Journal of Computer Science

Original Research Paper

A Time Series Forecasting for the Cumulative Confirmed and Critical Cases of the Covid-19 Pandemic in Saudi Arabia using Autoregressive Integrated Moving Average (ARIMA) Model

[1]Samer H. Atawneh, [2]Osamah A.M. Ghaleb,
[3,*]Ahmad MohdAziz Hussein, [4]Mohammad Al-Madi and [5]Bilal Shehabat

[1]College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia
[2]College of Engineering and Computer Science, Almustaqbal University, Qassim, Saudi Arabia
[3]Deanship of E-Learning and Distance Education, Umm Al-Qura University, Mecca, Saudi Arabia
[4]Faculty of Computer Studies, Arab Open University, Jeddah, Saudi Arabia
[5]Common First Year, King Saud University, Riyadh, Saudi Arabia

Article history
Received: 17-08-2020
Revised: 28-09-2020
Accepted: 29-09-2020

Corresponding Author:
Ahmad Mohdaziz Hussein,
Deanship of E-Learning and Distance Education, Umm Al-Qura University, Mecca, Saudi Arabia
Email: ahmadalmahsiri80@gmail.com
amihussein@uqu.edu.sa

Abstract: Reviews at present, different machine learning techniques and algorithms have been applied for predicting significant factors of the Coronavirus Disease-2019 (COVID-19) such as the outbreak and diagnosis. In this study, the most accurate time series forecasting model, namely, the Autoregressive Integrated Moving Average (ARIMA) model is used to forecast the expected cumulative number of confirmed and critical cases in Saudi Arabia for the upcoming months. Additionally, the dataset is collected from the King Abdullah Petroleum Studies and Research Centre (KAPSARC). Acquiring the number of expected cases within a short period is considered crucial as it provides an important knowledge that can be applied by the health sector in containing the COVID-19 pandemic and forming the proper precautions and strategies that are concerned on the public health system. The main finding of this research is that the number of cumulative confirmed cases is expected to increase at a high rate in the upcoming two months, while the number of critical cases is forecasted to increase at a smaller rate compared to the total number of cases. To evaluate the performance of the adopted model, different statistical matrices as the R Squared, Mean Squarer Error, Root Mean Square Error and Mean Absolute Error are used in this research. It is found to be proven from the findings that the proposed model generates an accurate prediction of the expected number of cumulative confirmed and critical cases in the upcoming months.

Keywords: COVID-19, Time Series Forecasting, Pandemic, ARIMA Model

Introduction

The new coronavirus, namely, the COVID-19, has been first identified in Wuhan, China at the end of 2019. Within a few weeks, it spread out all over the world (El Homsi *et al*., 2020). In March 2020, the World Health Organisation (WHO) has declared this virus as a global pandemic. According to the WHO, COVID-19 has the responsibility of injuring millions of people over the world since the first notification of COVID-19 until these days. According to the statistics of the WHO, this pandemic has infected millions of people across the globe since the first notification of it until these days (WHO, 2020). The rapid growth of COVID-19's cumulative confirmed and critical cases has attracted attention for many researchers due to its colossal importance from the medical sector where different Machine Learning (ML) algorithms have been applied to create different forecasting time series models. Having applied these algorithms into many areas, forecasting time series models are used to predict the number of cumulative confirmed and critical cases in the upcoming two months. Based on the Saudi Ministry Of Health (MOH), Saudi Arabia has registered more than 200 thousand confirmed cases until July 9th, 2020 (Arabia, 2020), which represents the 14th position of the total number of confirmed cases and the fourth position in Asia according to the world’s ranking. Due to the vast spread of COVID-19 around the world during the past few months, the latest updates demonstrate that the daily number of infected cases increases exponentially.

Many countries are suffering from the coronavirus outbreak with an increase in the number of infected and death cases. Consequently, many countries have applied very restricted precaution measures including the lockdown and curfew measures to control the effects of such an outbreak. This includes suspending schools and universities and closing borders and alleviating the number of international flights until further notice. The purpose of these preventive measures is to reduce the likelihood of physical contact among people and thus, the transmission of COVID-19 can be efficiently managed (Yousaf *et al*., 2020). However, with the emergence of these restrictions and precautionary measures, the number of infected cases remains increasing daily.

Although scientists and health organizations around the globe are further working to produce a vaccine for COVID-19, a treatment medicine is still yet undetected until now and the applied precautionary measures are insufficient for controlling this severe outbreak. Health institutions are required to collect further details on the behaviours of Coronavirus patients. Additionally, details on spreading factors, symptoms and treatment methods are considered to forecast the number of infected cases in the future so that they could provide necessary medical supplies and equipment for the confirmed and critical cases. Further, the novel nature of COVID-19 adds great uncertainty on the expected time of the disappearance of this pandemic. Therefore, a short-term prediction is extremely significant even with slight signs and symptoms, which can predict the upcoming days for ensuring more effective control of different economic, social and health issues (Yousaf *et al.*, 2020). The epidemiological time series prediction plays a significant role in health public systems where policy makers can develop appropriate strategic plans for avoiding any possible epidemic in the future (Ribeiro *et al*., 2020). In this context, Artificial Intelligence (AI), Machine Learning (ML) and deep learning methods play an important role in developing efficient forecasting models that can accurately predict all necessary factors based on the previously collected data. At present, several studies that use AI and ML methods in a pandemic prediction (for predicting the cases of Covid-19) have started to emerge due to the effectiveness of these methods in understanding current patients’ behaviours. Moreover, these studies can accurately forecast newly infected cases, critical cases and recovered cases shortly.

This research aims to assist the Saudi MOH in estimating the medical staff, necessary supplies and equipment that are required for COVID-19’s confirmed and critical cases in Saudi Arabia. This estimation is based on developing an effective time-series forecasting model that can predict the future confirmed and critical cases. The ARIMA model represents one of the most effective time-series prediction models, which is used in this research for forecasting purposes such as the one that is proposed in this research. This research uses the dataset of COVID-19’s infected cases, which are made available until July 7th, 2020 (Arabia, 2020). The results of this research are expected to alert the policy makers in the Saudi MOH to prepare themselves for withstanding the future of COVID-19’s critical cases and reacting accordingly by using the proper precautions and strategies.

The foreseeable contributions of this study are summarised as follows:

- To predict the potential cumulative number of confirmed and critical COVID-19 cases in Saudi Arabia by developing an efficient multi-days-ahead prediction model
- To assist the policy makers of the Saudi MOH in conducting their decision-making processes that rely on the accuracy of the proposed forecasting model to manage the COVID-19 pandemic and to form the strategies that are related to the public health systems

The remaining sections of this paper are outlined as follows. Section 2 presents the secondary research, which forms the literature review. Details on the methods and datasets that are used in the paper are described in section 3. The obtained findings are discussed in section 4. Section 5 provides a detailed discussion of the entire research. Finally, Section 6 concludes this paper and highlights the suggested future research.

Literature Review

Many AI techniques have far been introduced to assist in tackling many issues not only in the domain of computer sciences but also in many other domains. One of those issues that have persistently emerged lately is the COVID-19 pandemic. The literature has currently witnessed extensive research for attempting to produce various approaches and techniques that can contribute to managing and tackling the enormous and abrupt growth of the COVID-19 pandemic across the globe (Alabool *et al*., 2020). This pandemic has also emerged sporadically throughout the entire world (Cruz and Dias, 2020) and according to (Rothe *et al*., 2020) asymptomatic people represent possible infectious causes of such a pandemic where the entire transmission dynamics can change for it.

Other researchers as (Cruz and Dias, 2020), provide a qualitative explanatory research starting from the outbreak to the pandemic itself. Their study is based on several research methods that are derived from (Yin, 1988). Further, their study includes four countries, which comprise China, Italy, US and Brazil. The four selected countries represent the dataset of their study as (N = 4), which forms an analysis unit to these countries. Their obtained results demonstrate that permanent lockdown affects adversely on the interaction of economic and social factors and cannot act as the only method of tackling the issues related to the pandemic.

In the AI domain, the evaluation of machine learning has played a significant role in the COVID-19 pandemic at its very early stages (Alabool *et al*., 2020). The aim of applying the machine-learning domain according to many researchers is to provide an analysis of the current cases that are produced by the pandemic and to study the inflectional effects within the upcoming days. In the context of this paper, this domain is adopted and used to predict the number of expected inflectional cases of the COVID-19 pandemic upon the following days when starting from any given day. For instance, (Gupta *et al*., 2020) aim at forecasting the expected number of infectious cases for a couple of weeks later starting from the same day of releasing the new cases. Similarly, (Alzahrani *et al*., 2020) forecast the daily estimated numbers of emerging cases in Saudi Arabia based on the use of four models in which their prediction is integrated. The accuracy of these models is assessed according to different accuracy measures, which comprise: Root mean square error (RMSE), Coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Relative Error (RMSRE). They demonstrate that the Autoregressive Integrated Moving Average (ARIMA) model is more effective in comparison to other models. However, a second rank is given to the ARMA model followed by the AR model and afterward, MA models. Furthermore, (Ribeiro *et al*., 2020) study another similar idea by involving six different approaches related to the field of machine learning. Such approaches comprise the CUBIST, RF, RIDGE, SVR, 229 and stacking ensemble, including the ARIMA statistical approach. These approaches are applied according to different time series that are forecasted with one, three and six days before the COVID-19 aggregately confirmed cases within 10 different Brazilian states based on increased daily cases. It is found to be proven that the obtained findings from their research recommends using the assessed models for monitoring and forecasting the persistent increase of the COVID-19 cases when such models assist managers within the decision-making support systems. Additionally, (Tuli *et al*., 2020) propose a cloud-based machine-learning model to forecast the spread of COVID-19 pandemic cases. Different measures are used for their model, which include the Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and Coefficient of determination (R^2). In particular, an ML-based enhanced model is used in their study to forecast any possible threats related to this pandemic across the globe. Their results find that applying iterative weighting for fitting the distribution of the Generalized Inverse Weibull model; a more effective fit is possibly achieved towards enhancing a prediction framework. In the same context, (Fayyoumi *et al*., 2020) develop an online questionnaire tool for collecting data of different prediction approaches according to machine learning and statistical models. The models aim to predict patients who are infected with COVID-19 taking into account their symptoms and signs. Their obtained results showed the best accuracy and precision through the performance of their models. Moreover, (Yadav *et al*., 2020) investigate the effects of the COVID-19's prediction on several countries, such as the USA. This prediction is conducted by determining the data-driven of the machine learning time series, which analyzes active, infected and healed cases to erupt the prediction.

In the same manner, (Ardakani *et al*., 2020; Li *et al*., 2020; Wu *et al*., 2020) analyse the infections and predictions of COVID-19 cases. These researchers have obtained their required data from different clinical sources such as hospitals, health clinical centers and so on. Although these researchers have investigated the infection of the COVID-19 pandemic, still their aims are different. For example, (Ardakani *et al*., 2020) provide comparisons of two COVID-19 pandemic groups. In particular, they compare between those infected and non-infected groups with this pandemic. Further, (Li *et al*., 2020) aim to provide a prediction of various critical cases of patients that are infected with COVID-19. On the other hand, (Wu *et al*., 2020) apply the key blood from different suspected patients with the same Computed Tomography (CT) information or with the same incurred symptoms to determine the infections produced by the COVID-19 pandemic. Additionally, all of the three indicated researchers have used the accuracy and specificity metrics as performance evaluation metrics. However, (Ardakani *et al*., 2020) and (Wu *et al*., 2020) use an additional metric, which is the sensitivity metric, while (Li *et al*., 2020) have rather used the AUROC to study and examine their obtained dataset.

In conclusion, it can be inferred from the literature that one of the uses of the machine-learning approach that has been of interest to many researchers since the emergence of the COVID-19 pandemic, is to apply different statistical models that assist in predicting infected and non- infected cases of this pandemic according to patients' arising symptoms and signs. The machine-learning model can provide the support of screening patients instantly (Fayyoumi *et al*., 2020) to provide critical analysis for the discussed literature, it is worth highlighting the investigations conducted by the current research when applying the machine-learning approach. In light of the foregoing research advancements, the approach is motivating for several studies, which may put an influence on its usefulness effectively. It can be deduced from the previously explained studies that the majority of researchers have provided different studies on performing a prediction of the COVID-19 infections on suspicious and non-suspicious patients. On the other hand, some other researchers have carried out a comprehensive study on determining the critical cases of infected patients. To summarise, it is also perceived from the literature that further investigations are studied and taken into consideration for forecasting various infectious cases based on the use and the proposal of different machine-learning approaches since these approaches aim at instantly detecting infected and non-infected cases of the COVID-19 pandemic effectively and efficiently. From this point, several types of research study the effect of predicting infected cases when using the machine-learning approach for such a purpose.

Materials and Methods

This section is comprised of two subsections. The first subsection presents the collected dataset that relates to COVID-19. The second subsection describes the ARIMA model that is used to forecast the number of critical cases in Saudi Arabia. During the occurrence of the pandemic, a significant question is to ascertain the progression and the point of inflection. Consequently, the proposed method generates a trusted model for forecasting the critical cases in the future so that it can alert the medical sector to read this sign.

Dataset Description

The dataset is collected from the King Abdullah Petroleum Studies and Research Center (KAPSARC) official web site (KAPSARC, 2020). It includes different statistics about the daily and cumulative COVID-19 cases such as confirmed, active, critical, and recovered and mortalities cases from different cities in Saudi Arabia. From this dataset, two subsets are used where the first one about the confirmed cases from the date of Mar. 2nd 2020 till Jul. 7th 2020 and the second subset about the cases in critical conditions from the date of Apr. 10th 2020 till Jul. 7th 2020. For validating purposes, 30% from both subsets are used as a testing set in this research.

ARIMA Model

Over the past few months, a growing number of publications have tried to predict the course and eventual magnitude of the COVID-19 pandemic, employing various methods (Giordano *et al*., 2020; Read *et al*., 2020; Zhou *et al*., 2020). One of these methods is the Autoregressive Integrated Moving Average (ARIMA) model (Gupta and Pal, 2020; Kumar *et al*., 2020). the Autoregressive Integrated Moving Average (ARIMA) which was built in 1970 by Gwilym, M. Jenkins and George E.P Box is effectively employed in forecasting (Chen *et al*., 2008; Alsharif *et al*., 2019).

The ARIMA is regarded as one of the widely employed models for visualization and infographics (time series). It specifically integrates the regressive procedure and the rolling (animated) average. It also permits the forecast of a specified time series by minding its lags as well as prediction inaccuracies. The optimum ARIMA model specifications are selected by utilizing the Akaike's Information Criterion (AIC) (Perone, 2020).

The Autoregressive (AR) model is the easiest and more extensively employed model framework. In the *AR* model, the present output Y_t is indicated by former parameters and values¸$\alpha t{-}p$, as defined in (1), where t represents time and p represents the order of the parameters. (1) where ($Y_{t\text{-}1}$) is stated by (2) A rarer model framework against the AR is known as the Moving Average (MA). In the MA model, the output Y_t is indicated concerning innovation input and is clarified with the weights B_q as expressed in (2). A further reinforced structure known as ARMA model that is indicated in (3) is achieved by integrating both AR and MA. An additionally improved model is known as Autoregressive Integrated Moving Average (ARIMA), in which variations are inserted at least not less than one time. The formula ARIMA model is defined in (4). This model has a lot of actual fruitful prediction instances in publications in various subjects like the one stated (Arabia, 2020) ARIMA model is recognized by designating the order for the three expressions: p for AR, q for MA and the number of different stages d.

A flawless Auto-Regressive (AR only) model is one where Y_t relies solely on its own lags. That is, Y_t is a function of the 'lags of Y_t':

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1 \tag{1}$$

where, $Y_{t\text{-}1}$ represents *lag1* of the series, β_1 represents the coefficient of lag1 that the model determines and α is the intercept term, which is determined as well by the model.

Similarly, a flawless Moving Average (MA only) model is such that Y_t relies solely on the lagged prediction inaccuracies:

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q} \tag{2}$$

where the error (fault) terms represent the errors (faults) of the autoregressive models of the respective lags. The following errors E_t and $E_{t\text{-}1}$ refer to the errors as of the following expressions:

$$\begin{aligned} Y_t &= \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_0 Y_0 + \epsilon_t \\ Y_{t-1} &= \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + .. + \beta_0 Y_0 + \epsilon_{t-1} \end{aligned} \tag{3}$$

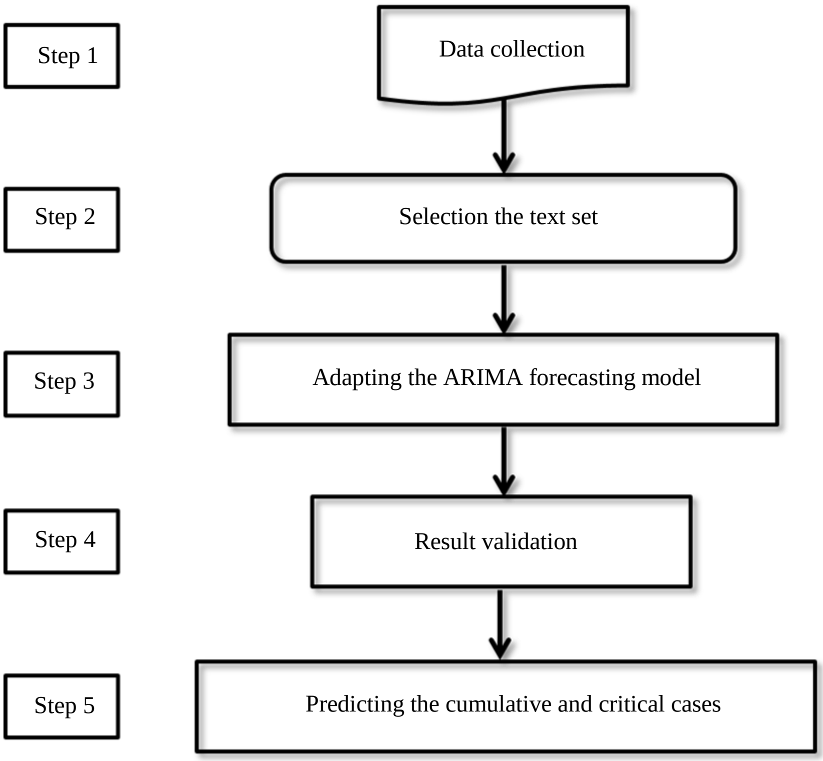

Fig. 1: A general framework for the proposed research

Those are AR as well as MA models in the order respectively.

An ARIMA model is one in which the time series is divergence at least not less than one time to ensure that it is static and the AR as well as the MA terms are integrated. So the expression turns out as:

$$\begin{aligned} Y_t &= \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t \\ &+ \phi_1 + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q} \end{aligned} \tag{4}$$

ARIMA model described in terms:

Forecasted Y_t = Constant + Linear combination Lags of Y (up to p lags) + Linear Combination of Lagged forecast errors (up to q lags) (Wang *et al*., 2015).

Figure 1 illustrates the general framework for the proposed research. The framework consists of five steps that start by collecting the dataset from reliable sources about COVID-19's cases in Saudi Arabia. In step 2, for the validation purposes, 30% of the collected data is used as a testing set. In step 3, the proposed research is represented by adopting the ARIMA model for predicting the number of cumulative confirmed and critical cases in the following day/month in Saudi Arabia. Step 4 involves the validation of the forecasted cases against the observed cases. In step 5, the forecasting results of this research are highlighted in detail.

Results

To validate the forecasted results of this study, 30% of the datasets are used as a testing set. The research applies two testing sets in order to test the forecasting validity based on the use of the ARIMA model. First, the model is applied to predict the number of cumulative confirmed cases. Second, comparisons are produced for the number of forecasted critical cases along with the real number of critical cases through the testing sets. As shown from Table 1, the observed number of the cumulative confirmed cases and the forecasted number of the cumulative confirmed cases are presented from 25/05/2020 to 25/06/2020 in Saudi Arabia. Similarly, Table 2 presents the observed number of cumulative critical cases and the forecasted number of cumulative critical cases from 05/06/2020 to 06/07/2020 in Saudi Arabia. Consequently, it is found to be proven from these results that the observed and forecasted numbers of the confirmed and critical cases are extremely approaching from each other. To compare the results that are obtained from Fig. 5, it is found that the predicted number of the cumulative confirmed cases (in red) highlights approaches from the observed number of the cumulative confirmed cases within a particular period. Further, Fig. 6 illustrates the forecasted number of cumulative critical cases (in red) highlights and the observed number of cumulative critical cases. From these results, it can be inferred that the predicted number is almost approaching from the real number of critical cases within a particular period.

Discussion

The Cumulative Confirmed and Critical Cases

To forecast the future cases of COVID-19, the ARIMA model is used to predict the future outbreak that is related to the COVID-19 cases for the next two months. Based on the daily and the cumulative number of confirmed cases, the charts of the intended cases are plotted (Figs. 2 to 4). It can be realised from the figures that the numbers of confirmed and critical cases are increased when the curfew was lifted on the 29th of May 2020. This increment in the number of cases occurred due to the number of factors, although the Saudi government has imposed several restricted precautionary measures, which resulted in reducing the number of daily cases in the mid of Jun. month.

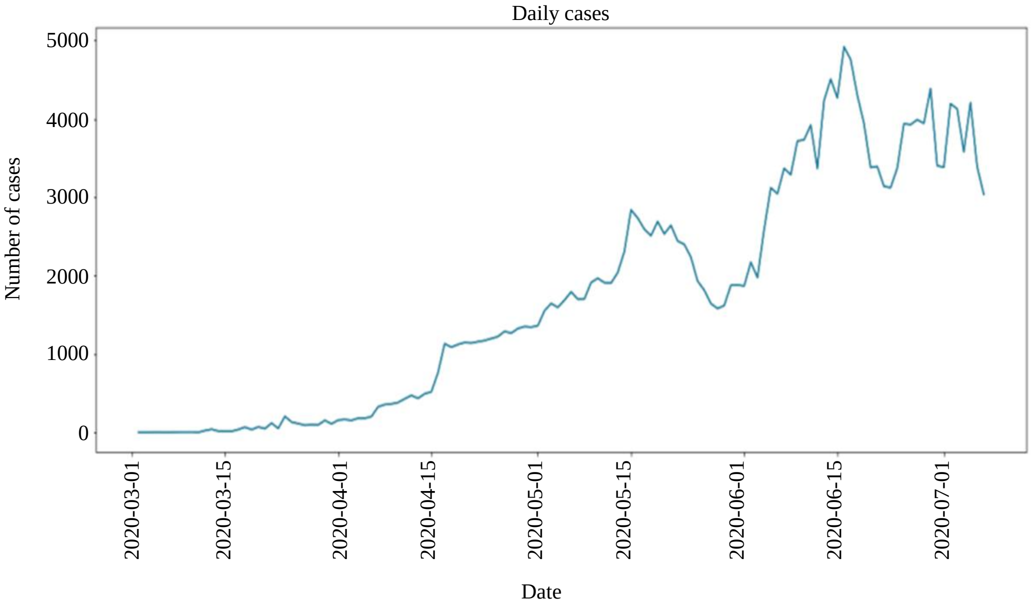

Fig. 2: The daily number of confirmed cases

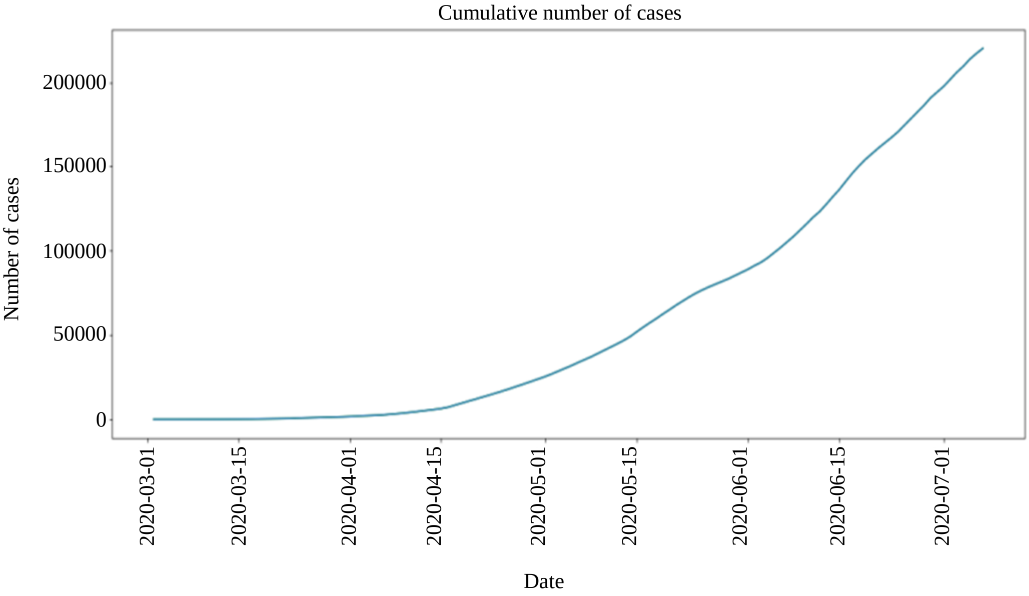

Fig. 3: The cumulative number of confirmed cases

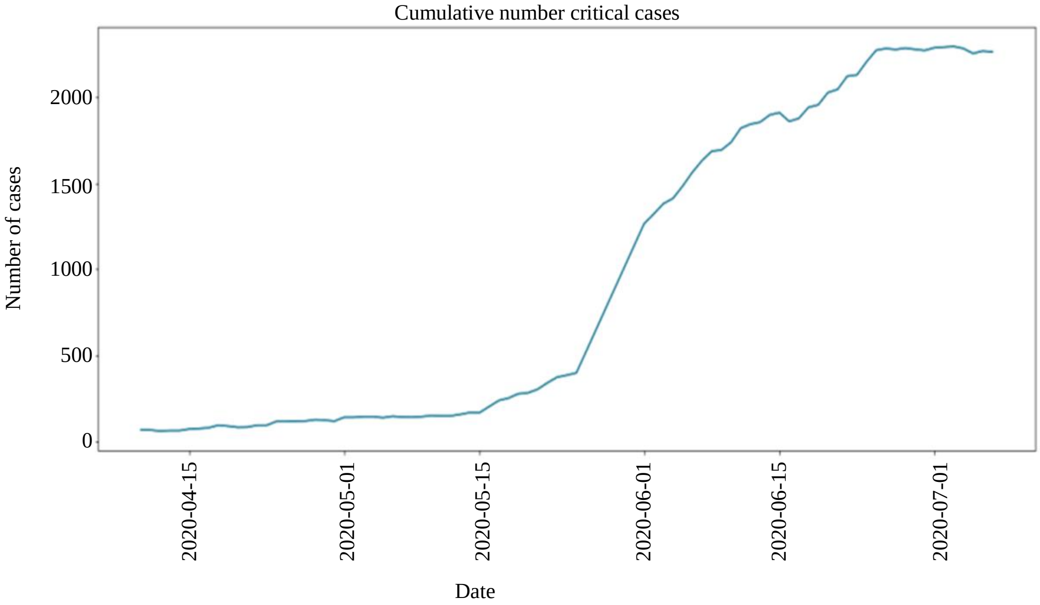

Fig. 4: The cumulative number of critical cases

Table 1: The validation of the model prediction for the cumulative confirmed cases

Date	The observed number of cumulative confirmed cases	The forecasted number of cumulative confirmed cases
25/05/2020	76726	77066
26/05/2020	78541	78880
27/05/2020	80185	80619
28/05/2020	81766	81803
29/05/2020	83384	83377
30/05/2020	85261	85058
31/05/2020	87142	87134
01/06/2020	89011	88976
02/06/2020	91182	91057
03/06/2020	93157	93312
04/06/2020	95748	95096
05/06/2020	98869	98218
06/06/2020	101914	101889
07/06/2020	105283	105051
08/06/2020	108571	108879
09/06/2020	112288	111752
10/06/2020	116021	115835
11/06/2020	119942	119716
12/06/2020	123308	123904
13/06/2020	127541	126562
14/06/2020	132048	131812
15/06/2020	136315	136259
16/06/2020	141234	140680
17/06/2020	145991	146093
18/06/2020	150292	150437
19/06/2020	154233	154555
20/06/2020	157612	157988
21/06/2020	161005	161071
22/06/2020	164144	164387
23/06/2020	167267	167499
24/06/2020	170639	170194
25/06/2020	174577	174233

Table 2: The validation of the mode prediction for the cumulative critical cases

Date	The observed number of cumulative critical cases.	The forecasted number of cumulative critical cases
05/06/2020	1484	1452
06/06/2020	1564	1540
07/06/2020	1632	1649
08/06/2020	1686	1705
09/06/2020	1693	1729
10/06/2020	1738	1758
11/06/2020	1820	1780
12/06/2020	1843	1849
13/06/2020	1855	1871
14/06/2020	1897	1876
15/06/2020	1910	1878
16/06/2020	1859	1887
17/06/2020	1877	1923
18/06/2020	1941	1929
19/06/2020	1955	1985
20/06/2020	2027	1991
21/06/2020	2045	2075
22/06/2020	2122	2083
23/06/2020	2129	2153
24/06/2020	2206	2185
25/06/2020	2273	2245
26/06/2020	2283	2249
27/06/2020	2277	2267
28/06/2020	2285	2270
29/06/2020	2278	2284
30/06/2020	2272	2287
01/07/2020	2287	2296
02/07/2020	2291	2297
03/07/2020	2295	2302
04/07/2020	2283	2310
05/07/2020	2254	2312
06/07/2020	2268	2315

The Validation of the Forecasting Model

Tables 1 and 2 highlight the number of forecasted cases against the number of real cases for the cumulative confirmed and critical cases. Additionally, the obtained results from the model validation of the total number of confirmed and critical cases are plotted in Fig. 5 and 6, respectively. It is demonstrated from these figures that the ARIMA model can accurately predict the number of future cases as the number of observed cases and the number of forecasted cases is near to the number of observed cases. In later stages, the accuracy of the adopted model is validated.

Forecasting COVID-19's Outbreak and Critical Cases

Respectively, The ARIMA model is utilized to predict the outbreak of COVID-19 and the number of critical cases in Saudi Arabia. In particular, the research focuses on predicting the number of future cases when the curfew was lifted. To achieve the objectives of the study, three steps are conducted. The first step is to determine the set of the ARIMA (*p*, *d*, *q*) parameters, which denote the notations for seasonality, trend and noise date. To find the optimal set of parameters that could result in the most effective performance for the proposed model, these parameters are calculated by using the greedy search which resulted in AIC values as 1441.69 for the cumulative confirmed cases set and 616.62 for the critical cases set. After that, the optimal values of the *p*, *d*, *q* parameters of the ARIMA model are fitted. Finally, the fitting results are identified for the model to check the odd behaviours as it can be found to be proven from the obtained results that the model fits for the cumulative confirmed and critical cases.

According to the number of daily-confirmed cases starting from the date of the first confirmed cases in Saudi Arabia until the present, the ARIMA model is used for forecasting the total number of future cases for the next two months (Fig. 7). From this figure, the number of future cases is increased and it is expected to reach 400,000 cases in the middle of September 2020. This implies that the spread of the COVID-19 pandemic has prominently increased and the number of total cases is likely to double in the next two months if the status is yet in its current situation.

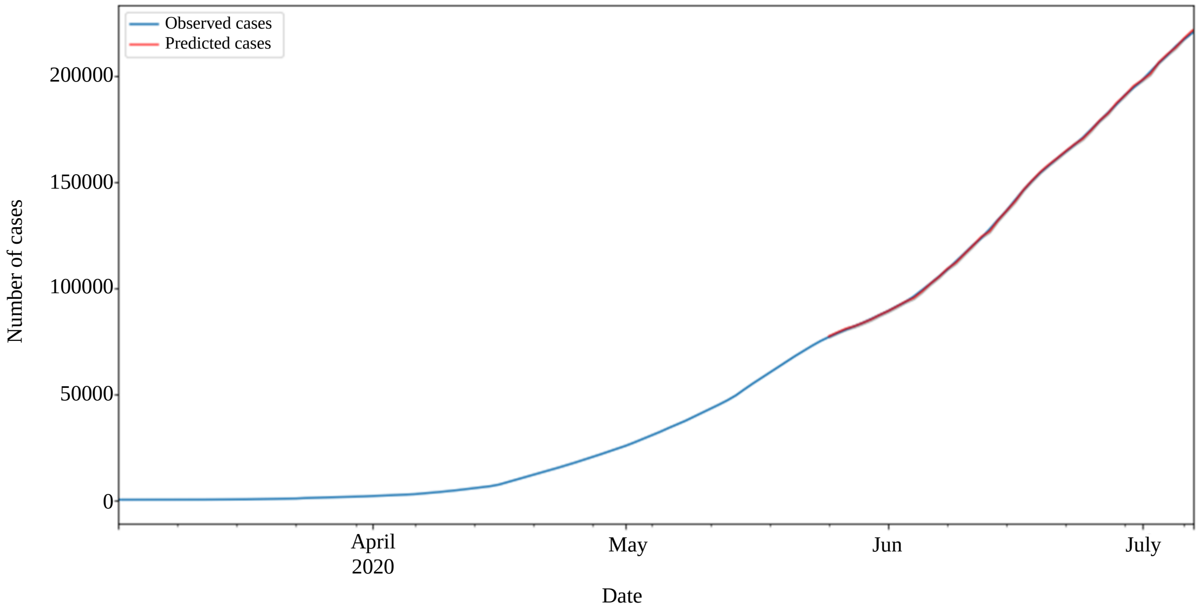

Fig. 5: Comparisons of the total number of predicted cases vs. the observed cases

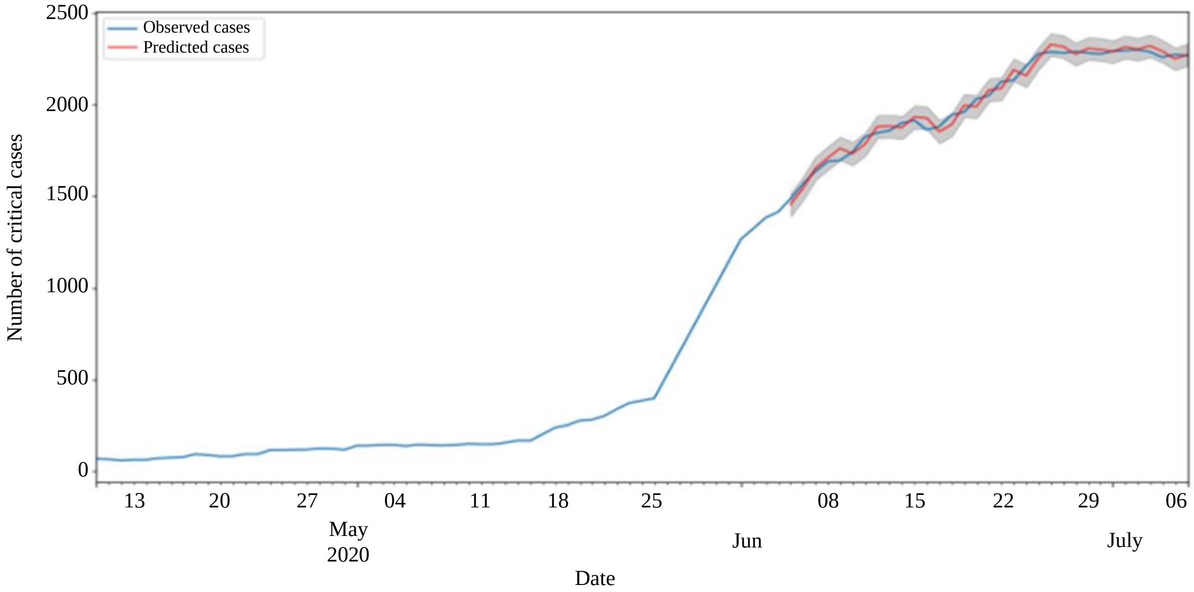

Fig. 6: Comparing the total number of predicted critical cases vs. the observed critical cases

Figure 8 shows the predicted number of future critical cases in Saudi Arabia for the next two months. It can be inferred from the figure that the current number of critical cases in Saudi Arabia is small compared to many other countries as the government provides efficient healthcare systems for all infected cases from the first day of the pandemic's emergence. Additionally, the number of critical cases is seen to be low in comparison with the number of confirmed cases in this country (approximately 1% of the total number of cases). Furthermore, the number of future critical cases for the next two months is predicted by applying the ARIMA model where the forecasting process that is conducted in this research demonstrates that the number of critical cases is likely to be increased, but with smaller numbers compared to the total number of predicted cases. The number of future critical cases is expected to reach around 3500 cases in September 2020.

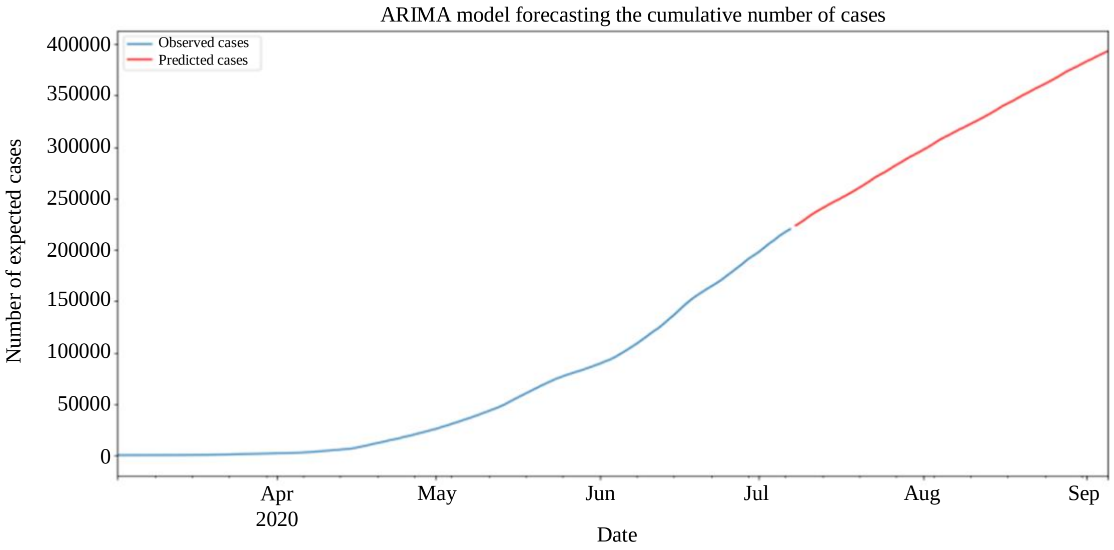

Fig. 7: The forecasted number of cumulative future cases

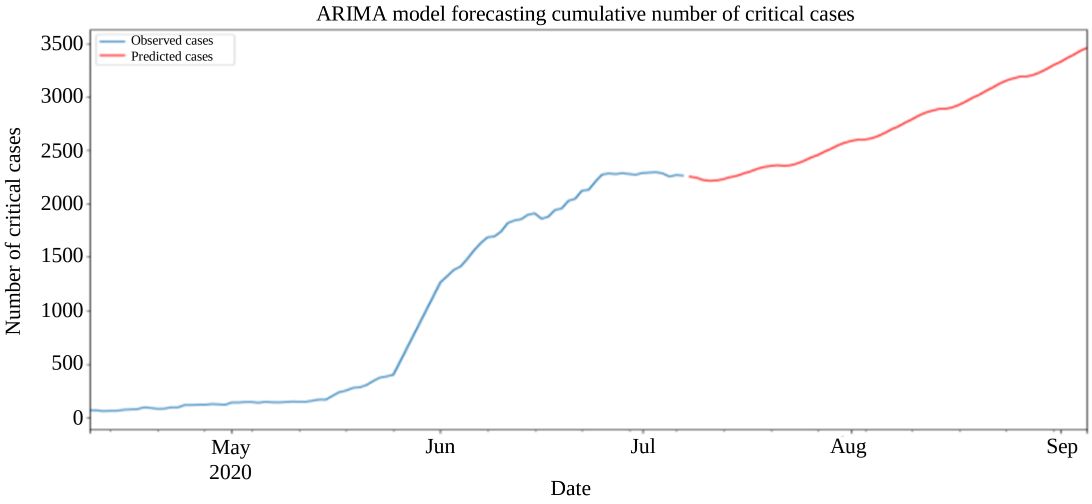

Fig. 8: The forecasted number of future critical cases

Performance Evaluation

To evaluate the performance of the forecasting process in this research, the ARIMA model is applied for using different statistical measures, particularly, regression metrics are calculated. It is found to be proven from the obtained results that the accuracy of prediction is high for the confirmed and critical cases (Table 3) In fact, the regression metrics that are used to evaluate the research’s forecasting for the cumulative confirmed and critical cases are briefly discussed as follows.

R Squared (R^2)

R squared (it sometimes refers to the coefficient of determination) is a key outcome of the regression analysis, which represents the square of the correlation (*r*) among the forecasted scores and real scores. Therefore, it represents the value between 0 and 1. The following formula represents the R^2:

$$R^2 = 1 - \frac{\frac{1}{K}\sum_{i=1}^{K}\left(Y_i - \hat{Y}_i\right)^2}{\frac{1}{K}\sum_{i=1}^{K}\left(Y - \bar{Y}_i\right)^2}$$

where K refers to the aggregate count of rows/observations, Y_i denotes the values of the real cases, $\hat{Y}_i$ refers to the values of the forecasted cases and $\bar{Y}_i$ denoted as following $\bar{Y}_i = \frac{1}{K}\sum_{i=1}^{K} Y_i$.

Table 3: The performances of the evaluated metrics

Cases set\Metric	*R* Squared	MSE	RMSE	MAE
Cumulative cases	99.99	193420.11	439.80	346.32
Critical cases	98.15	1150.42	33.92	29.94

Table 4: Comparisons of the performances among different forecasting models

Model name	R^2	MSE	RMSE	MAE
ARIMA	99.87	9533.82	97.74	71.58
ARMA	87.00	13740.5	117.22	73.33
AR	69.00	32504.5	180.29	160.4
MA	46.00	58105.1	241.05	190.87

Mean Square Error (MSE)

The MSE is likewise called as Mean Square Deviation (MSD) and is utilized to evaluate the mean error squares. For example, the square of the distinction between real and predicted values. The following formula represents the MSE formula:

$$MSE = \frac{1}{K}\sum_{i=1}^{K}\left(Y_i - \hat{Y}_i\right)^2$$

Root Mean Square Error (RMSE)

The RMSE is a common metric that is used for measuring the divergences between the observed values and predicted values by a model. After that, the results of the entire divergences are divided by the total observed values. The smaller the value of the RMSE, the closer it is from the observed values:

$$RMSE = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(Y_i - \hat{Y}_i\right)^2}$$

Mean Absolute Error (MAE)

The MAE is a statistical measure that is defined as a metric for measuring the average of absolute values of the divergences between the observed and corresponding forecasted values:

$$MAE = \frac{1}{K}\sum_{i=1}^{K}\left|Y_i - \hat{Y}_i\right|$$

As shown in Table 3, the values of the evaluation metrics prove that the accurate prediction of ARIMA model. The high value of R^2 indicates the quality of the model prediction as it proves the high fitting of our model.

To evaluate the performance accuracy of the ARIMA model in comparison with other time series forecasting models, the proposed model in this research is compared with another model that is proposed by (Alzahrani *et al*., 2020) to forecast the daily and cumulative cases in Saudi Arabia before the curfew was lifted. The performance differences between the ARIMA, ARMA, AR and MA models are evaluated by using the same test set for the confirmed cases in the Saudi Arabia, which includes the data of confirmed cases from 6 April 2020 until 20 April 2020. The outcomes in Table 4 reveal that the ARIMA forecasting model achieves the lowest MSE, RMSE and MAE values. Additionally, it can be inferred that the value of R^2 is approaching the value ‘1’, which has the highest value in comparison with the entire different models, which are indicated in the Table 4. Consequently, this model excels other models, while the ARMA model is ranked second, trailed by AR and the last one is MA models.

Since the above results demonstrate a highly accurate prediction of cumulative confirmed and critical cases, this research provides crucial and validated information regarding the expected outbreak of COVID-19 in Saudi Arabia for the upcoming two months. To achieve generality this information can be utilized by several concerned sectors within the country.

Conclusion and Future Research

The outcomes highlight the potential role of machine learning prediction algorithms. The time-series forecasting models are crucial to predict the future. In this research, the ARIMA model is used to forecast the COVID-19 for the expected number of cumulative confirmed and critical cases of the upcoming months. This model assists the health sector to react accordingly by using appropriate precautions and strategies. By using the dataset that is collected about the COVID-19 cases, the ARIMA model demonstrates that the outbreak of COVID-19 pandemic is will increase in the next two months. On the other hand, the number of critical cases is expected to remain under control with the same increased rate during the past months. The test set has been used to validate the forecasting result; the forecasted cases are compared with the observed cases in the test set since the results show that the numbers of the forecasted cases are very near to the number of observed cases. To evaluate the performance of the ARIMA forecasting model, different statistical metrics that include the R squared, Mean Square Error, Root Mean Square Error and Mean Absolute Error are used. The results of the performance evaluation show that the applied model is highly accurate in predicting the cumulative confirmed and critical COVID-19 cases in Saudi Arabia and can provide crucial knowledge for interested sectors, such as the health sector. It can be inferred that the model demonstrates a highly accurate prediction of the expected numbers of the cumulative confirmed and critical cases. This can be seen from the obtained results that represent 99.99% for the cumulative confirmed cases and 98.15% for the critical cases. Forecasting the correlation between different patients' attributes (e.g., age, gender, blood group, chronic disease and so forth) and the critical cases can be a suggestion for future research. By using different machine-learning prediction algorithms, the important factors that can influence patients' cases in a short period can be predicted.

Acknowledgment

The authors would like to thank the Deanship of Scientific Research at the Saudi Electronic University for funding this research project (No. ELI-CCI20145) through the initiative of supporting E-Learning in Higher Education.

Author's Contributions

Ahamd MohdAziz Hussein: Conceptualization, methodology, formal analysis, writing-original draft preparation, supervision, project administration.

Samer H. Atawneh: Conceptualization, investigation, writing - original draft preparation.

Osamah A.M. Ghaleb: Methodology, software, validation, writing - original draft preparation.

Mohammad Al Madi: Conceptualization, investigation, writing - original draft preparation, writing-review and editing.

Bilal Shehabat: Validation, investigation, writing-original draft preparation.

Conflicts of Interest

"The authors declare that they have no conflicts of interest to report regarding the present study."

References

Alabool, H., Alarabiat, D., Abualigah, L., Habib, M., Khasawneh, A. M., Alshinwan, M., & Shehab, M. (2020). Artificial intelligence techniques for Containment COVID-19 Pandemic: A Systematic Review.

Alsharif, M. H., Younes, M. K., & Kim, J. (2019). Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea. Symmetry, 11(2), 240.

Alzahrani, S. I., Aljamaan, I. A., & Al-Fakih, E. A. (2020). Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. Journal of infection and public health, 13(7), 914-919.

Arabia, M. O. H. K. (2020). COVID 19 Saudi Arabia [WWW Document]. Minist. Heal. https://covid19.moh.gov.sa/

Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., & Mohammadi, A. (2020). Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Computers in Biology and Medicine, 103795.

Chen, P., Yuan, H., & Shu, X. (2008, October). Forecasting crime using the arima model. In 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery (Vol. 5, pp. 627-630). IEEE.

Cruz, B. S., & de Oliveira Dias, M. (2020). COVID-19: from outbreak to pandemic. Global Sci J, 8(3).

El Homsi, M., Chung, M., Bernheim, A., Jacobi, A., King, M. J., Lewis, S., & Taouli, B. (2020). Review of Chest CT Manifestations of COVID-19 Infection. European Journal of Radiology Open, 100239.

Fayyoumi, E., Idwan, S., & AboShindi, H. (2020). Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan. Machine Learning, 11(5).

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nature Medicine, 1-6.

Gupta, R., & Pal, S. K. (2020). Trend Analysis and Forecasting of COVID-19 outbreak in India. medRxiv.

KAPSARC, 2020. Saudi Arabia Coronavirus disease (COVID-19) situation [WWW Document]. https://datasource.kapsarc.org/explore/dataset/saudi-arabia-coronavirus-disease-covid-19-situation/

Kumar, P., Kalita, H., Patairiya, S., Sharma, Y. D., Nanda, C., Rani, M., ... & Bhagavathula, A. S. (2020). Forecasting the dynamics of COVID-19 Pandemic in Top 15 countries in April 2020: ARIMA Model with Machine Learning Approach. medRxiv.

Li, P., Zhang, W., Jiang, X., Zhang, Y., Sun, L., Chen, X., & Shi, Y. (2020). Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. J. Clin. Virol. 128, 104431.

Gupta, R., Pandey, G., Chaudhary, P., & Pal, S. K. (2020). SEIR and Regression Model based COVID-19 outbreak predictions in India. medRxiv.

Perone, G. (2020). An ARIMA model to forecast the spread of COVID-2019 epidemic in Italy. arXiv preprint arXiv:2004.00382.

Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A., & Jewell, C. P. (2020). Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. MedRxiv.

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos, Solitons & Fractals, 109853.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., ... & Seilmaier, M. (2020). Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. New England Journal of Medicine, 382(10), 970-971.

Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing. Internet of Things, 100222.

Wang, W. C., Chau, K. W., Xu, D. M., & Chen, X. Y. (2015). Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. Water Resources Management, 29(8), 2655-2675.

WHO. (2020). WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/

Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., ... & Zhao, M. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. medRxiv.

Yadav, D., Maheshwari, H., & Chandra, U. (2020). Outbreak prediction of covid-19 in most susceptible countries. Global Journal of Environmental Science and Management, 6(Special Issue (Covid-19)), 11-20.

Yin, R.K. (1988). Case Study Research: Design and Methods. Sage Publications., Newbury Park, CA:

Yousaf, M., Zahir, S., Riaz, M., Hussain, S. M., & Shah, K. (2020). Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan. Chaos, Solitons & Fractals, 109926.

Zhou, T., Liu, Q., Yang, Z., Liao, J., Yang, K., Bai, W., ... & Zhang, W. (2020). Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV. Journal of Evidence-Based Medicine, 13(1), 3-7.