International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

# Applying the ARIMA Deep Learning Algorithm to Predict the Coronavirus in the Kingdom of Saudi Arabia using Time Series Data

**Afrah Owehan Al-Rashedi[1], Mohammed Abdullah Al-Hagery[2]**

[1]Master Student, Department of Computer Science, College of Computer,Qassim University, Qassim, KSA
afrahalr12[at]gmail.com

[2]Professor, Department of Computer Science, College of Computer,Qassim University, Qassim, KSA
*hajry[at]qu.edu.sa,*
*drmalhagery[at]gmail.com*

**Abstract:** ***The COVID-19 pandemic has posed a significant threat to humanity, with irreversible societal consequences. Research is being conducted to anticipate the development or return of the pandemic at any later time and discover more effective vaccinations, and consequently reduce the death toll. Maybe in the future context, accurate COVID-19 prediction using deep learning is gaining increasing attention as deep learning approaches are more successful in dealing with non-linear situations. Time series prediction of COVID-19, in terms of the estimated number of confirmed, death, and recovered cases, is performed in our study utilizing the ARIMA model. Short-term infected cases are all predicted in the proposed methodology. We used daily data from April 1, 2020, to May 31, 2021, to train and evaluate the models for our study. The models used in this study are data-driven, and we use two MAPE, and R2 metrics to assess our models' predictive performances. We aim to evaluate and contrast the abilities of ARIMA the model in interpreting complex time series trends, and ultimately forecasting new cases for the future period of 14 days. Our methods and predicted consequences will aid in the prevention of COVID-19 pandemic infections.***

**Keywords:** COVID-19, Deep Learning, Autoregressive Integrated Moving Average (ARIMA), Artificial Intelligence (AI), Time series data.

## 1. Introduction

We also know that the outbreak of this Coronavirus may cause serious threats to the lives of individuals and society as a whole because there are no specific, practical and proven treatments to combat this virus [1]. Potential antiviral therapies, such as plasma transfusion, are critical and are being carefully implemented in the clinical sector [2] and by taking preventive measures such as hand washing and maintaining social distancing between individuals. Moreover, health officials have a special obligation to control chronic situations to prevent disease outbreaks [3]. The production of vaccines in a new effective way is necessary, despite the importance of the antiviral drug [4]. Neither a curative drug nor a prophylactic immunization has been significantly and effectively achieved. The outbreak is negatively affecting the economies of countries around the world. Unfortunately, many types of the virus have recently been discovered all over the world, and there is no cure for this disease yet. We know that the Kingdom of Saudi Arabia is one of the countries that has been negatively affected by this epidemic, so it is necessary to develop a special model to predict cases of infection with this coronavirus, and here is the role of artificial intelligence in effectively predicting the number of cases. In this research, time series data were used with the Autoregressive Integrated Moving Average (ARIMA) algorithm to predict the cases of Coronavirus in the Kingdom of Saudi Arabia in the short. Two measures were used to evaluate the performance of this algorithm. The paper aims to predict the number of coronavirus cases in Saudi Arabia using a deep learning algorithm. The rest of the paper is organized as follows: Section II provides a review of the literature. The process of the methodology is presented in the III section and explains the data set, pre-processing, and algorithm chosen and defines the performance measures used to evaluate the performance of the model, The IV section contains a discussion of the algorithm results. Finally, the V section demonstrates the conclusion.

## 2. Related Works

Many literary studies focus on the prediction of the Coronavirus in many parts of the world, as it is concerned with analyzing the spread of the virus in the current situation. In this section, we will discuss the most important literary publications and research contributions at the present time.

In this work, Hawas used only the RNN Deep Learning algorithm to predict the next 30 and 40 days of Brazil's confirmed cases[5].Tiwari et al. in 2020, created a machine learning model that forecasts the number of cases infected with the Coronavirus and recovery cases. Based on data from China, the number of deaths in the Indian state. Whereas the results indicated that the prediction for the virus reached its peak from April 2020, starting in the third week and ending in the fourth week. The Indian government benefits from this research, as it takes appropriate decisions related to mitigating the spread of Coronavirus and limiting its spread[6]. This study uses a clustering algorithm known as K-Means; this is an unsupervised machine learning algorithm. The aim is to gather COVID-19 data for various predictive variables and concepts. The model assisted in studying countries infected by the Coronavirus or are very likely to be affected by it shortly[7]. On the other hand, Mai

et al. developed a joint-based model in 2020, which includes a CNN in moreover a random tree, as well as a support vector machine (SVM), Random Forest (RF), besides the integrated perspective is multi-layered and with a result of a CT scan of the chest and non-imaging clinical information with the aim of early prediction of the infection of a Coronavirus patient. CNN ran on a computerized tomography image. Other algorithms for Coronavirus have been classified using clinical information that is invisible and a set of outputs from the CNN algorithm and various other algorithms to predict the patient's infection with the virus. Using these three models, the diagnostic tool can quickly detect patients with Coronavirus[8]. The study's datasets were obtained from Johns Hopkins University's CSSE, for cases that have been confirmed. Also, the algorithms of RNN, LSTM, and VAE were utilized by Zeroual et al.[9].Likewise, Hu et al. developed a model for epidemiological prediction. The model data was 15,384 cases, and 36,602 clinically confirmed and laboratory-confirmed cases, respectively. The data type was time series and was collected from Chinese news[10].Gozes et al. developed a model to diagnose Covid-19. The model's data was 157 patients. the data type was CT scan images of lungs and was collected from Testing Dataset Source at the hospital in Wenzhou, China, Chainz, El-Camino Hospital[11].Kırba and colleagues also executed a comprehensive comparative assessment of some of the available Use time-series algorithms for modeling and predicting the Coronavirus accumulated confirmed cases and the overall rates of increase in some European countries. The Auto-Regressive Integrated Moving Average (ARIMA) is one of the algorithms used, The Long-Short Term Memory (LSTM) and the Nonlinear Auto regression Neural Network (NARNN)[12].

## 3. The Method

The flowchart shown in Figure 1 is designed to demonstrate the methodology steps, which presents the sequence of the study which consists of the following:
Step1: Data Collection: the greeting of the history of COVID-19 in Saudi Arabia from the world health organization
Step 2: Pre-Analysis: some analysis steps were made on the data to discover hidden patterns.
Step 3: Processing: processing data and cleaning it from missing values or not important variables.
Step 4: Time Series Extracting: we extract a day, month and year from data and make it into separate columns to analyze it.
Step 5: data scaling: Scaling data is very important to get good performance when applying the model.
Step 6: Building Model: It is the core step in our study, where the model is built based on one algorithm; ARIMA .
Step 7: Predicting Outcome: For the model, we predict the outcomes of different cases in total Saudi Arabia.
Step 8: Result Analysis: After applying the model, we analyzed the result based on different measures.

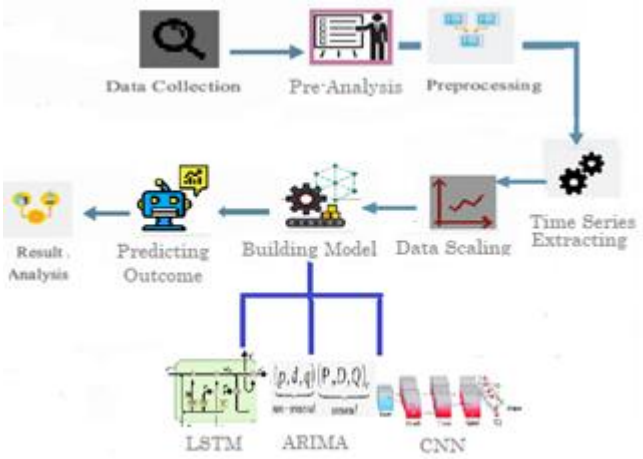


**Figure 1:** Methodology general steps

***Research Datasets***

Some governments have published a variety of publicly accessible data sources. Also, actual and real-time observations are available to be used for up-to-date real-time evaluations of the COVID-19 events forecasting by researchers of interest. Saudi Arabia's government is one of the first governments to make all data related to Coronavirus infections publicly available to all interested researchers. The data provided complete transparency to support scientific research related to this pandemic. Such datasets can be downloaded through the website of Saudi Arabia's Ministry of Health through <https://covid19.moh.gov.sa/>.
The time-series data of the overall Saudi Arabia cases of the COVID-19 and the cases associated with each city and region are being collected.

For the analysis that will be implemented in this thesis, there are three independently different time-series datasets gathered as follows:

1) Confirmed cases (newly infected cases).
2) Recovered cases.
3) Death/mortality cases.

Time Series is one of the data analysis types. It deals with time data. Our data depends on time and it is considered historical data. Figure 2 shows the case type of COVID-19 distributed from April 1, 2020, to May 31, 2021. There are two types of the number of data reporting: daily cases and cumulative cases.

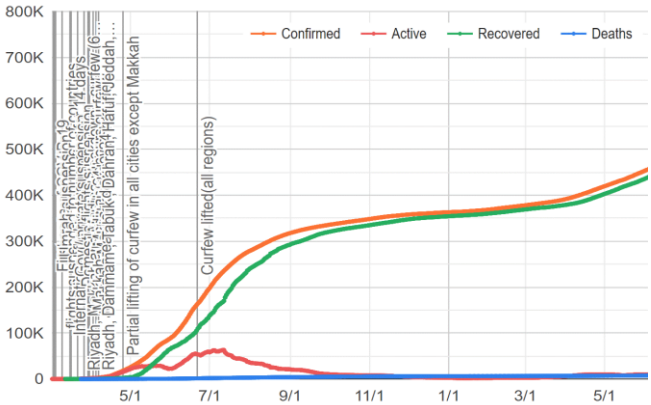


**Figure 2:** Data of COVID-19 cases collected in Saudi Arabia (Total Cases - Cumulative) from April 1, 2020, to May 31, 2021

***Data Preprocessing Steps***

The pre-processing of any data set is important to obtain better results and remove all the defects of the data set, as the pre-processing process of data in our study was as follows:

Step 1: In this step, we process the missing values and drop them.

Step 2: We sort the date in ascending order starting with April 1, 2020 .

Step 3: Extraction of day, month, and year from the date column to achieve analysis.

Step 4: Split weekend days than weekdays and apply pivoting on the Indicator column to get case type in different columns, then fill missing values with zero.

Step 5: Create a new dataset for each separate case daily and cumulatively for each city in Saudi Arabia.

***Selection of the deep learning algorithms***

We selected ARIMA Deep Learning algorithms, which will be explained briefly in the following section:

ARIMA is a forecasting algorithm that depends on the guess that previous values carry inherent information and can be used to predict future values. To predict $x_t$ using given past values, a predictive model will be based on the equation (1)

$$p(x_t \mid x_{t-1}, \ldots, x_1) \quad (1)$$

***ARIMA, Foundational Component***

An ARIMA model can be understood by outlining each of its components as follows:

Autoregression (AR): using the dependent relationship between an instance and some number of lagged instances.

Integrated (I): Using differencing of raw instances to make the time series stationary.

Moving Average (MA): using the dependency between an instance and a residual error from a moving average model actionable to lagged instances.

***The ARIMA model parameters***

Each component in ARIMA functions as a parameter with a standard notation. For ARIMA models, a standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

1) p is the order of the AR term.
2) q is the order of the MA term.
3) d is the number of differences.

***Performance evaluation metrics***

Several statistical performance metrics can be used to evaluate the predictive efficacy of established models. the goodness of fit , R-squared (R2)and mean absolute percentage error (MAPE)are the two metrics we have studied in this study.

## 4. Results Generation and Discussion

This section mainly focuses on the implementation phase and the results of this research. The model will be built on the basis of a deep learning algorithm and ARIMA. In this section, the algorithms have been applied in different experiments. Finally, the performance of the algorithm is calculated using six different metrics that fit the model. These measures are mean absolute error(MAE), normalized mean absolute error (nMAE) mean squared error(MSE), normalized root mean square error (nRMSE) R2 squared, mean absolute percentage error (MAPE).

***ARIMA model Deep Learning (Setup & Training)***

***Selecting the ARIMA parameters***

Table 1 lists the parameters used to construct the ARIMA model, along with their values. ARIMA (1,1,1) is a model with one Auto-Regressive (AR) term, one first-order difference, and one Moving Average (MA) term applied to the z variable, which indicates the linear trend in the data.

This section describes the ARIMA model's prediction for future COVID-19 cases. The forecasting period is 14 days long.

**Table 1:** Proposed scheme with parameters and their values.

| Method | Parameters | Values |
|---|---|---|
| ARIMA | (p, d, q) | (1,1,1) |

***Building, training, and testing ARIMA***

By fitting the training dataset, the ARIMA method is used to predict the number of new confirmed, recovered, and mortality cases on the test dataset. This section describes the ARIMA model's forecasting findings for future COVID-19 cases. The dataset is split into training and testing, and the test dataset's forecasted and actual values are reported.

With the forecasted confirmed, dead, and recovered cases of COVID-19 by using the ARIMA model, good forecasting efficiency and precision are attained. We summarize the different values of the error metrics in Table 2. From Table 2, the MAPE values of the confirmed, death, and recovered cases for the ARIMA model were found to be 21.4404, 16.2702, and 34.9793, respectively. By further looking at the error measures of the ARIMA model, it can be observed that ARIMA has an excellent performance in predicting the death cases, but poor performance on confirmed and recovered cases. For example, ARIMA's R2 values for predicting the number of death cases are 0.8792. For the confirmed cases, the values of the R2 are 0.1569, while for recovered 0.3967. The scatter plots that compare the actual and the predicted values of the three cases on the testing sets can be seen in Figures 3 ,4 and 5.

**Table 2:** Performance Evaluation of the test dataset forecasting using the ARIMA.

| Model | Predicted variable | MAPE | $R^2$ |
|---|---|---|---|
| ARIMA | Confirmed | 21.4404 | 0.1569 |
| | Deaths | 16.2702 | 0.8792 |
| | Recovered | 34.9793 | 0.3967 |

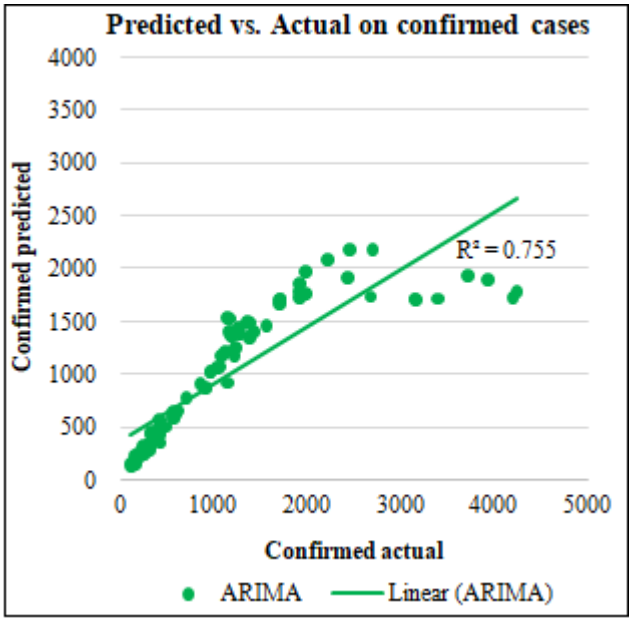


**Figure 3:** Scatter plot of the confirmed cases by the ARIMA on the testing dataset

As we see from the scatterplot, predicted confirmed cases show some higher-order polynomial relationship, rather than the first-order linear relationship, with the actual confirmed cases. The regression line in the plot, which has been drawn assuming a linear relationship and placed in the plot for comparison with other similar plots, is therefore not valid.

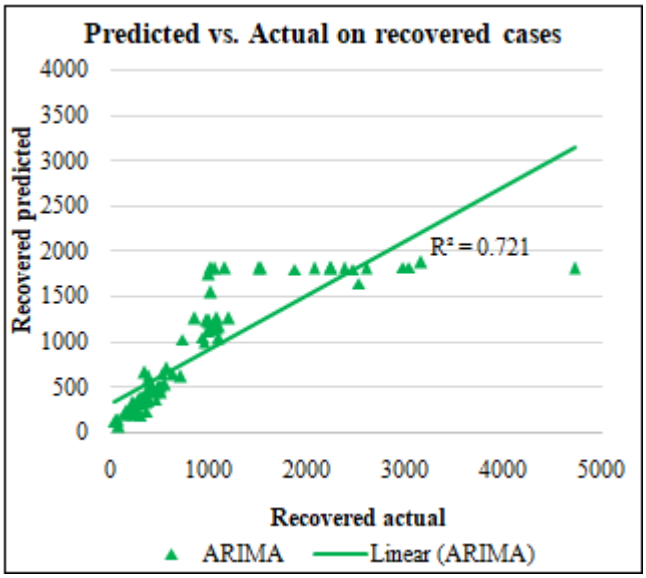


**Figure 4:** Scatter plot of the recovered cases by the ARIMA on the testing dataset

As in the performance in Confirmed cases, ARIMA prediction of Recovered cases shows some non-linear function relationship with the actual cases. When the number of cases is around 1,000 with small differences, ARIMA made predictions of a large range from 1,000 to 1,800. On the other hand, for actual cases from about 1,100 to about 5000, ARIMA makes an almost constant prediction of about 1,800 –1,900, making the overall scatterplot a stagewise appearance.

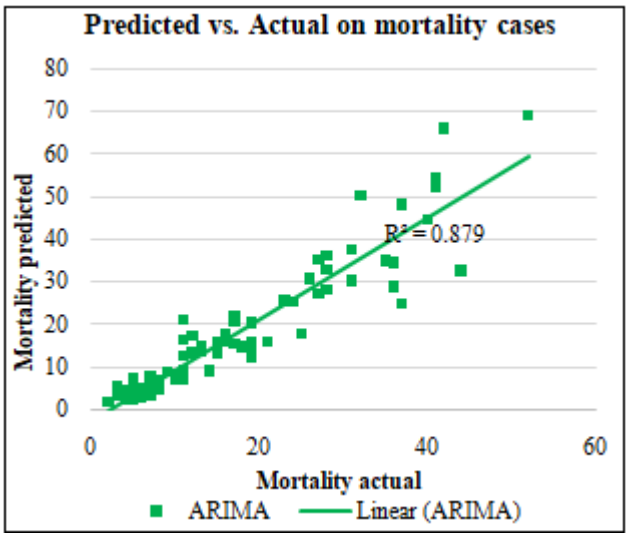


**Figure 5:** Scatter plot of the mortality cases by the ARIMA on the testing dataset

In contrast to the performances in Confirmed and Recovered cases, ARIMA's prediction in the mortality cases is very good, finally managing to establish a linear relationship between the actual cases and predicted cases. Also, the points in the scatterplot closely follow the regression line for the number of cases up to 30, before starting to show performance deterioration for the number of mortality cases beyond 30.

In the next three plots, we present the test data versus the predicted data for the three cases as line plots along a common index axis to see how closely one line is superimposed over the other as a visual indication of the prediction accuracy. As the plots suggest, there are larger spreads between the actual data and prediction by ARIMA in confirmed and recovered cases, indicating less accuracy in the prediction, while the accuracy in mortality cases is very good.

## 5. Conclusion

Based on a review of the medical community's research on the transmitting properties of COVID-19 disease, These methods play a critical role in the study and prediction of disease trend scenarios when applying ARIMA deep learning algorithm to model the time series data of daily new confirmed COVID-19 cases.We found that it can be observed that ARIMA has an excellent performance in predicting the death casesbased on MAPE measure, but poor performance on confirmed and recovered cases . This means that it can be used to predict deaths in the Kingdom of Saudi Arabia.

## References

[1] S. K. Kumaravel et al., "Investigation on the impacts of COVID-19 quarantine on society and environment: Preventive measures and supportive technologies," 3 Biotech, vol. 10, no. 9, pp. 1–24, 2020, doi: 10.1007/s13205-020-02382-3.
[2] X. Jiang et al., "Towards an artificial intelligence framework for data-driven prediction of coronavirus

clinical severity," Comput. Mater. Contin., vol. 63, no. 1, pp. 537–551, 2020, doi: 10.32604/cmc.2020.010691.

[3] C. C. Lai, T. P. Shih, W. C. Ko, H. J. Tang, and P. R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges," Int. J. Antimicrob. Agents, vol. 55, no. 3, p. 105924, 2020, doi: 10.1016/j.ijantimicag.2020.105924.

[4] H. Li, S. M. Liu, X. H. Yu, S. L. Tang, and C. K. Tang, "Coronavirus disease 2019 (COVID-19): current status and future perspectives," Int. J. Antimicrob. Agents, vol. 55, no. 5, p. 105951, 2020, doi: 10.1016/j.ijantimicag.2020.105951.

[5] M. Hawas, "Generated time-series prediction data of COVID-19′s daily infections in Brazil by using recurrent neural networks," Data Br., vol. 32, p. 106175, 2020, doi: 10.1016/j.dib.2020.106175.

[6] S. Tiwari, S. Kumar, and K. Guleria, "Outbreak Trends of Coronavirus Disease-2019 in India: A Prediction," Disaster Med. Public Health Prep., vol. 14, no. 5, pp. e33–e38, 2020, doi: 10.1017/dmp.2020.115.

[7] L. Chaudhary and B. Singh, "Community Detection using Unsupervised machine learning technique on COVID -19 dataset," Soc. Netw. Anal. Min., pp. 1–9, 2021, doi: 10.1007/s13278-021-00734-2.

[8] X. Mei et al., "Artificial intelligence–enabled rapid diagnosis of patients with COVID-19," Nat. Med., vol. 26, no. 8, pp. 1224–1228, 2020, doi: 10.1038/s41591-020-0931-3.

[9] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, "Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study," Chaos, Solitons and Fractals, vol. 140, p. 110121, 2020, doi: 10.1016/j.chaos.2020.110121.

[10] Z. Hu, Q. Ge, S. Li, L. Jin, and M. Xiong, "Artificial intelligence forecasting of covid-19 in China," arXiv, pp. 1–20, 2020, doi: 10.18562/ijee.054.

[11] O. Gozes, M. Frid, H. Greenspan, and D. Patrick, "Rapid AI Development Cycle for the Coronavirus ( COVID-19 ) Pandemic : Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis Article Type : Authors : Summary Statement : Key Results : List of abbreviati," arXiv:2003.05037, 2020.

[12] İ. Kırbaş, A. Sözen, A. D. Tuncer, and F. Ş. Kazancıoğlu, "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches," Chaos, Solitons and Fractals, vol. 138, 2020, doi: 10.1016/j.chaos.2020.110015.