performance of clustering algorithms using Dunn index

Fig. 4 shows the result of performance testing of K-Means, K-Medoids and Mean Shift, it is noticeable that in K-Means and Mean Shift performed well in each cluster. Thus, K-Means had shown a promising result after gaining more than 2.5 Dunn index in the 16th cluster. This shows that among the three clustering algorithms, K-Means performs well in clustering the COVID-19 cases dataset.

Fig. 5 shows the geospatial analysis of COVID-19 in the municipality of Los Baños. The intensity of red color shades corresponds to the cluster where each barangay belongs in every bi-weekly analysis. As shown in the figure, there are 2, 3, 4, and 5 clusters found in the dataset.

In the first bi-weekly of the 1st quarter of the year (a), there are 2 clusters found while there are 3 present clusters on the next bi-weekly (b). Likewise, on the 2nd quarter of the year, the same number clusters are still found. The first bi-weekly of the quarter (c) still got 2 clusters while the following bi-weekly got (d) got 3 clusters. For the 3rd quarter, the first bi-weekly (e) produces a total of 5 clusters and the next bi-weekly (f) got 4 clusters. In the 4th quarter, both the two bi-weekly (g and h), got only 2 clusters. All of these clusters contain density which is represented by the intensity of the color. Meaning, the darker the color the higher the risk each cluster in COVID-19. Moreover, it is noticeable that barangay Batong Malake always got the higher risk in the disease. Hence, it is frequently clustered with barangay San Antonio, Tuntungin Putho and Mayondon. This epidemiological scenario shows that Batong Malake must have a contribution in the spread of COVID-19 in its neighboring barangay specifically, Tuntungin Puho and San Antonio. Moreover, Fig. 5 shows that among the barangay in Los Baños, Laguna, Philippines, Batong Malake is the most vulnerable in COVID-19 followed Tuntungin Putho and San Antonio thus, Bagong Silang is the safest barangay in the disease.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig 5. Result of spatial analysis of COVID-19 using K-Means.

IV. Conclusion

In this study, clustering approach in data mining is conducted to analyze the spread of COVID-19 in the municipality of Los Baños, Laguna in the Philippines. Geopandas is a library that can be used for spatial analysis. While Dunn index is a metrics for finding which clustering, algorithms is most suitable for specific dataset. Among the tested clustering algorithms, it is found that K-Means is the most promising algorithm to cluster the COVID-19 cases in Los Baños. Hence, using silhouette score it is found tha 2 and 3 is the optimal number for each cluster done by K-means algorithm in this study.

Using GeoPandas, it is found in this study that barangay Batong Malake is associated with barangay Mayondon, San Antonio, and Tuntungin Putho. Likewise, it is the most vulnerable barangay in COVID-19. Looking at the epidemiological scenario, Batong Malake can be concluded as the main source of the spread of COVID-19 in the municipality. Being the central barangay for different necessary transaction during the COVID-19 pandemic, Batong Malake may become the source where of the spread where each citizen of Los Baños, Laguna, Philippines unintentionally made physical contacts with each other.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

The author would like to acknowledge the local government unit of Los Baños, Laguna, Philippines for entrusting their data to accomplish this study. Likewise, the author would like to give gratitude to the Laguna State Polytechnic University – Los Baños Campus for allowing the conduct of this research.

References

[1] World Health Organization, "COVID-19 weekly epidemiological update," WHO, 2021.

[2] Representative Office of the Philippines, "Coronavirus disease 2019 (COVID-19) situation report #75," World Health Organization, Philippines, 2021.

[3] A. S. K. Rashid, H. N. Abduljabbar. and B. Alhayani, "Coronavirus disease (COVID-19) cases analysis using machine-learning applications," *Applied Nanoscience*, vol. 13, 2021.

[4] B. Paital, K. Das, and S. K. Parida, "Inter nation social lockdown versus medical care against COVID-19, a mild environmental insight with special reference to India," *Sci Total Environ*, vol. 728, 2020.

[5] K. Hapal, "The philippines' COVID-19 response: Securitising the pandemic and disciplining the pasaway," *Journal of Current Southeast Asian Affairs*, vol. 40, no. 2, pp. 224–244, 2021.

[6] J. M. V. Seventer and N. S. Hochberg, "Principles of infectious diseases: Transmission, diagnosis, prevention, and control," *International Encyclopedia of Public Health*, no. 10, pp. 22–39, 2017.

[7] S. A. Alanazi, M. M. Kamruzzaman, M. Alruwaili, N. Alshammari, S. A. Alqahtani, and A. Karime, "Measuring and preventing COVID-19 using the SIR model and machine learning in smart health care," *J Healthc Eng*, 2020.

[8] J. Budd, B. S. Miller, E. M. Manning, V. Lampos, M. Zhuang, M. Edelstein, G. Rees, V. C. Emery, M. M. Stevens, N. Keegan, M. J. Short, D. Pillay, E. Manley, and I. J. Cox, "Digital technologies in the public-health response to COVID-19, *Nat Med*, vol. 8, pp. 1183–1192, 2020.

[9] T. H. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Hospital Journal*, vol. 6, no. 2, pp. 94–98, 2019.

[10] K. Chitra and D. Maheswari, "A comparative study of various clustering algorithms in data mining," *International Journal of Computer Science and Mobile Computing*, vol. 6, no. 8, pp. 109–115, 2017.

[11] K. D. Bailey, "Numerical taxonomy and cluster analysis," *Typologies and Taxonomies*, pp. 35–65, 1994.

[12] K. M. Kumar and A. R. M. Reddy, "An efficient k-means clustering filtering algorithm using density based initial cluster centers," *Information Science*, vol. 418–419, pp. 286-301, 2017.

[13] M. Azarafza, M. Azarafza, and H. Akgün, "Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran," *Journal of Applied Science, Engineering, Technology, and Education*, vol. 3, no. 1, 2021.

[14] K. G. Soni and A. Patel, "Comparative analysis of K-means and K-medoids algorithm on iris data," *International Journal of Computational Intelligence Research*, vol. 13, no. 5, pp. 899–906, 2017.

[15] C. Y. Lee and E. K. Antonsson, "Dynamic partitional clustering using evolution strategies," in *Proc. 2000 IEEE International Conference on Industrial Electronics, Control and Instrumentation, 21st Century Technologies*, vol. 4, pp. 2716–2721, 2000.

[16] Y. Wang, C. Dong, Y. Hu, C. Li, Q. Ren, X. Zhang, H. Shi, and M. Zhou, "Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: A longitudinal study," *Radiology*, vol. 296, no. 2, 2020.

[17] A. P. G. D. Souza *et al*., "A spatial-temporal analysis at the early stages of the COVID-19 pandemic and its determinants: The case of Recife neighborhoods, Brazil," *PLoS ONE*, vol. 17, no. 5, 2022.