Journal of Computer and Communications, 2022, 10, 32-50
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Fighting against COVID-19: Who Failed and Who Succeeded?

Hussein Baalbaki[1], Hassan Harb[1], Ali Jaber[1], Chamseddine Zaki[2], Chady Abou Jaoude[3], Kifah Tout[1], Layla Tannoury[4]

[1]Faculty of Sciences, Lebanese University, Beirut, Lebanon
[2]College of Engineering and Technology, American University of the Middle East, Egaila, Kuwait
[3]Ticket Lab, Faculty of Engineering, Antonine University, Baabda, Lebanon
[4]Faculty of Business Administration and Economics, Lebanese University, Rashaya, Lebanon
Email: hsinbaalbaki@gmail.com, hassan.harb.1@ul.edu.lb, ali.jaber@ul.edu.lb, chamseddine.zaki@aum.edu.kw, chady.aboujaoude@ua.edu.lb, ktout@ul.edu.lb, layla.tannoury@ul.edu.lb

How to cite this paper: Baalbaki, H., Harb, H., Jaber, A., Zaki, C., Jaoude, C.A., Tout, K. and Tannoury, L. (2022) Fighting against COVID-19: Who Failed and Who Succeeded? *Journal of Computer and Communications*, **10**, 32-50.
https://doi.org/10.4236/jcc.2022.104004

Received: January 2, 2022
Accepted: April 12, 2022
Published: April 15, 2022

Abstract

Recently, governments and public authorities in most countries had to face the outbreak of COVID-19 by adopting a set of policies. Consequently, some countries have succeeded in minimizing the number of confirmed cases while the outbreak in other countries has led to their healthcare systems breakdown. In this work, we introduce an efficient framework called COMAP (COrona MAP), aiming to study and predict the behavior of COVID-19 based on deep learning techniques. COMAP consists of two stages: clustering and prediction. The first stage proposes a new algorithm called Co-means, allowing to group countries having similar behavior of COVID-19 into clusters. The second stage predicts the outbreak's growth by introducing two adopted versions of LSTM and Prophet applied at country and continent scales. The simulations conducted on the data collected by WHO demonstrated the efficiency of COMAP in terms of returning accurate clustering and predictions.

Keywords

COVID-19, Data Clustering and Prediction, Co-Means, ANOVA, LSTM, Prophet

1. Introduction

Despite the tremendous human efforts and the massive expenditures, the health care systems at both local and international levels are still suffering from a severe gap between the provided services and the real needs. This can be clearly shown with the appearance of the COVID-19 that affected more than 22 million people

and caused the death of 777 thousand until August 20, 2020. Against such a pandemic, governments were unable to defend their citizens. Thus, their rescue plans consisted of administering lockdown and awareness campaigns about respecting social distancing, sanitizing, washing hands, and wearing protective masks. The healthcare ecosystem had no sufficient staff, hospitals, and beds to face this ongoing tragedy, where some countries lost control of this epidemic. Hence, researchers have focused on studying the behavior of COVID-19 in different points of view [1] [2] [3]. It has already been widely recognized that data science—tools and techniques—can potentially have a vital role in tracking the behavior of COVID-19 in terms of the geographic distribution of the outbreak and the evolution of its hot spots.

On the one hand, some authors such as [4] [5] [6] [7] propose data clustering and analysis techniques for COVID-19. For instance, the authors of [4] propose a data analysis method aiming to find a correlation between the temperature condition and the cases' situations for different regions of China. The proposed method uses the K-means algorithm in order to show the trend effects of temperature on each region. The results showed that the temperature condition could not be the only significant factor for the outbreak of COVID-19. In [5], the authors introduce a clustering model based on the generalized K-means to group the state-level data sets for the spread of the epidemic in the United States. To select the optimal clusters number, the authors use two selectors, e.g., Akaike information criterion and Bayesian information criterion, and they prove that the nations are optimally grouped into six clusters regarding the outbreak of the pandemic. In [6], the authors propose a data analysis framework based on the Python language that allows cleaning up, classifying, and visualizing the COVID-19 data. Mainly, the proposed framework uses a data exploratory approach, in particular clustering and bivariate analysis. This approach verified that the rate of deaths is linearly proportional to the percentage of elderly persons in most countries. On the other hand, the authors of [8] [9] [10] [11] are dedicated to predicting the outbreak growth of COVID-19. For instance, the authors of [8] propose a hybrid artificial intelligence technique for predicting the COVID-19 spread called ISI, improved susceptible-infected. Mainly, ISI uses natural language processing combined with the long short-term memory to forecast the infection rates caused by the virus's propagation. According to the proposed technique, the authors prove that the patients touched by COVID-19 had a higher contamination rate after eight days of infection. In [9], the authors use four traditional prediction models (linear regression, exponential smoothing, support vector machine, and LASSO) to forecast the spreading growth of COVID-19. By applying such techniques on the confirmed cases, death, and recovery rates, the authors show that the exponential smoothing gives the best forecasting accuracy while the support vector machine gives the worst one. The authors of [10] propose a model that benefits from the relationships between the nearest countries to estimate the progress of COVID-19. The model takes as input several metrics such as the number of cas-

es, population per 1 M people, ARIMA parameters, and the polynomial function, then it finds, as output, a correlation between the coronavirus spread and the population in each country.

Unfortunately, despite the signficant number of proposed techniques, the virus's outbreak continues to grow without any clear understating of the current situation and its future progress. In this paper, we tackle a new trend in studying the spread of COVID-19 by proposing an accurate framework called COMAP consisting of two stages: clustering and prediction. Unlike the most clustering techniques, the first stage does not require the initial number of clusters, but it dynamically finds the optimal one according to the virus's spread across the countries. On the other hand, we adopted two well-known deep learning techniques, e.g., LSTM and Prophet, applied in several domains to the COVID-19 case to raise the precision of its outbreak.

The rest of this paper has the following structure. In Section 2, we describe the problem according to the WHO logging. Sections 3 and 4 detail the clustering and prediction stages proposed in our framework. Section 5 exposes the simulation results and discusses the performance of the COMAP framework. Lastly, the paper is concluded in Section 6.

2. COVID-19 Logging and Problem Formulation

One of WHO's main missions consists of monitoring the global health system and collecting the data related to dangerous diseases and viruses. Mainly, with the emergence of COVID-19 in January 2020, WHO started to gather all information related to the outbreak of the pandemic and the number of infected persons, to closely supervise its progress and severity. Daily statistics were published by each country and reported to the WHO, which, in turn, categorizes the data with attributes describing the number of daily confirmed, death, and recovered cases. Therefore, we mathematically formulate the analysis problem of COVID-19 as follows: given the set of countries $\mathcal{C}=\{C_1, C_2, \cdots, C_\alpha\}$, where α is the total number of countries recorded in the WHO table. Each country $C_i \in \mathcal{C}$ has its own set of records for each attribute collected starting from the emergence of the virus until the date of this paper submission, e.g. $A_i=\{r_1, r_2, \cdots, r_\beta\}$; β is the total number of records collected for an attribute and r_k is the number of cases detected during the k^{th} day. Hence, the record sets of all countries according to a specific attribute is defined as: $\mathcal{A}=\{A_1, A_2, \cdots, A_\alpha\}$. Therefore, our objective is to study and analyze the data sets in $\mathcal{A}$ using clustering and prediction approaches in order to understand the behavior of COVID-19 in the world countries.

3. COVID-19 Clustering Stage

In machine learning, clustering is a set of unsupervised algorithms used to classify unlabeled data into clusters. These algorithms are integrated into various kinds of applications [12] [13] [14]. One of the most well-known clustering al-

gorithms used in many data science and machine learning classes is K-means [15]. However, K-means faces two main challenges: first, the selection of the cluster number (K) that is a crucial decision because it determines the accuracy of the obtained clusters; second, the convergence criterion function that can highly affect the number of iterations thus, increasing the computation process. Therefore, to conquer these difficulties, we propose a new version of K-means, called Corona-based K-means or Co-means, that adapts the analysis of variance (ANOVA) with the Bartlett test to the traditional K-means algorithm. Co-means has an objective to group the countries having similar COVID-19 spread into the same cluster. This allows the WHO to study the severity of COVID-19 inside each cluster and verify the government procedures' seriousness, efficiency, and people's adherence. In the next sections, we first recall the ANOVA model and Bartlett test, followed by the description of the new clustering algorithm called Co-means. Then, we introduce a new metric that allows us to assign the risk level of COVID-19 for the countries inside each cluster.

3.1. ANOVA and Bartlett Test

In statistics, the analysis of variance (ANOVA) is one of the most effective ways to determine whether the data sets belong to the same population and have the same variance, e.g., the null hypothesis, or not. Hence, it comes the statistical test's role by specifying a threshold for the significance level of the variance between data sets. In our work, we focus on the Bartlett test [16] that does not assume any constraint about the normality of data which works based on the following two steps.

3.1.1. Variance Computation

The first step aims to calculate the variance, V, between the COVID-19 data sets for all countries according to the Bartlett equation [16]:

$$V = \frac{(N-\alpha)\ln\left(\sigma_p^2\right) - \sum_{k=1}^{\alpha}(n_k - 1)\ln\left(\sigma_k^2\right)}{1 + \frac{1}{3(\alpha-1)}\left(\sum_{k=1}^{\alpha}\left(\frac{1}{n_k - 1}\right) - \frac{1}{N-\alpha}\right)} \quad (1)$$

where:

- α indicates the countries' number
- n_k indicates the records' number collected for the country C_k
- σ_k^2 is the record variance of the country C_k
- N is the total number of records collected for all countries and it is calculated as follows:

$$N = \sum_{k=1}^{\alpha} n_k$$

- σ_p^2 is the pooled variance that can be calculated as:

$$\sigma_p^2 = \frac{1}{N-\alpha}\sum_{k=1}^{\alpha}\sigma_\alpha^2$$

3.1.2. Significance Decision

In order to check its significance, the Bartlett test compares the value of V to a defined threshold $V_{\alpha-1,\gamma}$ calculated according to the Chi-square table. Subsequently, $V_{\alpha-1,\gamma}$ indicates the critical value of the variance corresponding to the degree of freedom $\alpha-1$ with significance level γ. Therefore, the decision is based on the following:

- if $V > V_{\alpha-1,\gamma}$ then the null hypothesis is rejected and the variance between the sets is significant.
- if $V \leq V_{\alpha-1,\gamma}$ then the variance between the sets is not significant and they have similar variances.

3.2. Corona-Based K-Means Algorithm: Co-Means

This section aims to integrate the ANOVA and Bartlett test to the traditional K-means in order to produce a more accurate clustering algorithm, e.g. Co-means. Indeed, Co-means is two-fold: first, it dynamically finds the optimal number of clusters without using the trial-and-error method adapted in most existing algorithms. Second, it uses a new convergence criterion to increase the accuracy of the obtained clusters. Basically, Co-means assumes that COVID-19 spreads similarly in all countries, thus, they are assigned to the same cluster, e.g., the initial cluster. Then, it recursively divides the initial cluster into small clusters every time a significant variance inside the new cluster is detected. The process of cluster division is stops when the null hypothesis is accepted in all the obtained clusters, which are used as a criterion function in our algorithm. Algorithm 0.0.2 describes the process of Co-means that takes, as input, the data sets of the daily confirmed cases for all countries, e.g., $\mathcal{A}$, and finds the clusters of countries having the same COVID-19 spread, e.g., $\mathcal{L}$. First, all countries are considered similar in terms of the outbreak, and they are assigned to the set of clusters T (lines 2-7). Then, for each cluster in T, we calculate the variance between the data sets based on the Bartlett formula, e.g., equation 1, as well as the Bartlett threshold according to the Chi-square table (lines 9-10). Consequently, if the calculated variance is less than the threshold, then Co-means considers that the countries have the same virus spread, so they are added to the final list of clusters $\mathcal{L}$ (lines 11-13). Otherwise, when the variance is greater than the threshold, the countries have different speeds of spread; so, we divide the current cluster into two subclusters by applying the K-means algorithm (line 15). This iterative process continues until no more clusters on T thus, obtaining the final clusters of countries with similar COVID-19 behavior.

3.3. Risk Degree Metric

After forming all clusters, we propose a clustering analysis method that allows studying the spread of COVID-19 of countries at the same cluster and evaluates its risk. Hence, we introduce a new metric called risk degree referred to as R_i that measures the impact of COVID-19 of each cluster L_i. Subsequently, R_i is calculated based on three parameters:

Algorithm 3.2 Co-means Algorithm.

Require: Set of countries: $\mathcal{C} = \{C_1, C_2, \ldots, C_\alpha\}$, Data sets of countries: $\mathcal{A} = \{A_1, A_2, \ldots, A_\alpha\}$, Significance level: γ.
Ensure: Clusters of countries: $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$.

$\mathcal{L} \leftarrow \emptyset$
$T \leftarrow \emptyset$ // temporary list of clusters
$L_1 \leftarrow \emptyset$
for each set $A_j \in \mathcal{A}$ **do**
  $L_1 \leftarrow L_1 \cup \{A_j\}$
end for
$T \leftarrow T \cup \{L_1\}$
repeat
  compute V for L_i based on equation 1
  $V_{\alpha-1,\gamma}$ = Chi-square $(|L_i| - 1, \gamma)$
  if $V \leq V_{\alpha-1,\gamma}$ **then**
    $\mathcal{L} \leftarrow \mathcal{L} \cup \{L_i\}$
    remove L_i from T
  else
    $T \leftarrow T \cup$ K-means $(L_i, 2)$
  end if
until no cluster $L_i \in T$
return $\mathcal{L}$

• The percentage of confirmed cases in each cluster, CC_i: this parameter aims to evaluate the effectiveness of the procedures taken by the countries of a cluster, and it is calculated as follows:

$$CC_i = \frac{\sum_{k=1}^{|L_i|} \text{confirmed cases of country } C_k}{\sum_{k=1}^{|L_i|} \text{population number of country } C_k}$$

• The percentage of death cases in each cluster, DC_i. This parameter allows us to validate the efficiency of the healthcare systems of countries in L_i in responding to the massive number of patients infected by the virus. Thus, the more the countries can hospitalize the patients and reduce the number of deaths, the more their system's efficiency is. DC_i is calculated as follows:

$$DC_i = \frac{\sum_{k=1}^{|L_i|} \text{number of deaths of country } C_k}{\sum_{k=1}^{|L_i|} \text{confirmed cases of country } C_k}$$

• The rapid spread of COVID-19 outbreak of each cluster, RR_i. This parameter indicates the progress of the COVID-19 outbreak in the cluster L_i. Thus, the more the epidemic spreads in a country, the more the severity of its impact will be. RR_i is considered as the mean variance of all countries in a cluster and calculated as follows:

$$RR_i = \frac{\sum_{k=1}^{|L_i|} \sigma_k^2}{|L_i|}$$

Based on the above equations, the risk degree R_i of a cluster L_i is calcu-

lated according to the following formula:

$$R_i = \left(CC_i\right)^2 \times \left(DC_i\right)^2 \times RR_i \tag{2}$$

Therefore, we assess the risk level of the COVID-19 according to the value of the risk degree, which mostly takes a value between 0 and 1. Subsequently, the risk assessment is shown in Table 1 based on three defined thresholds, e.g., ε_i where $i \in [1,3]$ and $\varepsilon_1 < \varepsilon_2 < \varepsilon_3$. Indeed, the thresholds are configured by the experts or the WHO, and their values can change from period to another depending on the progress of the epidemic. Then, each cluster L_i is assigned a risk level and description according to its risk degree R_i. For instance, the countries of a cluster are considered in a risk level 1 with a safe situation if its risk degree equal to 0 while countries inside a cluster with a risk degree more than ε_3 are considered in high critical situation with risk level of 5.

4. COVID-19 Prediction Stage

In the second stage of our framework, we aim to study and analyze the spread of COVID-19 from the prediction point of view. While the clustering offers an overview of the countries affected by the pandemic with its current spread, the prediction approach allows governments and WHO to estimate its future progress. Hence, estimating the number of infected persons in the near future can help the relevant authorities in two directions; on the one hand, they can estimate the number of patients to be hospitalized, thus the need for medical equipment such as PCR, ventilators, masks, and gloves, required to withstand the corona outbreak. The prediction also helps determine the needed medical staff and resources to serve the estimated number of patients, such as nurses, physicians, hospitals, and medical centers. On the other hand, the authorities can determine the hot zones most affected by the epidemic, thus, they decide about the precautions, procedures, and essential policies to reduce the virus's spread. This may include the isolation of some geographical regions, closing of crowded places, restrictions on people's movements, and preparation of quarantine places for infected people. This paper introduces two prediction methods that can efficiently estimate the epidemic outbreak on the scales of countries and continents: Long Short-Term Memory (LSTM) and Prophet.

Table 1. Evaluation of COVID-19 impact based on risk degree metric.

Risk level	Risk degree	Risk description
1	$R_i = 0$	safe
2	$R_i \le \varepsilon_1$	low
3	$\varepsilon_1 < R_i \le \varepsilon_2$	medium
4	$\varepsilon_2 < R_i \le \varepsilon_3$	serious
5	$R_i > \varepsilon_3$	high

4.1. Recall of LSTM Method

Long Short-Term Memory (LSTM) [17] is an enhanced variant of repetitive neural network that deals with complicated cases related to processing, prediction and classification of time series. Typically, the architecture of the LSTM is based on the concept of cell state where data in the cell state can be added, removed, or reset using structures called gates. Subsequently, we distinguish between three gates (e.g. input, output and forget) for each cell state (**Figure 1**). As shown, every LSTM cell takes three variables as inputs: the input of the current time x_t, the previous cell state u_{t-1} and the hidden state h_{t-1}. Furthermore, a cell state consists of several neural layers where each one defines three variables: 1) $\mathcal{B}$ or the blocks' number, that indicates the neuron's capacity; 2) $\mathcal{T}$, or the time steps' number, that indicates the size of the input vector (x_t) used in the prediction of the next time step (y_t); 3) the features' number ($\mathcal{F}$) that indexes the time step.

4.2. Adapting LSTM to COVID-19

The rapid spread of the virus is highly dependent on the procedures set by each country. Hence, the consecutive numbers of daily infected persons are mostly correlated, making LSTM one of the ideal methods to forecast the future progress of coronavirus based on the previous collected data. Indeed, our objective is to predict the outbreak using LSTM at country-based and continent-based levels. At the local level, e.g. country-based, the goal of prediction is to help the governments avoid the worst scenario and take the precautions procedures before losing control. At the global level, e.g. continent-based, the prediction can help the WHO understand the effect of neighboring countries on the pandemic's spread; this will help provide global standards and measures such as limiting travel between countries, and airports closing, to reduce the outbreak among countries.

In order to apply LSTM, we divide the data records of each country (respectively continent) into two parts: training and testing. Given the record set $A_i = \{r_1, r_2, \cdots, r_\beta\}$ of a country C_i then: $\theta\%$ of the records are taken for the training of LSTM method while the remaining $(\beta - \theta)\%$ records are used in testing the precision of the obtained results. Obviously, the more the percentage of the training part is, the more the LSTM accuracy will be. Therefore, we adopt the LSTM to forecast the outbreak of COVID-19 according to the following steps:

1) LSTM configuration: in this step, we adapt all the parameters of LSTM before training the data model. Indeed, the configuration includes two phases: data normalization and assigning the cell state parameters. On the one hand, data must be normalized to scale up their values into the same range [0, 1]. This is an essential task when studying the COVID-19 data due to the difference between the population numbers of countries, which highly affects the number of infected persons. For instance, the United States' confirmed cases are much greater

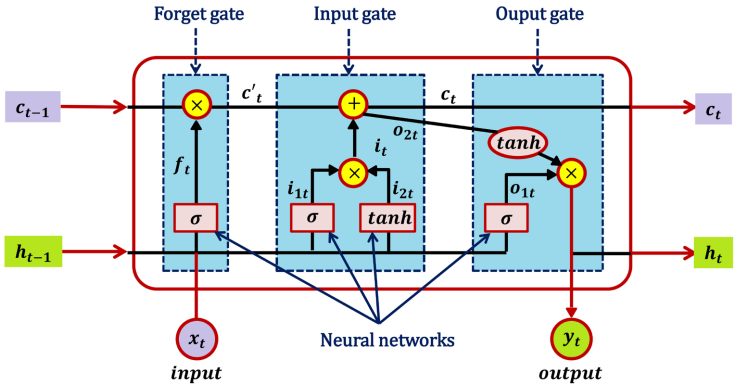


Figure 1. Architecture of LSTM cell.

than those in Spain but both countries may have an approximate spread level. In this work, we normalized the COVID-19 records of all countries using the MinMaxScaler algorithm before training and predicting the outbreak progress. Therefore, each record $r_j \in A_i$ is normalized according to Min-Max scaling based on the following equation:

$$r'_j = \frac{r_j - \mathcal{A}_{\min}}{\mathcal{A}_{\max} - \mathcal{A}_{\min}} \tag{3}$$

where $\mathcal{A}_{\min}$ and $\mathcal{A}_{\max}$ represent the lowest and highest record values in $\mathcal{A}$, respectively. On the other hand, this step also includes the assignment of the cell parameter values, e.g., $\mathcal{B}$, $\mathcal{S}$ and $\mathcal{F}$. Indeed, in our simulation, we assign several values for each parameter in order to study its impact, except for $\mathcal{F}$ that is fixed to 1 since we only predict the confirmed cases feature.

2) LSTM training: in fact, studying and analyzing the previous data is an essential step to estimate its future behavior. Hence, it comes the importance of the training phase in LSTM aims to understand the records' trends for each country and the correlations existing between the successive ones. Indeed, to train the data, we need to set up two concepts in LSTM: the error function and the optimization algorithm. Typically, the error function is used to estimate the model's loss, while the optimization algorithm aims to reduce the current state's error using a recursive process. In this work, the Mean Square Error (MSR) is set as the error function for our model and we used Adam optimizer as the optimization algorithm [18].

3) LSTM prediction: after training the data based on the first θ records of each country, the last step aims to predict the future progress of the outbreak for the next days, weeks or months. Subsequently, the prediction process of LSTM takes into account the last $\mathcal{T}$ trained records to estimate the future ones.

4.3. Adapting Prophet to COVID-19

Prophet [19] is an open source forecasting tool built by Facebook's Core Data Science team in 2008 and implemented in Python and R languages. Its core is composed of an additive model describing daily, weekly, and yearly seasonality,

while consider holiday effects. Prophet can handle the missing values and outliers through the correlation discovered in the given historical data. Indeed, the Prophet method's accuracy increases if a strong seasonal effect in the time series is detected, or the historical data's increasing size. Upon its release, Prophet found its way into several domains but it mostly shines in the business where it is flexible to fit various business problem scenarios.

To apply the Prophet algorithm over COVID-19 datasets, we adopt the same scenario of the LSTM method, where each country and continent's records are divided into training and testing parts. Then, in order to predict the future of the epidemic, the Prophet combines the main trends components in the following regression equation:

$$p(t) = d(t) + y(t) + w(t) + h(t) \tag{4}$$

where:

- $d(t)$: indicates the piecewise linear trend used by Prophet to select the change points in the data to detect changes in trends automatically.
- $y(t)$: indicates the yearly seasonal trend that calculates, based on Fourier series, constant and predictable variations for one year.
- $w(t)$: indicates the weekly seasonal trend that calculates, based on dummy variables, constant and predictable variations for one week.
- $h(t)$: is the list of important holidays or events determined by the user. In our simulation, if a daily statistics of a country is not reported to the WHO, then it is considered as a missed or holiday data.

5. Results and Discussion

The performance of our technique, COMAP, is evaluated in terms of understanding the behavior of the COVID-19 outbreak. We implemented COMAP based on the Python language under the Anaconda framework, and we used the worldwide data collected by WHO, as described in Section 2. In our simulation, the performance of both clustering and prediction stages are tested based on the data of all countries. However, in the prediction stage, we only show the results of three countries (USA, Italy and Iran) scattered on various continents and they are highly affected by the COVID-19.

Table 2 summarizes the parameter description and configuration adapted in our simulations.

Table 2. Simulation configuration.

Parameter	Description	Values
γ	significance level of Bartlett	0.1, 0.05, 0.01
θ	percentage of training data	60, 70, 80
$\mathcal{B}$	number of blocks	fixed to 50
$\mathcal{T}$	number of time steps	fixed to 15
$\mathcal{F}$	number of features	fixed to 1 (confirmed cases)

5.1. Performance Evaluation of Clustering Stage

This section studies the performance of the Co-means algorithm according to the following metrics.

5.1.1. Countries' Distribution over Clusters

Indeed, one of the most advantages of the Co-means algorithm is its ability to dynamically find the optimal number of clusters based on the Bartlett threshold criterion. Figure 2 shows the number of obtained clusters along with their sizes, e.g., the number of countries inside each one, when varying the significance level values. The obtained results show that cluster formation is highly affected by the variation of γ allowing more understanding of the COVID-19 outbreak. We mainly observe that most of the obtained clusters have a size less than 10 while few exceed this value.

Furthermore, the following observations are eminent:

• The number of clusters K increases with the increasing of the value of γ. For instance, K increases from 34 to 46 when γ varied from 0.01 to 0.1. This is because the Bartlett test becomes more flexible regarding the variance between records when the significance level value increases.

• The average number of countries per cluster decreases with the increasing of γ. Subsequently, each cluster contains an average of 6.2, 5.1 and 4.5 countries when γ changed to 0.01, 0.05 and 0.1 respectively. This will decrease the variance between countries' records in each cluster and enhance the understanding of the outbreak in such countries.

5.1.2. Cluster Analysis According to Risk Degree Metric

In Figure 3, we show the severity of the epidemic outbreak worldwide based on the risk degree metric calculated for each cluster, after applying Co-means. As we can see, COVID-19 has almost attacked all countries but with different impact levels. Apart from the Antarctica continent and according to R_i metric, we observe the following: 56.5% of countries are at the low level, 26.7% are at the medium level, 6.2% are seriously affected by the epidemic, and 10.5% are of high

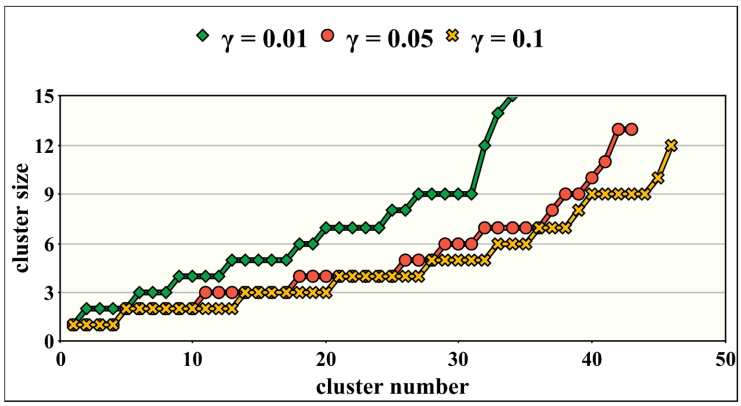


Figure 2. Countries' distribution over clusters in the function of the significance level.

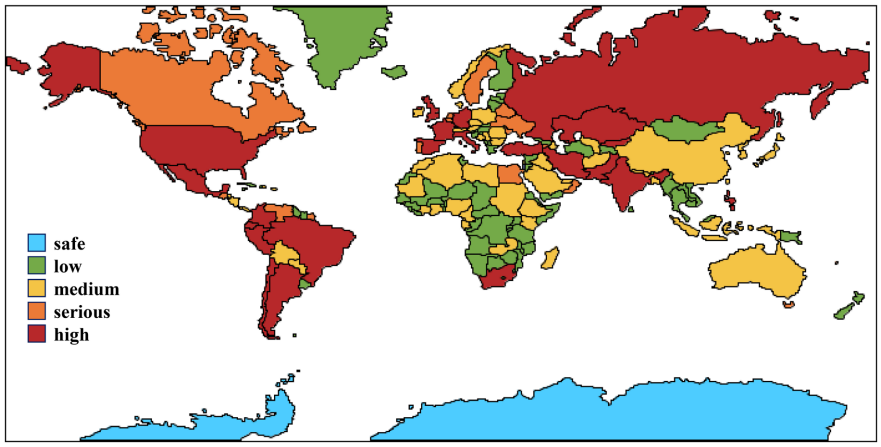


Figure 3. COVID-19 severity based on the risk degree metric.

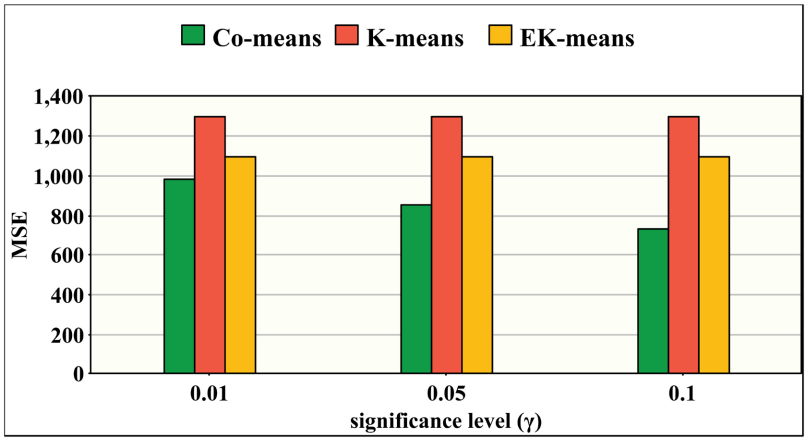


Figure 4. Comparison of clustering accuracy.

outbreak level. The results also confirm our mechanism's behavior by assigning the same R_i value for countries having similar COVID-19 spread. For instance, we can see that the USA, Spain, Italy, Brazil, and Iran are assigned to the same cluster with the same risk level and the COVID-19 are quickly spread in such countries. Otherwise, we can notice that the number of confirmed cases is slowly increased in countries such as Finland, Mongolia, Niger, Mali, Thailand, and others. Furthermore, we can also notice that the countries in America (respectively in Africa) are more (respectively less) affected by COVID-19 than those in other continents.

5.1.3. Clustering Accuracy

An important metric to asses any clustering algorithm is by calculating the accuracy. A good clustering accuracy is obtained when the difference within clusters is minimized and maximized between clusters. This work focuses on the mean square error (MSE) as a well-known and most used metric to calculate the clustering accuracy. **Figure 4** shows the average MSE of all clusters by applying traditional K-means, Co-means and the Enhanced K-means (EK-means) proposed in [4]. The obtained results show that Co-means outperforms K-means

and EK-means in terms of reducing the clustering error in all cases. Subsequently, Co-means optimizes the clustering accuracy up to 77% and 49% comparing to K-means and EK-means, respectively. We can also observe that Co-means ensures a high level of similarity within the clusters, which strongly confirms the behavior of our mechanism by grouping countries having similar COVID-19 spread in the same cluster. Furthermore, we notice that the accuracy of Co-means increases with the increase of γ value because of the decreasing average countries per cluster (see **Figure 2**).

5.2. Performance Evaluation of Prediction Stage

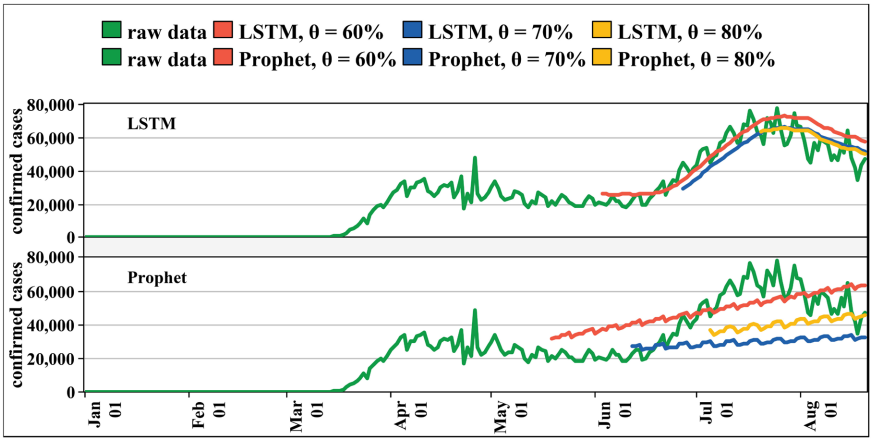
This section studies the performance of both prediction techniques (LSTM and Prophet) according to the following metrics.

5.2.1. Prediction Accuracy at Country Level

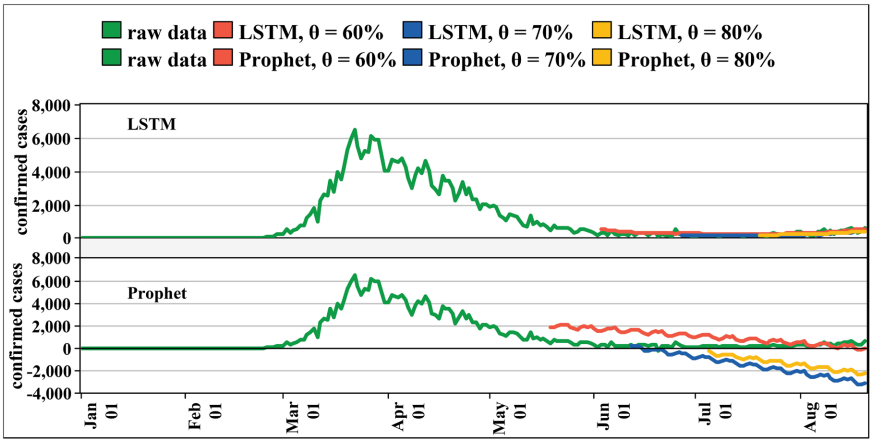
The prediction approach plays a vital role in combating COVID-19 due to its ability to provide an approximate view about the future number of infected persons and determine the procedures needed to reduce the number of deaths. In **Figure 5**, we show LSTM and Prophet's accuracy in our mechanism to predict the confirmed cases in three different countries while varying the training data percentage. The results show a common behavior of the COVID-19 in all countries; at the beginning of its emergence, the epidemic exponentially spreads reaching a critical point, then it decreases for a certain amount of time. This indicates that most countries, including the mentioned ones, did not seriously face the outbreak of this virus, and they did not implement efficient policies and procedures to limit its spread. In addition, the people's commitment in such countries was insignificant, which led to quickly increasing the number of deaths and leading to healthcare system's failure in some.

According to **Figure 5**, we also observe that LSTM gives more accurate results than those of Prophet for different countries and θ values. This is due to two reasons; first, the future values of confirmed cases in LSTM are estimated based on the last $\mathcal{T}$ registered records, while the Prophet method takes into account the trend variation from the beginning of the set of records. Second, LSTM uses a back propagation approach in each record prediction during the training stage to tune the inputs' weights while Prophet is highly based on seasonality and yearly conditions that are not yet existing in COVID-19 data. Furthermore, the results show the following observations:

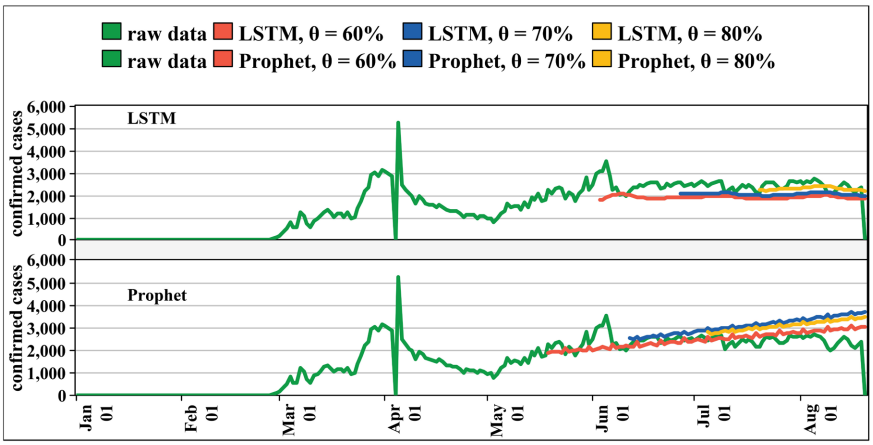
- The accuracy of both methods increases with the increasing of the percentage of the training data. This is because each method will further analyze and understand the variation existing in the data records leading to increasing the accuracy of the predicted data.
- Each method's accuracy can differ from one country to another depending on the obtained records' variation. For instance, the accuracy of LSTM in the case of Italy is greater than those obtained in other countries, while the most accurate results of Prophet are noticed in the case of Iran.



(a) [USA]



(b) [Italy]



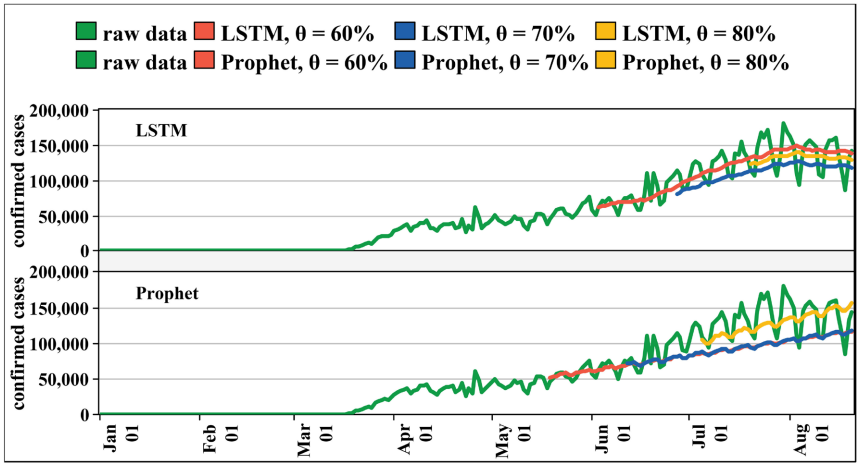
(c) [Iran]

Figure 5. Prediction accuracy of LSTM and Prophet methods at the country level.

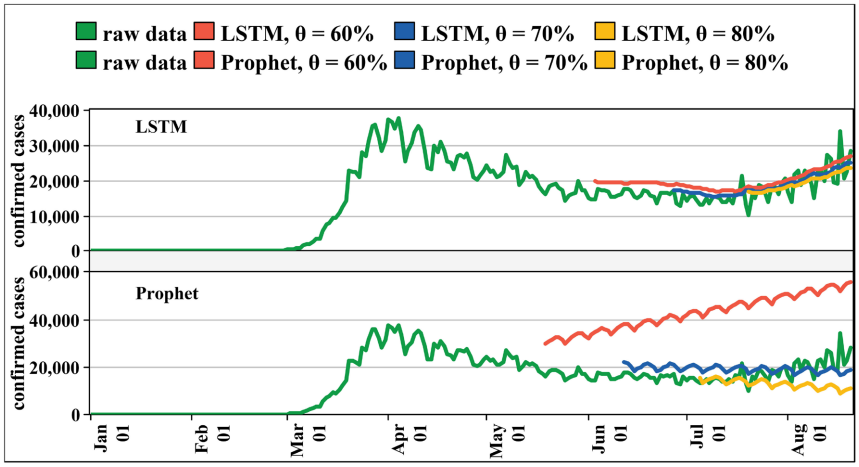
5.2.2. Prediction Accuracy at Continent Level

Sometimes, the outbreak of any disease will not only affect the people of a country but it may propagate to its neighbors. Hence, monitoring and predicting the spread of the COVID-19 at the global level is considered as one of the missions of WHO. In **Figure 6**, we show the accuracy of LSTM and Prophet methods in

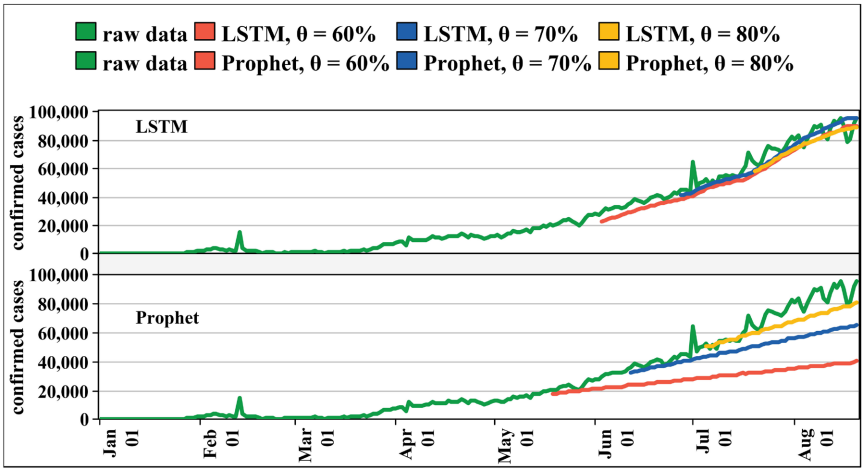
predicting the number of confirmed cases in each of the main continents, e.g., America, Europe, Asia and Africa. Similar to the results obtained at the country level, LSTM shows more prediction accuracy compared to the Prophet methods for various continents and values of the significance level. Obviously, we can



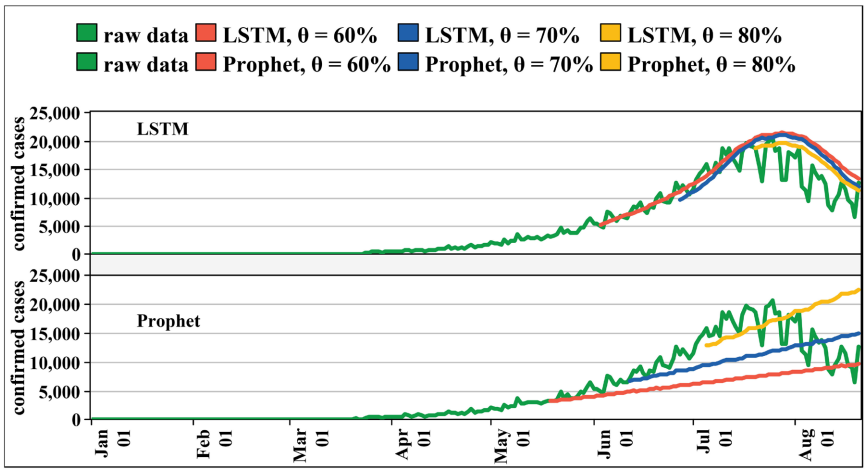
(a) [America]



(b) [Europe]



(c) [Asia]



(d) [Africa]

Figure 6. Prediction accuracy of LSTM and Prophet methods at the continent level.

observe that America is largely suffering from the outbreak of COVID-19 while Africa indicates the minimum number of confirmed cases. Furthermore, the results show that the curves of America and Asia are increasing linearly, while those of other continents are linearly increasing followed by a linear decrease. This may result in strong procedures, including the shut down of most airports, and limiting the travel between countries, adopted recently by the European and African Unions.

Based on the results of Figure 6, the following observations are also eminent:

• The accuracy of each method increases with the increasing size of the training data (for the same reasons described in Figure 5).

• The accuracy of the Prophet method increases with the increasing size of the raw data. For instance, the predicted data of America and Asia is more accurate than those of Europe and Africa because the population numbers of the first ones are much more than those of the last ones. The main reason for that is the significant variation in the high size of data compared to that existing in small data sets. Thus, the accuracy of the Prophet will increase with the increase of the variation between data.

• The LSTM method will maintain a high accuracy level, independent of the data size. This is because LSTM is dependent on the previous data and not the whole ones.

5.3. More Evaluations

The objective of this section is to discuss our proposed mechanism more, while discussing the performance of both prediction and clustering stages under different conditions and circumstances.

From the data prediction point of view, LSTM and Prophet give accurate results for the outbreak of the epidemic at both country and continent levels. However, LSTM ensures more prediction accuracy independent of the variation existing among the data or its size. While, the accuracy of Prophet becomes more

important when increasing the size of data or the size of the training data set.

From the data clustering point of view, Co-means ensures a high level of clustering accuracy, and it efficiently assigns countries having similar COVID-19 spread to the same cluster. Then, by evaluating the epidemic progress in each cluster using risk degree metric, we can conclude that the healthcare systems in most countries failed to combat the COVID-19 where the procedures and policies adopted were not sufficient to reduce its impact. Otherwise, few countries succeeded in fighting against the COVID-19 outbreak, and minimize the number of confirmed and death cases. Subsequently, we analyzed the procedures taken by such countries and we found that most of them adapted common strategies and policies that can be summarized as follows: 1) closing of the educational institutions in the early stage of the outbreak and switching to the online learning; 2) closing of the most entertaining places; 3) shut down of airport; 4) avoiding any type of people gathering; 4) maintaining social distancing; 5) constraints on people movements; 6) tracking of infected people through various kind of technologies; 7) obligation of following up the precautions such as mask, and hand sanitizers; 8) minimizing the number of employees in official and private sectors; 9) a hard penalty in case of policies violation.

6. Conclusions

Unfortunately, COVID-19 will not be the last epidemic that will attack our world; thus, new diseases and viruses will continue to appear and spread in the future. Hence, governments and authorities are requested to further invest in the public healthcare while researchers and the community should make more efforts in designing efficient algorithms to analyze the disease's behavior. In this paper, we proposed an efficient framework called COMAP, aiming to track the outbreak of the epidemic and the effectiveness of the procedures and policies settled by worldwide countries. COMAP is composed of two stages: clustering and prediction. The first stage proposed the Co-means algorithm that allows to group countries having similar outbreak behavior in clusters, then a spread speed metric is proposed to evaluate the severity of the epidemic in each cluster. The second stage introduces two advanced deep learning methods, LSTM and Prophet, aiming to predict the future behavior of COVID-19 at country and continent levels. Through simulation on the data collected by WHO, we demonstrated the efficiency of COMAP in terms of returning accurate and realistic clustering and predictions.

Although its high efficiency, COMAP has several limitations that may be enhanced in future work. First, we seek to increase the performance of the Co-means algorithm proposed in the clustering stage. This may happen either by testing another statistical test rather than Bartlett or by finding new criteria that ensure more accurate clustering. Second, we seek to increase the prediction accuracy in the second stage of our mechanism. This may happen either by taking into consideration additional factors, such as geographical information of countries, when

using LSTM method or by testing another predictive model.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Rossman, H., Keshet, A., Shilo, S., Gavrieli, A., Bauman, T., Cohen, O., Shelly, E., Balicer, R., Geiger, B. and Dor, Y. (2020) A Framework for Identifying Regional Outbreak and Spread of COVID-19 from One-Minute Population-Wide Surveys. *Nature Medicine*, **26**, 634-638. https://doi.org/10.1101/2020.03.19.20038844

[2] Nguyen, T.T. (2020) Artificial Intelligence in the Battle against Coronavirus (COVID-19): A Survey and Future Research Directions. *Preprint*, **10**, 1-27. https://doi.org/10.36227/techrxiv.12743933

[3] Kumar, A., Gupta, P.K. and Srivastava, A. (2020) A Review of Modern Technologies for Tackling COVID-19 Pandemic. *Diabetes & Metabolic Syndrome*: *Clinical Research & Reviews*, **14**, 569-573. https://doi.org/10.1016/j.dsx.2020.05.008

[4] Siddiqui, M.K., Morales-Menendez, R., Gupta, P.K., Iqbal, H.M., Hussain, F., Khatoon, K. and 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**, 1017-1024. https://doi.org/10.22207/JPAM.14.SPL1.40

[5] Zhang, T.L. and Lin, G. (2020) Generalized k-Means in GLMs with Applications to the Outbreak of COVID-19 in the United States. *Computational Statistics & Data Analysis*, **159**, 107217. https://doi.org/10.1016/j.csda.2021.107217

[6] Imtyaz, A., Haleem, A. and Javaid, M. (2020) Analysing Governmental Response to the COVID-19 Pandemic. *Journal of Oral Biology and Craniofacial Research*, **10**, 504-513. https://doi.org/10.1016/j.jobcr.2020.08.005

[7] Sethi, M., Pandey, S., Trar, P. and Soni, P. (2020) Sentiment Identification in COVID-19 Specific Tweets. 2020 *International Conference on Electronics and Sustainable Communication Systems* (*ICESC*), Coimbatore, 2-4 July 2020, 509-516. https://doi.org/10.1109/ICESC48915.2020.9155674

[8] Zheng, N.N., Du, S.Y., Wang, J.J., Zhang, H., Cui, W.T., Kang, Z.J., Yang, T., Lou, B., Chi, Y.T. and Long, H. (2020) Predicting COVID-19 in China Using Hybrid AI Model. *IEEE Transactions on Cybernetics*, **50**, 2891-2904. https://doi.org/10.1109/TCYB.2020.2990162

[9] Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B., Aslam, W. and Choi, G.S. (2020) COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access*, **8**, 101489-101499. https://doi.org/10.1109/ACCESS.2020.2997311

[10] Hernandez-Matamoros, A., Fujita, H., Hayashi, T. and Perez-Meana, H. (2020) Forecasting of COVID19 Per Regions Using ARIMA Models and Polynomial Functions. *Applied Soft Computing*, **96**, 106610. https://doi.org/10.1016/j.asoc.2020.106610

[11] Liu, D.B., Clemente, L., Poirier, C., Ding, X.Y., Chinazzi, M., Davis, J.T., Vespignani, A. and Santillana, M. (2020) A Machine Learning Methodology for Real-Time Forecasting of the 2019-2020 COVID-19 Outbreak Using Internet Searches, News Alerts, and Estimates from Mechanistic Models. arXiv preprint arXiv: 2004.04019 .

[12] Landauer, M., Skopik, F., Wurzenberger, M. and Rauber, A. (2020) System Log Clus-

tering Approaches for Cyber Security Applications: A Survey. *Computers & Security*, **92**, 101739. https://doi.org/10.1016/j.cose.2020.101739

[13] Senouci, O., Harous, S. and Aliouat, Z. (2020) Survey on Vehicular Ad Hoc Networks Clustering Algorithms: Overview, Taxonomy, Challenges, and Open Research Issues. *International Journal of Communication Systems*, **33**, e4402. https://doi.org/10.1002/dac.4402

[14] Ghosal, A., Nandy, A., Das, A.K., Goswami, S. and Panday, M. (2020) A Short Review on Different Clustering Techniques and Their Applications. In: Mandal, J. and Bhattacharya, D., Eds., *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing*, Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_9

[15] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

[16] Snedecor, G.W. and Cochran, W.G. (1967) Statistical Methods. The Iowa State University Press, Iowa. https://doi.org/10.1097/00010694-196809000-00018

[17] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[18] Kingma, D.P. and Ba, J.A. (2019) A Method for Stochastic Optimization. arXiv preprint arXiv: 1412.6980.

[19] Facebook (2008) Prophet Tool. https://facebook.github.io/prophet/