PalArch's Journal of Archaeology of Egypt / Egyptology

COVID-19 PANDEMIC DATASETS BASED ON MACHINE LEARNING CLUSTERING ALGORITHMS: A REVIEW

Halat Ahmed Hussein[1], *Adnan Mohsin Abdulazeez*[2]

[1]Collage of Science University of Duhok, Kurdistan Region, Iraq

[2]Duhok Polytechnic University Duhok, Kurdistan Region, Iraq

E-mail: Helat.ahmed@uod.ac, adnan.mohsin@dpu.edu.krd

Halat Ahmed Hussein, Adnan Mohsin Abdulazeez. Covid-19 Pandemic Datasets Based On Machine Learning Clustering Algorithms: A Review-- Palarch's Journal Of Archaeology Of Egypt/Egyptology 18(4), 2672-2700. ISSN 1567-214x

Keywords: COVID-19 Datasets, Clustering, Machine Learning, K-Means. Partitioning Clustering, Hierarchical Clustering

ABSTRACT

COVID-19, standing for Corona Virus Disease in 2019, was a major problem emerged in 2019 and 2020. This has led many intellectuals and scientists to think creatively of ways to eradicate its negative effects. Accordingly, many people have been referring to this pandemic on social media and other media outlets. Within this reference, people have given a lot of data and predictions. Computer intelligence and digital analysis is also one of the fields that has taken this into consideration using clustering algorithms. Clustering can be defined as an approach to place identical data in one community or cluster and separate unsimilar data in another group. There are many Clustering algorithms used for clustering COVID-19 Pandemic datasets. The aims of this paper are to give an overview of the clustering algorithms used in case of COVID-19 datasets, to show how these algorithms help to provide accuracy for clustering the COVID-19 Pandemic, and provide an explanation of the algorithms used for the purpose of either lessening or controlling COVID-19 that are discussed in different papers. Moreover, it details the datasets in terms of different variables like temperature, countries, media outlets including social media, and present the findings of these papers, and which clustering algorithms used and the accuracy of these algorithms. It is found out from the present overview that the clustering algorithm *k-means* is used widely in different types of the COVID-19 datasets with high accuracy.

INTRODUCTION

Computer science is one of the fields, like other fields, which paid attention to COVID-19. People have discussed COVID-19 in all media outlets including

social media. The main process used in computer science for dealing with COVID-19 is Clustering and its algorithms to refer to and deal with COVID-19(Zhang et al., 2020)(Qader Zeebaree et al., 2021). Computer intelligence and digital analysis is also one of the fields that has taken this into consideration using clustering algorithms. Clustering is an important feature that is also applied to learning data. The unsupervised clustering is defined as segmenting the data into clusters that contain data of the same characteristics, which mainly means sorting the data to make homogenous groups(Kushwaha et al., 2020). Clustering algorithms are applied in different fields namely, image segmentation, data cleaning and exploratory analysis, information retrieval, web pages grouping, market segmentation and scientific and engineering analysis(Mobasher, 2007)(Pham & Afify, 2007). Clustering can also be used as a preprocessing step to identify pattern classes for subsequent supervised classification (Garcia, 2020). Clustering is the subject of active research in several fields such as statistics, pattern recognition, machine learning, and data mining.

As stated earlier, Clustering is an approach which classifies the data used into similar and unsimilar group(Dodd, 2014)(Bargarai et al., 2020). That is to say homogenous clusters(Arora & Chana, 2014). To put it simply, a clustering algorithm places a large number of data points having all characteristics in common in a smaller number of classes. That could be referred to as cleaning the data because data sharing the same properties are in different categories and vice versa. Clustering has many uses, including segmentation of images and of large sets of data in social media like tweeter, health care field, market segmentation, and review of publications about COVID-19 (Krishnachandran, 2007). It can be used for data sorting and for exploratory analysis.

Clustering can also be used to categorized pattern corporations for next supervised class as a preprocessing phase (Iskandar Fitri, Refly Asmar, 2020). The current paper focuses on clustering the COVID-19 datasets in different fields where very broad data sets are typical with several attributes of various kinds. This puts on the clustering algorithm difficult computational conditions. A number of clustering algorithms that satisfy these criteria have recently appeared, and several of these algorithms have been successfully implemented to real-world data mining applications and they constitute the core subject of this paper. Clustering, similar to other methods, played an important role in finding out more about the causes and the conditions of the virus. That was an attempt to clean the noisy data spread worldwide in order to inform biological fields where researchers were making every effort to know how the virus lives outside the human body, and to know impact of the different variables like the temperature, populations, and on the spread of COVID-19. In addition to that the cleaning data led to results that can be useful to control the spread of the virus and help health units to make the right decisions to fight this virus. Thus, the paper is organized as follows: section 2 is a presentation of COVID-19 pandemic, section 3 provides a review of different clustering techniques used with COVID-19, and section 4 is an assessment of the application of these algorithms. Section 5 concludes the paper with a summary of some of the key research findings in clustering in regard to COVID-19.

COVID-19

Wuhan in China, in the last two months of 2019, witnessed a pandemic of pneumonia of no obvious causes and origins. However, it was later identified as caused by a new coronavirus(Panwar et al., 2020). It later spread to many other countries rapidly. The novel virus resembles MERS coronavirus and SARS coronavirus and it is given the abbreviated name as COVID-19 to stand for CORONA VIRUS DISEASE IN 2019(Panwar et al., 2020). World Health Organization (WHO) assigned the label pandemic to this disease, which was caused by the SARS-CoV-2 virus, because it is was highly infectious and hence became an issue of concern and debate all over the world(Liu et al., 2020).

This has caused risks to different countries especially those with poor health systems. It is known that pandemics grow with high speed. However, they cannot grow rapidly forever(Wyplosz, 2020). Eventually, the virus will finish either because most people have already been infected/killed or because we will obtain the ability to control it(Oniani et al., 2020). As the situation was different from one country to another due to not following the same restrictions and regulations or not responding to COVID-19 in similar ways, different conditions were posed throughout the whole world. SARS-CoV-2 spreads when people get in close contact with each other especially those infected with it during family and friend gatherings (Shen et al., 2020).
To control such virus, early detection of these gatherings and isolation of infected people is a preliminary must. As a vital prevention measure, very new and modern geospatial tools as important digital tools are used to point out the precise locations where patients with COVID-19 reside (Rossi et al., 2020). These methods support an up-to-date clustering and help in monitoring the spread of COVID-19 in terms of time and space. This can help in creating strategies that can give a good knowledge to epidemiologists and decision makers to intervene on the local levels(Yazdani et al., 2020).

Machine learning, similar to other methods, played an important role in finding out more about the causes and the conditions of the virus. That was an attempt to clean the noisy data spread worldwide in order to inform biological fields where researchers were making every effort to know how the virus lives outside the human body, and to know impact of the different variables like the temperature, populations, and on the spread of COVID-19. In addition to that the cleaning data led to results that can be useful to control the spread of the virus and help health units to make the right decisions to fight this virus (Kushwaha et al., 2020).

Clustering Algorithms

Machine Learning (ML) has become the most important technique used in every area of the computational work. Learning from unbalanced data sets has become a crucial problem in machine learning in recent years and is often used in various applications such as computer security, engineering, biomedicine, and healthcare (Abdulqader et al., 2020) one of the machine

learning technique is clustering. Clustering is to place a group of similar objects (in some sense) in one set i.e. cluster and another group of objects sharing similar characteristics in another group i.e. cluster(Nath & Levinson, 2014)(Abdulqader et al., 2020)(Zeebaree et al., 2018). One can classify Clustering algorithms into three parts: hierarchical, partition and density-based clustering of which CURE is one good example. One can further classify the hierarchical clustering method into agglomerative and divisive techniques. If we describe CURE, it is the partition algorithm that has k-means and Fuzzy algorithms while density- based has DBSCAN and OPTICS algorithm(Garcia, 2020)(Zeebaree et al., 2018). Figure (1) illustrates the overview of clustering methods in general (Arora & Chana, 2014). There are four categories of these techniques (Mobasher, 2007)(Najim Adeen et al., 2020): hierarchical methods, partitioning methods, density-based methods, and grid-based methods(Nath & Levinson, 2014).

Partitioning methods are techniques whose main benefit is to improve clustering quality that is better than the original one. Their main task is summarized in the RELOCATION of data points between the clusters(Y. Li et al., n.d.) (Zebari et al., 2020). Hierarchical methods' main task, on the other hand, is structuring clusters progressively by either joining the smaller clusters with larger ones, or by partitioning the larger clusters. Another type of clustering method is the density-based ones which connect regions that mark sufficiently high data points to make clusters(Pham & Afify, 2007)(Zebari et al., 2020). Grid-based methods divide the data space into a finite number of partition cells that form a grid structure, and then use the dense grid cells to form clusters. This is one way to make clustering more efficient. It is noteworthy to point out that there are traditional clustering methods that, until the present time, are not applied directly to the COVED-19 datasets available up to date(Hu et al., 2020)(Rahman et al., 2020). In the following section, we will present an overview of the clustering methods used for clustering datasets related to COVID-19 only, although the following figure shows other clustering methods not applied to COVID-19. The presentation of the other clustering is given to enrich the paper with more information about clustering.

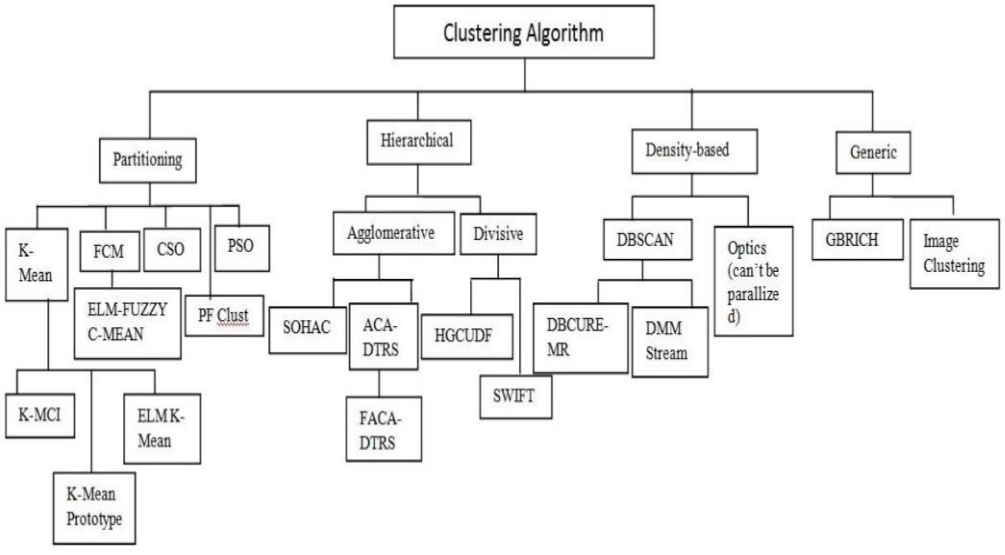

Figure 1: Overview of The Machine Learning Clustering Methods

K-means

The simplest unsupervised learning algorithms is the K-mean clustering algorithm. It helps classify a set of existing data into a certain number of groupings to identify K clusters. The first step of the algorithm is to assign the midpoint of the clusters arbitrarily as K points(Oheneba-Sakyi, 1992)(Zeebaree et al., 2017). Then, in the second step, sets of given data are connected with the nearest center and then by taking the average of the data point close to the midpoint, a cluster is produced. (Godinho et al., 2015). The process is repeated for each midpoint. The third step is represented in repeating the second stage until the midpoints reach some fixed points. The midpoint of the cluster (the assigned center) represents each cluster. The equation of the K-means method is shown in the following equation where $x_i^{(j)}$ is the data points fitting to the *jth* cluster, C_j is the midpoint of the *jth* cluster, k is the number of clusters and n_j is the number of data in cluster j(Jahwar, 2021)

$$E = \sum_{j=1}^{k} \sum_{i=1}^{nj} \left\| x_i^{(j)} - C_j \right\|^2 \quad (1)$$

A six-stage K-means clustering flowchart is shown in figure (2). It comprises six essential stages. The first stage represents the preliminary value of centroids: Let (C1, C2, ….) indicates centroids harmony. At the second stage, the distance among objects is calculated taking two things into consideration: the cluster centroid and the objects in the cluster. The distance Euclidean is used and the distance matrix is then calculated with the iteration 0. Each object is represented by a column in the matrix of distance. In other words, in the first row, the distance of matrix matches the distance of every object to the second row and the first centroid is a sign to the distance of every object in the second centroid. In the third row, which is the clustering of objects, every object is allocated based on the least distance. In the fourth row which is iteration-1, the centroids are determined by identifying the components of all groups and the new centroid of every set. It is computed on the basis of these memberships that are new. In the fifth row, the process is repeated from step 2. In the sixth row, the last iteration grouping is compared and this iteration states that groups are not moved by the objects and thus K- means clustering computation means that it has become stable and there is no need of iteration anymore (Jahwar, 2021)

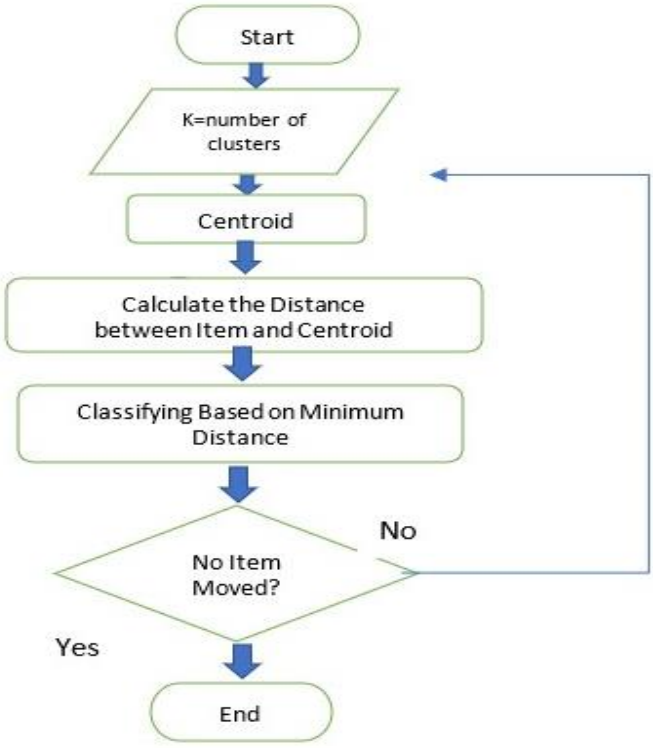

Figure 2: K-Means Clustering Flowchart (Zeebaree et al., 2017)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

The objectives of the DBSCAN approach are to classify the candidate points into main points, boundary points, or outliers. Under the given two parameters, eps and minpts, DBSCAN will effectively return clusters of arbitrary shapes. DBSCAN's clustering process is to decide the number of points within the predefined eps neighborhood distance of a given point if the qua assumes a set of sample points(Mahesh Kumar & Rama Mohan Reddy, 2016). DBSCAN is the most remarkable density-based clustering method. In this method, the clusters are identified as areas of higher density than the rest of the data. Objects needed to separate clusters in these sparse areas are mostly regarded to be noise and border points(Duan et al., 2007).

Hierarchical clustering algorithms

Hierarchical clustering is a self-explanatory term as it is a method of building a hierarchy i.e., groups of clusters in a dataset. This method is also termed as hierarchical cluster analysis or (HCA). The hierarchical clustering functions in the following way: It merges or combines clusters at one lower level producing clusters at a higher level(Rokach & Maimon, 2006). More obviously, the clusters at each level of the hierarchy are formed by combining clusters at the next lower level(Murtagh, 1983). Being formed in this way, at the last lowest level, each cluster contains one single observation but at the highest level, there will be only one cluster where the whole data is saved. Hierarchical clustering has two methods: the agglomerative method (or the bottom-up method)(Yu et al., 2004). In this method, the process starts from

the bottom and at each level recursively a selected pair of clusters are joined into one single cluster resulting in a grouping at the next higher level minus one cluster(Rokach, 2010). The second method is the divisive method which, contrary to the previous one, starts from the top and hence a top-down method. It is one of the popular algorithms of divisive (DIVISIVE ANALYSIS) DIANA(Nietto & Do Carmo Nicoletti, 2017).

Self -Organizing Map (SOM)

The SOM models reduce the dimensions of a dataset into a map. In other words, it is a method of visualizing the data as a map. It is associated with the nodes of a regular cluster and groups the data of the same features together i.e. it reduces data dimensions and display the data in the form of clusters(Kohonen, 2013). Also, SOM is one of the neural networks that has a set of neurons organized on 2D grid(Pampalk, 2001). All neurons are associated with all input gadgets with the resource of the use of a weight vector. The weights are determined through iterative critiques of a Gaussian network function, with the cease end result of creating a 2D topology of neurons to model(Villmann, 1999)(Zeebaree et al., 2019).

Long Short-Term Memory LSTM

The LSTM is an enhancement of comprising a processor that decides whether or not the knowledge is usable, of which the cell is named as the functioning portion(Jahangir et al., 2020). The input layer door, the forget-gate, and the output layer door are three doors in a cell. Both the input layer door and forget-gate function on the cell state. The function of the input doors, however, is to selectively record new ones (Karadayi et al., 2020).The LSTM is advanced and developed Recurrent Neural Network (RNN) that solve the problem of the RNN to handle and remember the sequences of the input states with a length higher with 10 times (Kratzert et al., 2018)

Fuzzy C-Means (FCM)

Fuzzy C-Means (FCM) set of rules is a clustering set of rules primarily based on partitioning, which assign concepts with the largest similarity to identical clusters, whilst those with minimal similarity among to distinct clusters(Wang et al., 2015) . Also, FCM is the maximum famous fuzzy clustering set of rules that is fantastically touchy to noise and outliers and length of the clusters(Ben Ayed et al., 2014). Therefore, many researches accomplished success over those issues. Possibilistic C-Means (PCM) is provided to address the records containing noise and outliers. Sensitivity to initialization and producing coincident cluster facilities are the principal issues with this set of rules(Min et al., 2018)(Omran et al., 2007). To recognize which clustering set of rules is higher than the others we should recognize their benefits and drawbacks as proven in table (1)(Tilson et al., 1988)(Villmann, 1999)

Table 1: Advantage and Disadvantage of Clustering Algorithms

Algorithm	Advantage	Disadvantage
K-means	• Fast, robust • Relatively efficient • Give best result if data well separated	• Need to define K value in the beginning • Not work with overlapping • Is not invariant to non-linear
Fuzzy C-mean	• Best result for overlapped data • Data point can belong to more than one cluster center	• Need to specify the number of the clusters
Gaussian (EM)	• Provides highly beneficial outcomes for the data set in the real world.	• Is high complex in nature
Hierarchical Clustering	• No need to specify the number of the cluster in the beginning • Easy to implement	• Time consuming • Sensitive to noise and outliers • Breaking large clustering
Quality Threshold	• Number of clusters is not specified apriori • All possible clusters are considered	• Time Consuming • distance and minimum number of elements in the cluster has to be defined
DBSCAN	• No require the number of clusters • Good with noise data • is able to find arbitrarily size and arbitrarily shaped clusters	• Fails in case of neck type of dataset
SOM	• Easily interpret the data and presentation • Handle several types of classification problems	• requires necessary and sufficient data • nearby data points behave similarly
LSTM	• can deal with long-term sequence dependencies • no need to set the parameters in the beginning	• it is more complex and time consuming • require memory

RELATED WORK

In this section, the related work of the various CVID-19 Pandemic clustering techniques is presented and explained. Divided this section in to three categories according to the clustering algorithms used

Partitioning Clustering

Rodrigo M.(Carrillo-Larco & Castillo-Cara, 2020) proposed a method to cluster the countries to groups dependent on the COVID-19 level of effect by using (K-means). The features used in this algorithm are disease prevalence estimates, air pollution, health system estimates, and economic crisis. The study depended on the open source of dataset of about 155 countries. Using this number of countries (155), the model of three PCA (principal component analysis) clustered the countries to 5 groups according to the number of those infected with COVID-19. The study didn't take the number of death or even the number of fatality rate into consideration. They compared the results of the K-means clustering by using ANOVA test. The results of the comparison found that the model with five to six clustering is ($p<0.001$). The proposed method is applied only to the confirmed cases.

Mohammad Khubeb Siddiqui(Siddiqui et al., 2020) used K-means clustering to analyze his data depending on the existence of association between the temperature and the three cases of COVID-19 disease(complete, suspected and death cases). The datasets used for this analysis were obtained from WHO for different regions of China. When they applied the K-means clustering, it was noticed that the clustering to 30 groups has occurred and the main finding was that temperature was not just an effective factor in the spread of the COVID-19 pandemic. With investigating the effect of temperatures in the three cases of the disease (confirmed, suspected and death), they suggested that other factors may have played an effective role in the spread of COVID-19.

Marichi Gupta(Gupta et al., 2021) worked on clustering social media tweets attempting to find out if the temperatures will affect the spread of COVID-19 or not. The datasets used were 166,005 English tweets posted from January 23 until June 22, 2020. To clean the data from the repeated words and topics from the tweets, they used a machine-learning algorithm. They removed (4803) repeated tweets, and the remaining tweets (23,752) were clustered. They used (K-means) methods to cluster the tweets. The clustering groups were K= 25, and each group was correlated with an output of the 20 keywords. They found that cluster 10 concerned the cold weather effect on COVID-19 spread. Cluster 24 discussed the impact of the high temperatures on the COVID-19 which discussed the different climates and the general distribution of the virus. Cluster 6 compared COVID-19 with influenza virus, cluster 20 and 22 grouped tweets about how COVID-19 became slow to spread and how the social distance affected the spread of COVID-19 while cluster 11,14,18,21 referenced different topics about scientific experts and Trump's contributions on tweeter about COVID-19. From this clustering process, the results of the work showed 19,33.5 % no effect, 26.1% having effect and 40.4% of uncertain effect of weather on COVID.

Ayman Imtyaz (Imtyaz et al., 2020) presents another user of the k-means clustering to know how the government responds to the COVID-19 pandemic, where K=5. They found a correlation between fatality rates and old age as they found out that the death rates/case fatality levels for the top 30 countries are directly proportional to the ratio of the old aged people (people over 65 years of age). Western European countries have had the most considerable death rate, while South Asia and the Middle East countries have shown the lowest death rate when all other variables are controlled.

To analyze the immune systems and the blood of the hospitalized patients with COVID-19, Yan Zhao (Zhao et al., 2020) used both the k-means and the agglomerative hierarchical clustering. By using three cytokines (IL-1RA and IL-10 and RANTES), they extracted 2 k-mean clusters across the patients and then allocated each cluster to patients with established disease intensity and assessed if the cluster assignment could differentiate the severity of the disease. They found that 9 out of 15 patients in cluster 2 had significant diseases, compared with just 1 out of 10 in cluster 1 ($P = 0.018$). Similarly, when they identified the patient clusters by hierarchical clustering, which is an alternative method to the study of k-means and does not determine the number of clusters, they noticed 2 different clusters (10 of 16 patients in cluster 2 had extreme disease, compared to 0 of 9 in cluster 1 [$P = 0.0028$]).

Chandu, Viswa (Chandu, 2020) used the K-means clustering to find if there is a correlation between the high public expenditure and COVID-19 cases. Their datasets comprised 107 COVID-19 report, population of the peoples age>65 years, and public health expenditure. The algorithm clustered the 89 countries into two group cluster 1 = 54 and cluster 2= 35. Analyzing the clustering, they found that America, European countries and Australia formed a large part of the cluster 2 with strong COVID-19 high death incidences, a higher proportion of the population of the world tested positive for COVID-19, a higher percentage of GDP invested as public health spending, and a higher percentage of the population over 65 years of age. However, as they pointed out, low level of death registration rates with comparatively weaker public health systems in place may be the reason for the observation of the low case fatality rate among countries in cluster 1. To test the hypothesis, they used t-test to check the distance of the cluster members from their respective cluster centers ($t =-3.96$; $p = 0.001$) for the inter-cluster discrepancy.

From table 2, it is obvious that many researchers used the K-means in different fields of COVID-19. From discussing and findings, we came to the conclusion that K-means algorithm is the most widely used algorithm for clustering in topics related to COVID-19 in different variables like the effects of the population on the spread of the virus in these countries with high populations(Kurniawan et al., 2020)

Thang Hoang(Ta et al., 2020) used affinity propagation and mean shift to cluster the lateen topic from AL Jazeera news articles to mark the impact of coronavirus on human beings. The list provides 3400 direct quotes from Al Jazeera news stories. After cleaning the datasets, removing repeated words, using ITEFS for clustering subjects into number, and clustering these topic

number in to groups, they used affinity propagation, and mean shift algorithms because of their advantage of not having to pre-define the number of clusters. In the experiment, affinity propagation makes further clusters that equate mean shift (with smaller numbers of objects in each cluster)

David (Oniani et al., 2020) used the Density-based spatial clustering of applications with noise (DBSCAN) and applied the t-distributed stochastic neighbor embedding (t-SNE) algorithm to reduce the embeddings for all entity nods into 2D space . t-SNE is not used for clustering. It is used to redundant the dimension of the embeddings the results of the DBSCAN algorithm 36 clusters and the silhouette score=0.128.

Mohammed Reza (Mahmoudi et al., 2020) applied the fuzzy clustering algorithm to identify the correlation between the size of the population and the spread of COVID-19 in different countries. They did this by rescaling the datasets of the COVID-19 based on the population of the United States of America. The rescaled datasets of COVID-19 of countries were then clustered by using fuzzy algorithms. The results of the clustering showed that the distribution of spreads in Spain and Italy was roughly identical and varied in comparison with other countries.

Hierarchical Clustering

The hierarchical clustering algorithm is used combine with other clustering algorithms. For example, Thanh Thi Nguyen(Nguyen et al., 2020) used the hierarchical method and DBSCAN to find the relationship between COVID-19, and bat CoVs ,and pangolin CoVs by clustering the genus of the COVID-19 virus and the bat and pangolin virus genus sequences. The researcher used 334 sequences of genus that were clustered to 12 groups among 12 major virus class. The hierarchical methods confirmed that the cut-off(C) parameter is an important variable that worked as a threshold and it is useful to change the value of C during the experiment. Compared to this method, DBSCAN needs to identify the core point and the minimum number of the neighbor's parameters need to be defined too. The results showed that SARS-CoV-2 belongs to the Batcoronavirus genus depending on the set 2 and set 8 of both algorithms.

Layth Rafea(Rafea et al., 2021) proposed a tool for clustering articles about COVID-19 making use of the title of the article and using the hierarchical method and K-means. He used the Hierarchical Agglomerative used to produce good clustering and performance with several parameters like n-cluster=10, affinity ='euclidean ', and linkage 'ward'. On the other hand, the K-means was applied to obtain the labels necessary for starting supervised classification techniques with the parameter clusters K=10, n_jobs =4 and verbose =10. The results of the clustering showed that Hierarchical total of 10 clusters and the K-means used to label the dataset where clusters K=10, n_jobs =4 and verbose =10. In the next steps, he used supervised classification methods as shown in table (2), and after that, to reduce the dimensionality of the dataset used t-SNET.

Zarikas (Zarikas et al., 2020) They depend on the different COVID-19 time-series to cluster the countries depending on different time series. Also, the study taking the populations of the different countries in considerations. The purpose of this study is to cluster the countries according to the spreads of the virus depending on the time series and populations of the countries by using hierarchical algorithm which is highly recommended to use it with the time series. If the researcher used other clustering algorithms with time series the results will no be accuracy as hierarchical.

Other Clustering Techniques

Other clustering algorithms are used like (Wang et al., 2015) (Kotsiantis & Pintelas, 2004). They missioned it as model base methods like SOM (Kotsiantis & Pintelas, 2004) Wei (W. T. Li et al., 2020) used the SOM clustering algorithm to diagnose models of COVID-19 in order to distinguish between the COVID-19 patients and influenza patients depending on the clinical datasets and the public COVID-19 datasets. The input to the SOM methods was N*P matrix, where N=398 patients and P= 48 variables, and the output of the SOM was only 27 variables where the value of P very significant (P<0.001). The researcher found that COVID-19 patients can be grouped into subtypes depending on the gender, registered symptoms, and the levels of immune cell, and the researcher achieved 97.9% separating COVID-19 from the influenza patients

Koffka (Khan & Ramsahai, 2020) compared nine clustering algorithms (k-Means, DBSCAN, SC , AC,GM, Brich, mean-shift, optics) by using the tweeters datasets (the link to the datasets available in table (2)). It contains1086 cases and (19) features. The results show that mean-shift is the best clustering method in the first steps in which the supervised learning is used to test the model. All methods preformed a good classification except for (Mean-shift, BRICH). For more details about the datasets, see table 2.

Table 2: Summarization of Literature Survey

No.	Author	Year	Objective of the study	Cluster type	Datasets	Results
1	(Carrillo-Larco & Castillo-Cara, 2020)	2020	cluster countries in groups with shared profiles of the COVID-19	K- means	https://doi.org/10.6084/m9.figshare.12030363.v1 https://creativecommons.org/licenses/by/4.0/legalcode	155 countries, five to six clustering, confirm COVID-19 cases (p<0.001)

2	(Gupta et al., 2021)	2020	Measuring twitters users whether the weather is factor of distribution of Covid -19	K-means K=25	Tweets from Countries across Full Dataset and Relevant Tweets Set	166.005 tweets collaborative Identify 28.555 40.4% uncertainty about weather 33.5 % no effect 26.1% effect
3	(Imtyaz et al., 2020)	2020	Measure government assessment to control the pandemic	k-Means	https://github.com/CSSEGISandData/COVID-19 testing data https://ourworldindata.org/grapher/number-of-covid-19-tests-per-confirmed-case	K=5 Countries in Western Europe showed the highest humanity rates. Countries in South Asia and the Middle East showed the lowest mortality rate (controlling for all other variables)
4	(Zhao et al., 2020)	2020	To analyze the immune systems and the blood of the hospitalized patients with COVID-	k-means/ agglomerate	seventy-one sufferers changed into accompanied up with weekly blood assessments from clinic	The chemokine RANTES (CCL5) turned into drastically elevated, from an early

			19. Dependent on these three cytokines (IL-1RA and IL-10 and RANTES)		admission	degree of the infection, in sufferers with slight however now no longer extreme disease. TNF-α and GM-CSF confirmed no vast variations among extreme and slight cases
5	(Siddiqui et al., 2020)	2020	Show relationship between temperature and diverse cases situation (suspected, confirmed, and death cases)	k-means	https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports https://www.worldometers.info/coronavirus/	found that temperature is not always main factor for the spread of COVID-19 pandemic, while exploring the impact of temperature in suspected, confirmed, lack of existence case
6	(Ta et al., 2020)	2020	mining hidden subjects from quotations in Al Jazeera's	affinity propagation/ mean shift	extracted from Al Jazeera news articles containing keywords	Different group of clustering depending on the following sentences

			information articles, associated with the COVID-19 pandemic		("coronavirus", "COVID-19", etc.) related to the COVID-19 pandemic.	((People, countries, Africa, coronavirus, community, South Africa, EU, COVID-19deal)
7	(Nikolopoulos et al., 2020)	2020	help policymakers and planners make better decisions during the ongoing and future pandemics	Curve Nearest Neighbor methods (CPC–NN)	https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases https://github.com/bigscity/nCov-predict https://www.worldlifeexpectancy.com/cause-of-death/lung-disease/by-country/	the overall performance of PC–NN/CPC–NN. The outcomes are higher than the Naïve method, with 8 fashions outperforming as a result.
8	(W. T. Li et al., 2020)	2020	to predict COVID-19 presence using only commonly recorded clinical variables	SOM	https://github.com/yoshihiko1218/COVID19ML/projects dataset with clinical variables for each patient	correlational analyses on this dataset and determined that person adult males with COVID-19 have higher serum neutrophil and leukocyte ranges

						than women with COVID-19.
9	(Khan & Ramsahai, 2020)	2020	Helped separate large amount of Twitter data to quickly find key points in the data before going into further classification.	k-Means DBSCAN SC AC GM Birch Mean Shift OPTICS	collected tweets matching hashtags linked to COVID-19 from a Kaggle dataset The dataset has 1086 cases with nineteen (19) features	k-Means 100.000 % DBSCAN 99.412 % SC 100.000 % AC 100.000 % GM 100.000 % Birch 99.804 % Mean Shift 74.063 % OPTICS 99.524 %
10	(Karadayi et al., 2020)	2020	development on unsupervised anomaly detection performance provides beneficial perception to suppress the resurgence of coved19 outbreaks	DBSCAN/ Conv LSTM	public Italian COVID-19 time series dataset(21 regions)(2 autonomous provinces) http://dati.istat.it/Index.aspx?lang=en&SubSessionId=50f87960-20d5-44f2-b405-5fe16f91da73	MinPtsLOF = 20, MinPtsLDBCAN = 30, LOFUB = 5, pct = 0.3 intensive care of COVID-19
11	(Prakash et al., 2020)	2020	To broaden a COVID-19 chance stratified version that classifies human beings into one-of-a-	K-Modes	https://arxiv.org/abs/1807.01514v1	20% symptomatic instances that had been categorized into cluster A

			kind chance cohorts, primarily based totally on their signs and validate the same			Cluster B had human beings without a symptom C had human beings with signs for the infection 92% categorized correctly
12	(Diversified Community Services, n.d.)	2020	intelligent optimization method to develop diversified TCM prevention programs for COVID-19	Fuzzy FCM	diversified TCM prevention programs developed Zhejiang Provincial Health Commission.	(1) the use of fitness big-information analytics to beautify clustering (2) incorporating extra TCM know-how to lessen the efforts of TCM experts; (3) combining clinical know-how with gadget getting to know to assess the outcomes of TCM
13	(Chandu, 2020)	2020	Find the relationship between the high public health expenditure (%GDP)	K- means	https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200506covid-	Grouped 89 countries into Class1 size 54 Class 2 size 35

			and COVID-19		19-sitrep-107.pdf?sfvrsn=159c3dc_2	
14	(Vadyala et al., 2020)	2020	tackle the issue of variance and precision in predicting the number of reported cases in the traditional SEIR model	, K-Means, long short-term memory (LSTM)	https://github.com/CSSEGISandData/COVID-19 https://www.louisiana-demographics.com/	K-Means-LSTM has a higher accuracy with an RMSE of 601.20 in which due to the fact the SEIR model with an RMSE of 3615.83.
15	(Shuai Yong, Jiang Chunxu, Yuan Can , Su Xinyi, 2020)	2020	to investigate the effect of countrywide epidemic rules at the modern-day epidemic results	K-Means Agglomerative Clustering (DBSCAN)	National strategy data	K-Mean clustering optimal solutions
16	(Iskandar Fitri, Refly Asmar, 2020)	2020	algorithm developed based on characterization from geo-location of the country	DBSCAN K-Means Elbow Method	WHO (World Health Organization)and worldometers.info/coronavirus/	DBSCAN is more relevant with this pandemic case where the distance or density from each points. K-Means the complexity is faster.

17	(Oniani et al., 2020)	2020	construct a computable co-prevalence network embeddings to help affiliation detection among COVID-19 associated biomedical entities	DBSCAN	https://github.com/shenfc/COVID-19-network-embeddings https://www.davidoniani.com/covid-19-network	unsupervised learning, sixty three clusters had been formed with silhouette rating of 0.128
18	(Lucas et al., 2020)	2020	Analysis longitudinal immunological correlates of disease	K -means	https://doi.org/10.1038/s41586-020-2588-y	cluster 1 n = 46 moderate cluster 2 = 50 cluster 3 = 16
19	(Nguyen et al., 2020)	2020	explores the COVID-19 virus source by using raw genomic orders of the virus	hierarchical clustering algorithm / DBSCAN	GISAID database (https://www.gisaid.org) accession numbers EPI_ISL_410538 - EPI_ISL_410543	Clustering of all 334 sequences of genus Hierarchical: C=7 Clustering =5 DBSCAN =3 cluster E=o.55
20	(Kamath, 2020)	2020	visualization of highly infected top 50 countries	K-means	https://www.who.int/docs/default-source/coronaviruse/situation-reports	50 counters Clustering =5 Class1=China and Italy And so on for the other counters
21	(Rafea et al., 2021)	2021	recognize the responses of nations to the pandemic	hierarchical clustering / k-means	https://pages.semanticscholar.org/coronavirus-research	No of Hierarchical clusters (=10), No of K

			through analyzing posted articles which are at once associated with COVID-19			means clusters =10,
22	(Mahmoudi et al., 2020)	2020	The spread of COVID-19 in different countries depending on the population of the countries	Fuzzy	WHO Datasets	There is high correlation between the size of the population and dead cases and confirms cases
23	(Zarikas et al., 2020)(Rahman et al., 2020)	2020	Clustering the countries depending on the time-series of the different countries.	Hierarchical	https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 final data can be found in Mendeley http://dx.doi.org/10.17632/kg72dst75p.1	The results of clustering different COVID-19 time series using other clustering algorithms shows fails and not correct
24	(Spurlock & Elgazzar, 2020)	2020	Clustering the social media like tweeter according to the keywords related to the COVID-19	K-means/ agglomerative	The dataset of tweeter (Tweepy Library) For (1000 users)	There are 15 groups classification depending on the keys related to the COVID-19.

25	(Kurniawan et al., 2020)	2020	Clustering the spread of the virus in different countries	K-means	COVID-19 Outbreak data collection	There are 5 groups of infected countries. And the accuracy of about 97%
26	(Doroshenko, 2020)	2020	Clustering the spread of the virus in Italy	K-means/ Hierarchal	https://github.com/pcm-dpc/COVID-19	The both algorithms did a good clustering. the K-means clustering depending on the geographical location. But hierarchical clustering gives high ratio of 97% for 2 classes
27	(Wolfe et al., 2020)	2020	To cluster the articles related to the virus from the first datasets, and cluster the human mobility and influence on the virus from the datasets available in MTI (Maryland Transportation Institute)	K-means	1- the COVID-19 dataset 2- data from the MTI (Maryland Transportation Institute).	The most important features get from the first datasets are (smoking, age, and asthma) For the second dataset the most important feature extracts are (spread of the virus, social distance,

						and testing capacity)
28	(Boluwade, 2020)	2020	To find the relationship between the quality of the air due to nitrogen dioxide (NO2), particulate matter (PM2.5) from one side and COVID-19 from the other side the study implemented in Africa	self-organizing map (SOM)	https://covid19.who.int/ https://ourworldindata.org/burden-of-disease	separated the African continent into five clusters based on air quality indicators (NO2, PM2.5), and COVID-19 . It is interesting to note that these five clusters did not follow the five geographical boundaries in Africa.

DISCUSSION

From the previous section we noticed that clustering techniques have been used widely to inform about COVID-19. The methods used can be categorized to hierarchical, partitioning, and density. As for the partitioning cluster techniques K-means has been used widely in different aspects. Almost all the researchers applied these methods because they are easy to be implemented. However, one of the drawbacks of this methods is that it does not handle noisy data. The other partition clustering methods used is Fuzzy-c mean. It is used when the datasets overlap and it gives good results. Added to that, another kind of k-means like ELM K-means is best suited among all methods as it finds best quality clusters and in less computational time. Concerning the hierarchical category of clustering technique, we can see that the agglomerative technique is used but the divisive is not used. The hierarchical has been used widely with the natural language processing, and genus sequence clustering. Other hierarchical clustering techniques like (ACA-

DTRS, FACA, and DTRS) have not been used. Also, these algorithms may be providing better results, and give a perfect number of clustering.

CONCLUSIONS

COVID-19 was a major problem that marked 2019 and 2020. This has led many intellectuals and scientists to think creatively of ways to eradicate its negative effect. Accordingly, many people have been referring to this pandemic on social media and other media outlets. In this paper, we have presented various clustering techniques which are now used for analyzing the COVID-19 pandemic related to different variables. It useful to know the impact of these variables on the COVID-19, and how the results of the clustering will helps to controlling on the spread of the virus , and supporting the heath unites to take the right decisions. All these recent methods are compared on the basis of purpose of the usage, the accuracy of the algorithms, and application fields. We found that not all these techniques are used and the application of new clustering methods in the future in different fields is necessary. Another point of importance is that researchers need to test the model and the clustering approach by using T test or F test.

REFERENCE

Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Machine learning supervised algorithms of gene selection: A review. Technology Reports of Kansai University, 62(3), 233–244.

Arora, S., & Chana, I. (2014). A survey of clustering techniques for big data analysis. Proceedings of the 5th International Conference on Confluence 2014: The Next Generation Information Technology Summit, 59–65. https://doi.org/10.1109/CONFLUENCE.2014.6949256

Bargarai, F. A. M., Abdulazeez, A. M., Tiryaki, V. M., & Zeebaree, D. Q. (2020). Management of wireless communication systems using artificial intelligence-based software defined radio. International Journal of Interactive Mobile Technologies, 14(13), 107–133. https://doi.org/10.3991/ijim.v14i13.14211

Ben Ayed, A., Ben Halima, M., & Alimi, A. M. (2014). Survey on clustering methods: Towards fuzzy clustering for big data. 6th International Conference on Soft Computing and Pattern Recognition, SoCPaR 2014, 331–336. https://doi.org/10.1109/SOCPAR.2014.7008028

Boluwade, A. (2020). Regionalizing Partitioning Africa's Coronavirus (COVID-19) Fatalities Using Environmental Factors and Underlying Health Conditions for Social-economic Impacts. 2nd Novel Intelligent and Leading Emerging Sciences Conference, NILES 2020, 439–443. https://doi.org/10.1109/NILES50944.2020.9257875

Carrillo-Larco, R. M., & Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. Wellcome Open Research, 5, 1–22. https://doi.org/10.12688/wellcomeopenres.15819.3

Chandu, V. (2020). Identification of spatial variations in COVID-19 epidemiological data using K-Means clustering algorithm: a global perspective. https://doi.org/10.1101/2020.06.03.20121194

Diversified Community Services. (n.d.). Diversified Community Services. October 2020, 62–73.

Dodd, R. (2014). by.

Doroshenko, A. (2020). Analysis of the distribution of COVID-19 in Italy using clustering algorithms. Proceedings of the 2020 IEEE 3rd International Conference on Data Stream Mining and Processing, DSMP 2020, 325–328. https://doi.org/10.1109/DSMP47368.2020.9204202

Duan, L., Xu, L., Guo, F., Lee, J., & Yan, B. (2007). A local-density based spatial clustering algorithm with noise. Information Systems, 32(7), 978–986. https://doi.org/10.1016/j.is.2006.10.006

Garcia, A. (2020). Clustering of Longitudinal data: Application to COVID-19 data.

Godinho, G. G., França, F. D. O., Freitas, J. M. A., Santos, F. M. L., Prandini, A., Godinho, A. C., & Costa, R. P. D. P. (2015). Resultado do tratamento cirúrgico artroscópico das rerrupturas do manguito rotador do ombro. Revista Brasileira de Ortopedia, 50(1), 89–93. https://doi.org/10.1016/j.rbo.2014.03.007

Gupta, M., Bansal, A., Jain, B., Rochelle, J., Oak, A., & Jalali, M. S. (2021). Whether the weather will help us weather the COVID-19 pandemic: Using machine learning to measure twitter users' perceptions. International Journal of Medical Informatics, 145(November 2020), 104340. https://doi.org/10.1016/j.ijmedinf.2020.104340

Hu, Z., Ge, Q., Li, S., Jin, L., & Xiong, M. (2020). Artificial intelligence forecasting of covid-19 in China. ArXiv, 1–20.

Imtyaz, A., Abid Haleem, & Javaid, M. (2020). Analysing governmental response to the COVID-19 pandemic. Journal of Oral Biology and Craniofacial Research, 10(4), 504–513. https://doi.org/10.1016/j.jobcr.2020.08.005

Iskandar Fitri, Refly Asmar, A. R. (2020). Data clustering Mapping of Global Covid 19 pandemic Based on Geo -location. Jurnal Mantik, 4(2), 511–520.

Jahangir, H., Tayarani, H., Sadeghi Gougheri, S., Aliakbar Golkar, M., Ahmadian, A., & Elkamel, A. (2020). Deep Learning-based Forecasting Approach in Smart Grids with Micro-Clustering and Bi-directional LSTM Network. IEEE Transactions on Industrial Electronics, 0046(c), 1–1. https://doi.org/10.1109/tie.2020.3009604

Jahwar, A. F. (2021). META-HEURISTIC ALGORITHMS FOR K-MEANS CLUSTERING : A REVIEW. 17(7), 1–20.

Kamath, R. S. (2020). COVID-19 Country Cluster Analysis : Machine Learning Approach. September. https://www.researchgate.net/publication/344047248

Karadayi, Y., Aydin, M. N., & Öğrenci, A. S. (2020). Unsupervised anomaly detection in multivariate spatio-temporal data using deep learning: Early detection of covid-19 outbreak in Italy. IEEE Access, 8, 164155–164177. https://doi.org/10.1109/ACCESS.2020.3022366

Khan, K., & Ramsahai, E. (2020). Categorizing 2019-n-CoV Twitter Hashtag Data by Clustering. International Journal of Artificial Intelligence & Applications, 11(4), 41–52. https://doi.org/10.5121/ijaia.2020.11404

Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks,

37, 52–65. https://doi.org/10.1016/j.neunet.2012.09.018

Kotsiantis, S. B., & Pintelas, P. E. (2004). Recent Advances in Clustering: A Brief Survey. WSEAS Transactions on Information Science and Applications, 1(1), 73–81.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall – runoff modelling using Long Short-Term Memory (LSTM) networks. 6005–6022.

Krishnachandran, V. N. (2007). Lecture Notes in Machine learing. In Integration The Vlsi Journal. Vidya Centre for Artificial Intelligence Research Vidya Academy of Science & Technology Thrissur - 680501. http://oro.open.ac.uk/6442/

Kurniawan, R., Abdullah, S. N. H. S., Lestari, F., Nazri, M. Z. A., Mujahidin, A., & Adnan, N. (2020). Clustering and Correlation Methods for Predicting Coronavirus COVID-19 Risk Analysis in Pandemic Countries. 2020 8th International Conference on Cyber and IT Service Management, CITSM 2020, 6–10. https://doi.org/10.1109/CITSM50537.2020.9268920

Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant Applications of Machine Learning for COVID-19 Pandemic. Journal of Industrial Integration and Management, 05(04), 453–479. https://doi.org/10.1142/s2424862220500268

Li, W. T., Ma, J., Shende, N., Castaneda, G., Chakladar, J., Tsai, J. C., Apostol, L., Honda, C. O., Xu, J., Wong, L. M., Zhang, T., Lee, A., Gnanasekar, A., Honda, T. K., Kuo, S. Z., Yu, M. A., Chang, E. Y., Rajasekaran, M. R., & Ongkeko, W. M. (2020). Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis. BMC Medical Informatics and Decision Making, 20(1), 1–13. https://doi.org/10.1186/s12911-020-01266-z

Li, Y., Tang, S., & Sa, V. R. De. (n.d.). Supervised Spike Sorting Using Deep Convolutional Siamese Network and Hierarchical Clustering. 1–10.

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine, 27(2), 1–6. https://doi.org/10.1093/jtm/taaa021

Lucas, C., Wong, P., Klein, J., Castro, T. B. R., Silva, J., Sundaram, M., Ellingson, M. K., Mao, T., Oh, J. E., Israelow, B., Takahashi, T., Tokuyama, M., Lu, P., Venkataraman, A., Park, A., Mohanty, S., Wang, H., Wyllie, A. L., Vogels, C. B. F., … Iwasaki, A. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature, 584(7821), 463–469. https://doi.org/10.1038/s41586-020-2588-y

Mahesh Kumar, K., & Rama Mohan Reddy, A. (2016). A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. Pattern Recognition, 58, 39–48. https://doi.org/10.1016/j.patcog.2016.03.008

Mahmoudi, M. R., Baleanu, D., Mansor, Z., Tuan, B. A., & Pho, K. H. (2020). Fuzzy clustering method to compare the spread rate of Covid-19 in the high risks countries. Chaos, Solitons and Fractals, 140, 1–9. https://doi.org/10.1016/j.chaos.2020.110230

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., & Long, J. (2018). A Survey of Clustering with Deep Learning: From the Perspective of Network Architecture. IEEE Access, 6(c), 39501–39514. https://doi.org/10.1109/ACCESS.2018.2855437

Mobasher, B. (2007). Data mining for Web personalization. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4321 LNCS, 90–135. https://doi.org/10.1007/978-3-540-72079-9_3

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. Computer Journal, 26(4), 354–359. https://doi.org/10.1093/comjnl/26.4.354

Najim Adeen, I. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Systematic review of unsupervised genomic clustering algorithms techniques for high dimensional datasets. Technology Reports of Kansai University, 62(3), 355–374.

Nath, V., & Levinson, S. E. (2014). Machine learning. In SpringerBriefs in Computer Science (Issue 9783319056050). https://doi.org/10.1007/978-3-319-05606-7_6

Nguyen, T. T., Abdelrazek, M., Nguyen, D. T., Aryal, S., Nguyen, D. T., & Khatami, A. (2020). Origin of Novel Coronavirus (COVID-19): A Computational Biology Study using Artificial Intelligence. BioRxiv. https://doi.org/10.1101/2020.05.12.091397

Nietto, P. R., & Do Carmo Nicoletti, M. (2017). Case studies in divisive hierarchical clustering. International Journal of Innovative Computing and Applications, 8(2), 102–112. https://doi.org/10.1504/IJICA.2017.084893

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2020). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. European Journal of Operational Research, xxxx. https://doi.org/10.1016/j.ejor.2020.08.001

Oheneba-Sakyi, Y. (1992). Determinants Of Current Contraceptive Use Among Ghanaian Women At The Highest Risk Of Pregnancy. Journal of Biosocial Science, 24(4), 463–475. https://doi.org/10.1017/S0021932000020022

Omran, M. G. H., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods. Intelligent Data Analysis, 11(6), 583–605. https://doi.org/10.3233/ida-2007-11602

Oniani, D., Jiang, G., Liu, H., & Shen, F. (2020). Constructing co-occurrence network embeddings to assist association extraction for COVID-19 and other coronavirus infectious diseases. Journal of the American Medical Informatics Association, 27(8), 1259–1267. https://doi.org/10.1093/jamia/ocaa117

Pampalk, E. (2001). Limitations of the SOM and the GTM. 2001, 1–11. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.4673

Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. (2020). A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. Chaos, Solitons and Fractals, 140, 110190. https://doi.org/10.1016/j.chaos.2020.110190

Pham, D. T., & Afify, A. A. (2007). Clustering techniques and their applications in engineering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 221(11), 1445–1459. https://doi.org/10.1243/09544062JMES508

Prakash, A., Muthya, S., Arokiaswamy, T. P., & Nair, R. (2020). Using Machine Learning to assess Covid-19 risks. MedRxiv. https://doi.org/10.1101/2020.06.23.20137950

Qader Zeebaree, D., Mohsin Abdulazeez, A., Asaad Zebari, D., Haron, H., & Nuzly Abdull Hamed, H. (2021). Multi-Level Fusion in Ultrasound for Cancer Detection based on Uniform LBP Features. Computers, Materials & Continua, 66(3), 3363–3382. https://doi.org/10.32604/cmc.2021.013314

Rafea, L., Ahmed, A., & Abdullah, W. D. (2021). Classification of a COVID-19 dataset by using labels created from clustering algorithms. Indonesian Journal of Electrical Engineering and Computer Science, 21(2502–4752), 164–173. https://doi.org/10.11591/ijeecs.v21.i1.pp164-173

Rahman, M. A., Zaman, N., Asyhari, A. T., Al-Turjman, F., Alam Bhuiyan, M. Z., & Zolkipli, M. F. (2020). Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices. Sustainable Cities and Society, 62, 102372. https://doi.org/10.1016/j.scs.2020.102372

Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. Data Mining and Knowledge Discovery Handbook. https://doi.org/10.1007/978-0-387-09823-4

Rokach, L., & Maimon, O. (2006). Clustering Methods. Data Mining and Knowledge Discovery Handbook, 321–352. https://doi.org/10.1007/0-387-25465-x_15

Rossi, R., Socci, V., Talevi, D., Mensi, S., Niolu, C., Pacitti, F., Di Marco, A., Rossi, A., Siracusano, A., & Di Lorenzo, G. (2020). COVID-19 Pandemic and Lockdown Measures Impact on Mental Health Among the General Population in Italy. Frontiers in Psychiatry, 11(August), 7–12. https://doi.org/10.3389/fpsyt.2020.00790

Shen, H., Fu, M., Pan, H., Yu, Z., & Chen, Y. (2020). The Impact of the COVID-19 Pandemic on Firm Performance. Emerging Markets Finance and Trade, 56(10), 2213–2230. https://doi.org/10.1080/1540496X.2020.1785863

Shuai Yong, Jiang Chunxu, Yuan Can , Su Xinyi, H. X. (2020). 2020 IEEE 6th International Conference on Control Science and Systems Engineering (ICCSSE) : July 17-19, 2020, Beijing, China. 2–5.

Siddiqui, M. K., Morales-Menendez, R., Gupta, P. K., Iqbal, H. M. N., Hussain, F., Khatoon, K., & Ahmad, S. (2020). Correlation between temperature and COVID-19 (suspected, confirmed and death) cases based on machine learning analysis. Journal of Pure and Applied Microbiology, 14(April), 1017–1024. https://doi.org/10.22207/JPAM.14.SPL1.40

Spurlock, K., & Elgazzar, H. (2020). Predicting COVID-19 Infection Groups using Social Networks and Machine Learning Algorithms. 2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile

Communication Conference, UEMCON 2020, 0245–0251. https://doi.org/10.1109/UEMCON51285.2020.9298093

Ta, T. H., Rahman, A. B. S., Sidorov, G., & Gelbukh, A. (2020). Mining Hidden Topics from Newspaper Quotations: The COVID-19 Pandemic. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12469 LNAI. Springer International Publishing. https://doi.org/10.1007/978-3-030-60887-3_5

Tilson, L. V., Excell, P. S., & Green, R. J. (1988). A generalisation of the Fuzzy c-Means clustering algorithm. Remote Sensing. Proc. IGARSS '88 Symposium, Edinburgh, 1988. Vol. 3, 10(2), 1783–1784. https://doi.org/10.1109/igarss.1988.569600

Vadyala, S. R., Betgeri, S. N., Sherer, E. A., & Amritphale, A. (2020). Prediction of the number of COVID-19 confirmed Cases based on K-means-LSTM. ArXiv.

Villmann, T. (1999). Benefits and Limits of the Self-Organizing Map and its Variants in the Area of Satellite Remote Sensoring Processing. Proc. of European Symposium on Artificial Neural Networks (ESANN'99), April, 111–116.

Wang, Y., Chen, Q., Kang, C., Zhang, M., Wang, K., & Zhao, Y. (2015). Load profiling and its application to demand response: A review. Tsinghua Science and Technology, 20(2), 117–129. https://doi.org/10.1109/tst.2015.7085625

Wolfe, G., Elnashar, A., Schreiber, W., & Alsmadi, I. (2020). COVID-19 Candidate Treatments, a Data Analytics Approach. 2020 4th International Conference on Multimedia Computing, Networking and Applications, MCNA 2020, 139–146. https://doi.org/10.1109/MCNA50957.2020.9264290

Wyplosz, C. (2020). 14 The good thing about coronavirus, book: Economics in the Time of COVID-19. www.cepr.org

Yazdani, S., Minaee, S., Kafieh, R., Saeedizadeh, N., & Sonka, M. (2020). COVID CT-Net: Predicting Covid-19 From Chest CT Images Using Attentional Convolutional Network. http://arxiv.org/abs/2009.05096

Yu, J., Huang, H., & Tian, S. (2004). Cluster validity and stability of clustering algorithms. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3138(3), 957–965. https://doi.org/10.1007/978-3-540-27868-9_105

Zarikas, V., Poulopoulos, S. G., Gareiou, Z., & Zervas, E. (2020). Clustering analysis of countries using the COVID-19 cases dataset. Data in Brief, 31, 105787. https://doi.org/10.1016/j.dib.2020.105787

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends, 1(2), 56–70. https://doi.org/10.38094/jastt1224

Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018). Gene Selection and Classification of Microarray Data Using Convolutional Neural Network. ICOASE 2018 - International Conference on Advanced Science and Engineering, 145–150.

https://doi.org/10.1109/ICOASE.2018.8548836

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019). Trainable Model Based on New Uniform LBP Feature to Identify the Risk of the Breast Cancer. 2019 International Conference on Advanced Science and Engineering, ICOASE 2019, 106–111. https://doi.org/10.1109/ICOASE.2019.8723827

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zeebaree, S. R. M. (2017). Combination of k-means clustering with genetic algorithm: A review. International Journal of Applied Engineering Research, 12(24), 14238–14245.

Zhang, W., Zhou, T., Lu, Q., Wang, X., Zhu, C., Sun, H., Wang, Z., Lo, S. K., & Wang, F.-Y. (2020). Dynamic Fusion based Federated Learning for COVID-19 Detection. 14(8), 1–9. http://arxiv.org/abs/2009.10401

Zhao, Y., Qin, L., Zhang, P., Li, K., Liang, L., Sun, J., Xu, B., Dai, Y., Li, X., Zhang, C., Peng, Y., Feng, Y., Li, A., Hu, Z., Xiang, H., Ogg, G., Ho, L. P., McMichael, A., Jin, R., … Zhang, Y. (2020). Longitudinal COVID-19 profiling associates IL-1RA and IL-10 with disease severity and RANTES with mild disease. JCI Insight, 5(13), 4–14. https://doi.org/10.1172/jci.insight.139834