International Journal of Mathematics and
Computer Science, **17**(2022), no. 1, 413–424

Machine Learning Techniques and Forecasting Methods for Analyzing and Predicting Covid-19

**Israa Ali Alshabeeb, Ruaa Majeed Azeez,
Wafaa Mohammed Ridha Shakir**

Technical Computer System Department
Babylon Technical Institute
Al-Furat Al-Awsat Technical University
Kufa, Iraq

e-mail: Inb.esr@atu.edu.iq, ruaa.humady@atu.edu.iq, inb.wfa@atu.edu.iq

(Received August 5, 2021, Accepted September 7, 2021)

Abstract

Covid-19 is the name of the coronavirus that initially appeared in Wuhan, China at the end of 2019. Since its appearance, many people over the world have been infected with various illnesses ranging from the common flu, cold, headache, cough and breathing syndromes, which represent the virus's worst and most serious illness because it causes difficulty breathing for the person and can increase the virus's death rate if treatment is not provided immediately. The goal of this project is to evaluate confirmed cases and deaths using machine learning techniques and predicting methodologies. The results show that the *k*-means method can analyze the data and provide accurate results, and that both of the Exponential and Moving methods can predict results. According to the findings, machine learning techniques and forecasting methods performed best when used in tandem.

1 Introduction

A coronavirus is a respiratory virus that may cause a wide range of diseases from the ordinary cold to breathing syndromes such as SARS (Severe Acute

Key words and phrases: Machine Learning, Forecasting Methods, Covid-19, *k*-means clustering.
AMS (MOS) Subject Classifications: 68Q25, 62M10, 60G25, 68W40, 62H30.
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

Respiratory Syndrome) and MERS (Middle East Respiratory Syndrome). It derives its name from the crown-shaped point that develop on its surface. Viruses of this sort are present in a broad variety of animal species such as bats and camels but they may also be found in humans [1]. This infectious disease that appeared for the first time in Wuhan, China led to an outbreak there and caused widespread worries all over the world [2]. The behavior of the virus is comparable to viral pneumonia which can raise the virus's death rate if no quick treatment is provided thus posing the world's largest threat to people's health [3]. Many studies have been done on this new virus since its discovery (for example, [4]). The aim of this study is to examine the evolution of cases for Covid-19 in Singapore and Indonesia and to establish a VARI model for the cases of Covid-19 in these two countries. The study's findings in these two countries revealed a strong positive association for instances of Covid-19 and that the model is accurate enough to predict cases in the future. Milad, Kolo, Aspoukeh, Hamad, and Bailey [5] used mathematical prediction models as well as artificial simulations based on particles. Time series models comprising of the Simple Exponential model, Holt's approach, and Brown's models had been used to forecast the region's future potential rates. The study's findings demonstrated how the illness had spread in the Kurdistan region in Iraq and what the rates were by comparing them to surrounding countries. The model indicated that the risk of second and third waves of infections may be quite high depending on the number of affected persons. In [6], the Decision Tree Algorithm was used to categorize clinical indications based on a data set of clinical indications. The J48 tree fared marginally better than the Hoeffding tree in terms of precision, recall, and accuracy. Meanwhile, based on the tree view data, the Hoeffding Tree looks simpler and has fewer nodes than J48. Sujath, Chatterjee, and Hassanien [7] proposed a model that might be useful for anticipating Covid-19 spread. They used methods of Multi-layer perceptron, linear regression, and Vector auto-regression on the Covid-19 Kaggle data to predict the epidemiological example of the disease and the rate of Covid-19 instances in India and by comparing the results of the forecast values with cases from John Hopkins University data, the MLP method gave better forecasting results than the other methods. While a convolutional neural network was used to evaluate Covid-19, pulmonary radiography pictures and differentiates confirmed infections from non-infected patients [8]. In [9], contacts tracing apps and their various architectures with terms of security, privacy concerns, and policy during Covid-19 were discussed. In [10], a prediction model was proposed that provided accurate forecasts for the dates when New Zealand was capable of limiting Covid-19 spread. The suggested model was used to estimate the dates when other countries will be able to control the spread of Covid-19. In [11], semiotics methods found that Instagram was the most used media during the pandemic of Covid-19 in Indonesia. In order to combat the disease, the Gaussian mixture model and the Fourier decomposition method (FDM) were used [12] to predict Covid-19 future cases based on the number of current cases in several countries. In this paper, the k-means clustering method, Exponential Smoothing, and Moving Average methods are used to analyze the data set for Covid-19. The rest of this paper is organized as follows: the second section illustrates the data and methods. The third section includes some results and discussions. Finally, a conclusion is presented in the fourth section.

2 Data and Methods

This research investigates the quantitative consequences of the pandemic by examining cases and fatalities in four different countries. The Covid-19 data set was obtained from [13] for the period October 2020 to the end of April 2021 for four countries: Iraq, the United States, India, and China. Figure 1 depicts the proposed approach.

2.1 k-means Clustering Algorithm

One of the popular machine learning methods and unsupervised approach is the k-means method which works to group and cluster the data into several groups based on the distance between the objects. Each group would have similar data where those objects that are not similar transferred to a different group. As a result, the aim of the k-means method is to split n objects into k groups, with each object belonging to the cluster with the closest mean value [14]. With k-means clustering, even huge data sets may be grouped. This approach has been utilized in a variety of disciplines, including image processing, data analysis, computer vision, and business. It defines the center value of each cluster as a mean value for each cluster. First, the number of clusters will be determined, followed by the calculation of the mean value for each cluster and the assignment of objects to the cluster with the greatest similarity based on the Euclidean distance between the cluster and the objects. Finally, until a point of convergence is reached, each object will be assigned to a cluster [15], [16]. The objective function for calculating the k-means approach is:

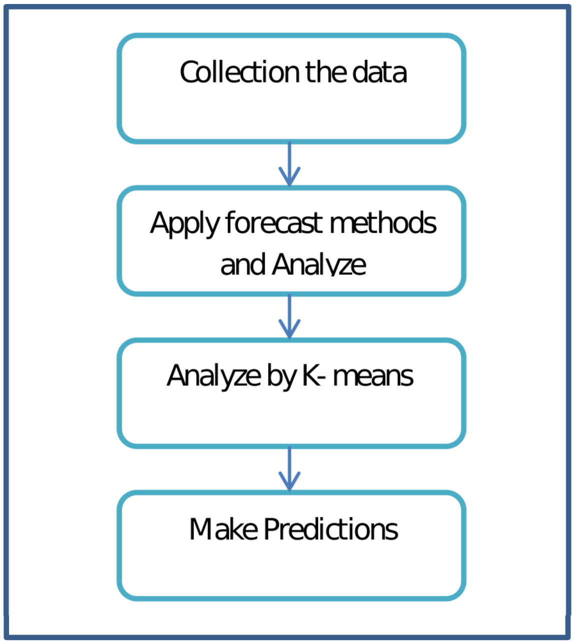

Figure 1: Proposed Method

$$fn = \sum_{i=1}^{k} \sum_{j=1}^{n} (xj - ci)^2 \tag{2.1}$$

The number of clusters is represented by k while the number of cases is represented by n, xj the point of the cluster, and ci the centroid point of the cluster. The centroid point represents the average value of each cluster.

2.2 Exponential Smoothing

Exponential smoothing is one of the time series techniques. In general, there are three forms of exponential smoothing: single exponential smoothing, double exponential smoothing, and triple exponential smoothing. A single type forecasts future value using weighted past values [17]. The Single Exponential Smoothing method's objective function is illustrated below:

$$F_t = \alpha X_t + (1 - \alpha)\, F_{t-1}, \tag{2.2}$$

where F_{t} represents the forecasted value, X_{t} represents the observed value, $Ft_{\text{-}1}$ is the predicted value with Exponential Smoothing at time $t - 1$, α is constant with $0 \leq \alpha \leq 1$.

2.3 Moving Average

Moving Average is a forecasting tool as well as a time series methodology. It is the mean of a set of numbers. Plotting Moving Averages is the easiest method to view them as this make it possible to compute it for any time period. The average value is computed multiple times for many subsets of data. Longer moving averages are preferable for long-term movements while shorter moving averages are preferable for short-term trading [15]. The mathematical expression for Moving Average is represented in Equation (2.3)

$$AV = \frac{(X_1 + X_2 + X_3 + .. + X_n)}{n}, \frac{(X_2 + X_3 + X_4 \ldots + X_{n+1})}{n}, \tag{2.3}$$

where X is the data and n represents the value of periodicity.

3 Result and Discussion

Figures 2-9 depict forecasting methods such as Exponential Smoothing and Moving average approaches estimated during a six-month period from October 1, 2020 to April 30, 2021 for cases and fatalities as well as a scatter plot of the data set. Table 1 displays the clustering results for k-means. The monthly change in cases and fatalities for each of the four nations was taken into account. The graphs were all created with the Xlstat tool and Excel. There are four steps in the simulation of Exponential Smoothing and Moving Average in Figures 2 and 6 for cases and deaths in China for both Exponential Smoothing and Moving Average. The first stage occurred before January 2021 when the number of cases began to rise. The second stage, which has the greatest ratio, happened in January 2021. The third stage occurred when the ratio began to fall after January 2021. The fourth stage came after February 2021 at which point China had control over the virus. The fundamental reproduction number of instances in China was 4100 before the control and 762 after the control. While the number of fatalities fell from 35 to 7 in April. In India, 6957939 individuals were affected in April 2021. The infections were under control, according to the model in figures 3 and 7, until March. In terms of fatalities, the number rose from 2765 in February to 4926 in April. The models in this article indicate that if the pandemic is not maintained under control, it will increase dramatically. According to the simulation in figures 4 and 8 for infection in Iraq, there are three critical time periods. The first was in October 2020 when cases began to decline until January 2021. However, there was a decrease of infection control after January.

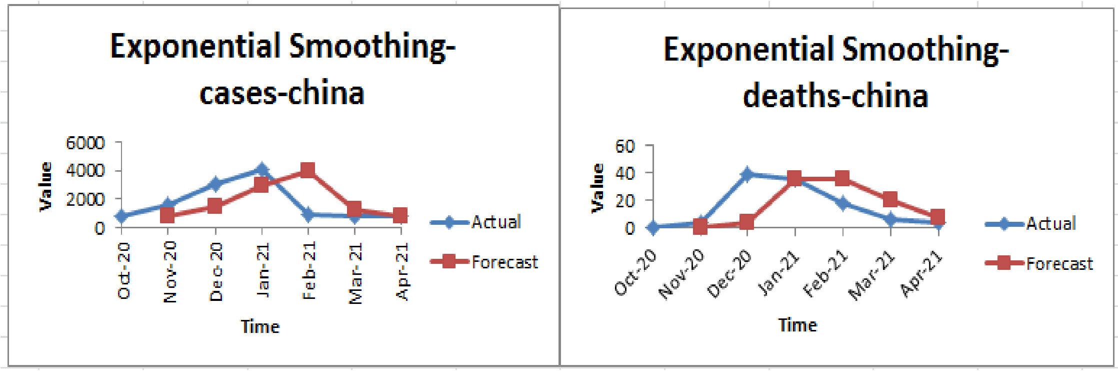

Figure 2: Exponential Smoothing for Cases and Deaths in China

As a result, the infection rate climbed from 24345 in January to 214275 in April, while fatalities jumped from 234 to 1142 in April. The models in this research anticipate an increase in cases until a strategy to decrease infections is discovered. In the United States, 6147690 illnesses were identified at the end of January. The death toll was at 97266. The United States was one among the worst-affected countries in terms of confirmed cases and deaths. Figures 5 and 9 depict the simulation of this finding. The infection rate had dropped to 1884761 cases and 23499 fatalities by the end of April. Table 1 can be explained by k-means capacity to evaluate data. In table 1, N denotes the number of instances whereas S is the size of the cluster for N. Table 1 displays the obtained results which provide a straightforward data analysis. In China, there were 222.000 instances of size 4 infections in January 2021 whereas there were 38.000 infections of size 7. As previously mentioned, this enhances the result of Exponential and Moving Average. The same may be said about India, Iraq, and the United States. According to table 1, the heights rate of infections in India was around 359355.000 with size 9 affected at the end of April while in Iraq the most hazardous number was 8055.385 with the size of 13. According to the results in all figures and table 1, the Exponential, Moving, and k-means methods produce correct results. The Exponential Smoothing model uses an alpha value of 0.9. As indicated in table 1, the clustering number for K was 4. When the alpha value for Exponential approaches one, the outcomes for Exponential and Moving become more comparable than when alpha is close to zero.

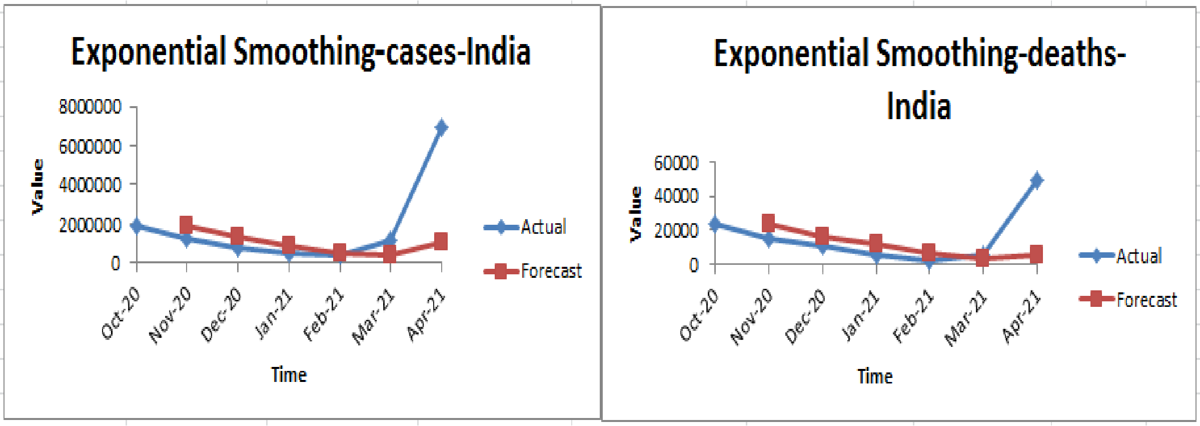

Figure 3: Exponential Smoothing for Cases and Deaths in India

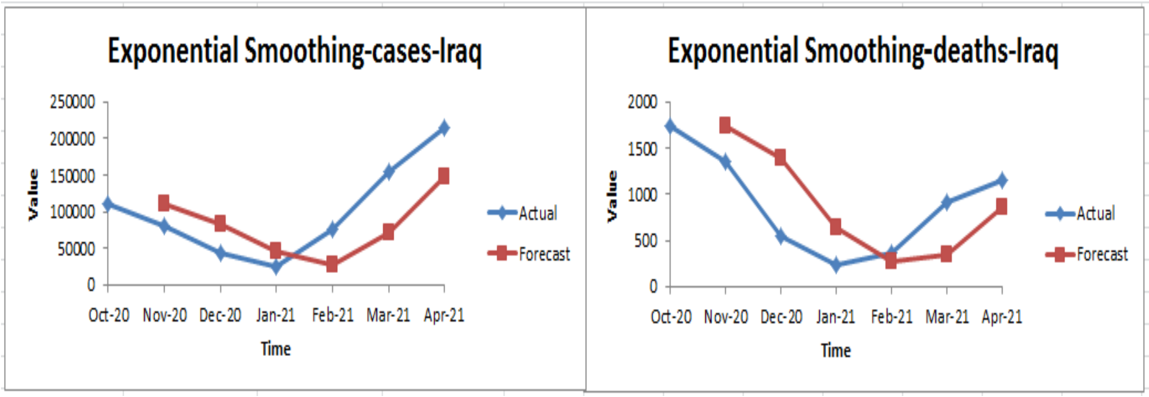

Figure 4: Exponential Smoothing for Cases and Deaths in Iraq

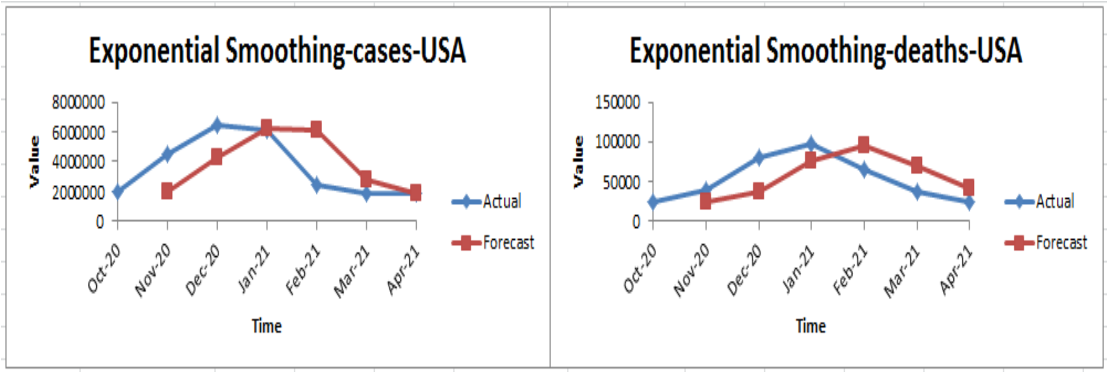

Figure 5: Exponential Smoothing for Cases and Deaths in United States

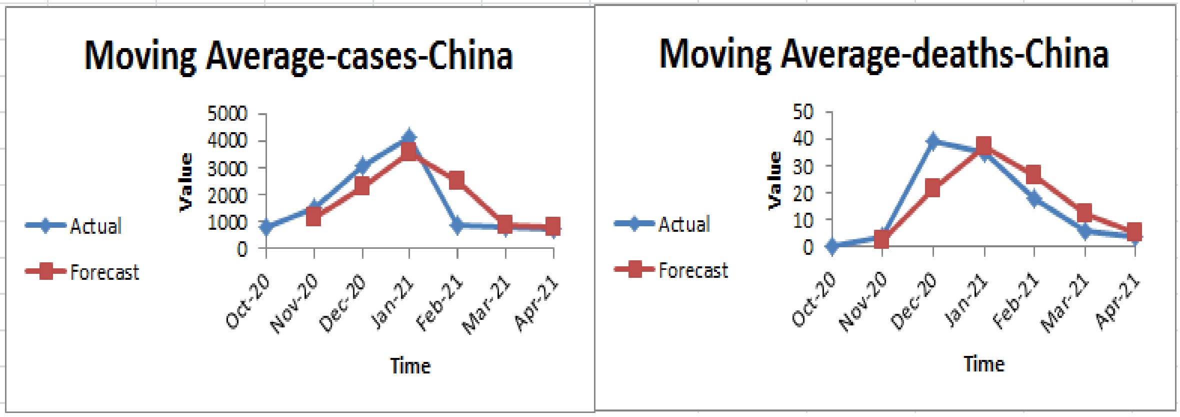

Figure 6: Moving Average for Cases and Deaths in China

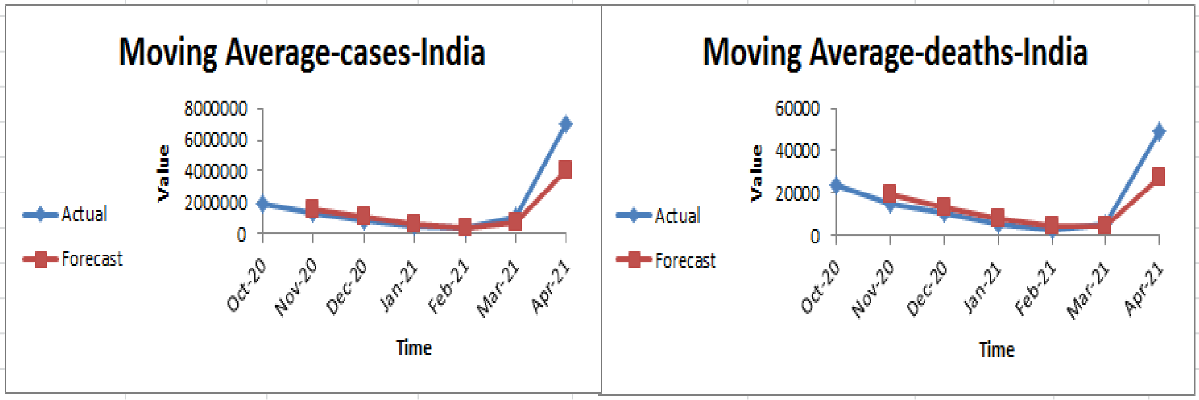

Figure 7: Moving Average for Cases and Deaths in India

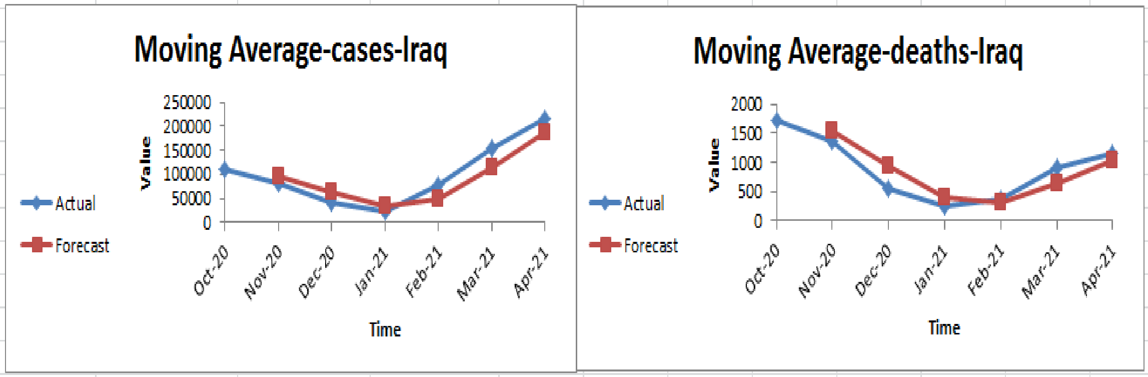

Figure 8: Moving Average for Cases and Deaths in Iraq

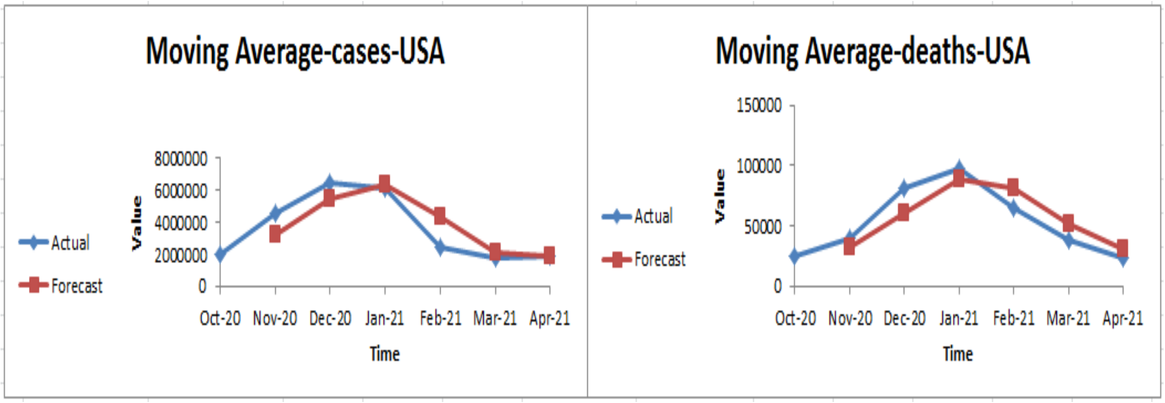

Figure 9: Moving Average for Cases and Deaths in the United States

Table 1 K-means for Confirmed Cases

	Oct-2020		Nov-2020		Dec-2020		Jan-2021		Feb-2021		Mar-2021		Apr-2021	
	S	**N**	**S**	**N**	**S**	**N**	**S**	**N**	**S**	**N**	**S**	**N**	**S**	**N**
							China							
C1	12	25.083	15	25.933	9	89.333	13	83.308	4	53.250	14	22.214	8	20.625
C2	10	17.400	6	47.000	6	123.500	8	127.000	6	41.500	10	16.700	7	38.000
C3	4	44.250	8	91.500	10	105.400	9	166.000	7	30.429	5	33.000	8	28.625
C4	5	33.800	1	133.000	6	77.000	4	222.000	11	20.636	2	62.000	7	14.571
							India							
C1	9	75550.556	15	45054.000	8	34749.375	10	17756.700	4	9043.750	9	16718.444	10	113714.000
C2	7	63810.000	8	39497.375	8	27681.750	2	37061.500	11	11663.364	7	25146.714	6	194634.000
C2	8	53586.625	4	49027.250	14	21744.000	3	6384.000	8	13511.750	7	42189.143	5	283760.000
C4	7	45168.571	3	30276.333	1	0	16	13755.875	5	16413.000	8	60950.875	9	359355.000
							Iraq							
C1	4	4490.000	4	2797.250	7	1987.000	7	905.143	8	1328.375	5	3789.200	5	5735.600
C2	16	3790.250	9	3456.444	4	1625.250	11	811.909	7	2341.000	10	4646.200	5	6420.800
C3	7	3134.857	13	2373.077	11	1265.727	7	754.429	6	3359.667	10	5247.700	7	6967.571
C4	4	2275.250	4	1693.000	9	934.333	6	632.833	7	4097.286	6	6258.333	13	8055.385
							USA							
C1	10	45504.100	6	107121.833	13	191926.538	8	141983.625	5	126109.600	12	59858.583	9	77985.667
C2	13	61379.846	9	135809.889	7	219866.429	5	277990.800	8	95909.875	8	71238.625	10	65192.600
C3	5	79445.000	9	166523.333	9	237960.000	9	223277.667	7	58736.857	5	40225.800	3	36351.667
C4	3	93329.000	6	190163.833	2	126640.500	4	179152.000	8	74088.250	6	54020.833	8	52738.625

4 Conclusion

This paper examined the number of confirmed cases and fatalities during the Covid-19 pandemic. Various approaches were used. Moving Average and Exponential Smoothing were based on the idea that the most recent data was generally the greatest forecaster of the future. The k-means model performed well on the data set. These methods provided a more complete picture of the pandemic's impact on each country. According to the findings, China managed the Covid-19 crisis successfully and has a high rate of recovery when compared to the other three countries. Covid-19 has had a particularly negative impact on the United States and India.

References

[1] Brunese Luca, Fabio Martinelli, Francesco Mercaldo, Antonella Santone, Machine learning for coronavirus COVID-19 detection from chest x-rays, Procedia Computer Science, **176,** (2020), 2212–2221.

[2] Deng Qing, Bo Hu, Yao Zhang, Hao Wang, Xiaoyang Zhou, Wei Hu, Yuting Cheng, Jie Yan, Haiqin Ping, Qing Zhou, Suspected myocardial injury in patients with COVID-19: evidence from front-line clinical observation in Wuhan, China, International Journal of Cardiology, **311,** (2020), 116–121.

[3] T. Aishwarya, V. Ravi Kumar, Machine Learning and Deep Learning Approaches to Analyze and Detect COVID-19: A Review, SN Computer Science, **2,** no. 3, (2021), 1–9.

[4] R. Fitriani, W. D. Revildy, E. Marhamah, T. Toharudin, B. N. Ruchjana, The autoregressive integrated vector model approach for Covid-19 data in Indonesia and Singapore, Journal of Physics: Conference Series, **1722,** no. 1, (2021), 012057.

[5] Abdullah Milad, Kamal Kolo, Peyman Aspoukeh, Rahel Hamad, James R. Bailey, Time Series Modeling and Simulating the Lockdown Scenarios of Covid-19 in Kurdistan Region of Iraq, Journal of Infection in Developing Countries, **15,** no. 3, (2021), 370–381.

[6] Naim Rochmawati, Hanik Badriyah Hidayati, Yuni Yamasari, Wiyli Yustanti, Lusia Rakhmawati, Hapsari PA Tjahyaningtijas, Yeni Anistyasari, Covid Symptom Severity Using Decision Tree. Third International Conference on Vocational Education and Electrical Engineering, IEEE,(2020), 1–5.

[7] R. Sujath, Jyotir Moy Chatterjee, Aboul Ella Hassanien, A machine learning forecasting model for COVID-19 pandemic in India, Stochastic Environmental Research and Risk Assessment, **34,** (2020), 959–972.

[8] Abdulaziz Alorf, The Practicality of Deep Learning Algorithms in Covid-19 Detection: Application to Chest X-ray Images, Algorithms, **14,** no. 6, (2021), 183.

[9] Ali Auwal Shehu, Zarul Fitri Zaaba, A Study on Contact Tracing Apps for Covid-19, Privacy and Security Perspective, Webology, **18,** no. 1, (2021), 341–359.

[10] Shiu Kumar, Ronesh Sharma, Tatsuhiko Tsunoda, Thirumananseri Kumarevel, Alok Sharma, Forecasting the spread of COVID-19 using LSTM network, BMC bioinformatics, **22,** no. 6 (2021), 1–9.

[11] J. Soekiman, Teguh Dwi Putranto, Daniel Susilo, Erica Monica A. Garcia, Economic Sector during the Covid-19 Pandemic, Indonesian Instagram Users Behavior, Webology, **18,** no. 1, (2021), 166–178.

[12] Amit Singhal, Pushpendra Singh, Brejesh Lall, Shiv Dutt Joshi, Modeling and prediction of Covid-19 pandemic using Gaussian mixture model, Chaos, Solitons and Fractals, **138,** (2020), 110023.

[13] https://ourworldindata.org/coronavirus-source-data

[14] Israa Ali Alshabeeb, Nidaa Ghalib Ali, Saba Abdulameer Naser, Wafaa M. R. Shakir, A Clustering Algorithm Application in Parkinson Disease based on k-means Method, Computer Science, **15,** no. 4, (2020), 1005–1014.

[15] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006.

[16] Jiawei Han, Micheline Kamber, Jian Pei, Data mining concepts and techniques, Third Edition, The Morgan Kaufmann Series in Data Management Systems, **5,** no. 4, (2011), 83–124.

[17] Kuljeet Singh, Sourabh Shastri, Arun Singh Bhadwal, Paramjit Kour, Monika Kumari, Anand Sharma, Vibhakar Mansotra, Implementation of exponential smoothing for forecasting time series data, Int. J. Sci. Res. Comput. Sci. Appl. Manag., Stud 8, (2019).