

ARIMA Model Estimation Based on Genetic Algorithm for COVID-19 Mortality Rates

Deif, Mohanad A.; Solyman, Ahmed A.A.; Hammam, Rania E.

Published in:
International Journal of Information Technology and Decision Making

DOI:
[10.1142/S0219622021500528](https://doi.org/10.1142/S0219622021500528)

Publication date:
2021

Document Version
Author accepted manuscript

[Link to publication in ResearchOnline](#)

Citation for published version (Harvard):
Deif, MA, Solyman, AAA & Hammam, RE 2021, 'ARIMA Model Estimation Based on Genetic Algorithm for COVID-19 Mortality Rates', *International Journal of Information Technology and Decision Making*, vol. 20, no. 6, pp. 1775-1798. <https://doi.org/10.1142/S0219622021500528>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please view our takedown policy at <https://edshare.gcu.ac.uk/id/eprint/5179> for details of how to contact us.

ARIMA MODEL ESTIMATION BASED ON GENETIC ALGORITHM: A CASE STUDY OF COVID-19 MORTALITY RATES *

MOHANAD A.DEIF †

*Dept. of Bioelectronics, Modern University of Technology and Information (MTI)
University, Egypt
Mohand.Deif@eng.mti.edu.eg*

RANIA E. HAMMAM

*Dept. of Bioelectronics, Modern University of Technology and Information (MTI)
University, Egypt
Rania.Hammam@eng.mti.edu.eg*

AHMED A. A. SOLYMAN

*Dept. of Electrical and Electronics Engineering, Istanbul Gelisim University,
Turkey
aaasahmed@gelisim.edu.tr*

Received (Day Month Year)

Revised (Day Month Year)

Communicated by (xxxxxxx)

This article presents a forecasting model for the mortality rates of COVID-19 in six of the top most affected countries depending on the hybrid Genetic Algorithm and Autoregressive Integrated Moving Average (GA-ARIMA). It was aimed to develop an advanced and reliable predicting model that provides future forecasts of possible confirmed cases and mortality rates (Total Deaths per 1 million Population of COVID 19) that could help the public health authorities to develop plans required to resolve the crisis of the pandemic threat in a timely and efficient manner. The study focused on predicting the mortality rates of COVID-19 because the mortality rate determines the prevalence of highly contagious diseases. The Genetic algorithm (GA) has the capability of improving the forecasting performance of the ARIMA model by optimizing the ARIMA model parameters. The findings of this study revealed the high prediction accuracy of the proposed (GA-ARIMA) model. Moreover, it has provided better and consistent predictions compared to the traditional ARIMA model and can be a reliable method in predicting expected death rates as well as confirmed cases of COVID-19. Hence, it was concluded that combining ARIMA with GA is further accurate than ARIMA alone and GA can be an alternative to find the parameters and model orders for the ARIMA model.

Keywords: COVID-19 Forecasting, ARIMA Model; Genetic Algorithm; Time-series; infection disease; pandemic

1. Introduction

Coronavirus disease (COVID-2019) is a severe ongoing pandemic threat that began in December 2019 in the capital city of China's Hubei province, Wuhan ¹⁻³. It has spread quickly across the world in a few months, which has led to an increase in the number of infected people. The number of deaths attributed to the disease has also increased, especially those with chronic diseases and the elderly ^{4 5}.

COVID-19 has fatal features as it is a very active and quickly spreading disease. Throughout the Chinese New Year immigration, in early and mid-January 2020, the virus spread to other provinces in China. The World Health Organization (WHO) ⁶ announced that the disease infection expanded to other different countries by overseas travelers. To prevent further spread, strong precautionary measures have been taken; however, the numbers of infected people are increasing worldwide even after lockdown all around the globe. Disastrously, as of April 15, 2020, there were 123,010 deaths, and about 2 million cases of infection were confirmed globally. As there is no known treatment for this virus version until now, effective arrangement for healthcare services and infrastructure is urgently needed to control the disease spreading rate. That is why modeling of approved infected cases, daily and future deaths, and estimates of probable new cases of infection are important for the health system.

To address the epidemic and planning for future health resources and materials demanded, mathematical tools and statistical models are needed to predict the cases in the short and long term future. Estimations of the expected risk of disease are vital for public health authorities to plan the medical services and other gears required to resolve the crisis in an efficient and timely manner. Such assessments can also suggest the severity and nature of treatments expected to reduce the outbreak ^{7 8}.

In the past, various mathematical approaches had been implemented with high accuracy for diverse purposes of prediction. Recently, several computational approaches have been employed, such as Time Series models ^{9 10}, Multivariate Linear Regression ¹¹, Gray Forecasting model ^{12,13}, Back-propagation Neural Network ¹⁴⁻¹⁶, Simulation models ^{17,18}, Automatic Regressive Integrated Moving Average (ARIMA) ^{19,20}, Long-Short Term Memory (LSTM) which is a type of Recurrent Neural Networks (RNN) ^{21,22} and Nonlinear Autoregressive Artificial Neural Networks (NARANN) ²³ to forecast epidemic cases. There are numerous common variables influencing epidemics and preventing them. Of this cause, the general distribution of the epidemic is characterized by patterns and uncertainty. Therefore, the computational methods listed are inadequate for the study of disease spreading randomness, and the simulations are challenging to generalize.

Automatic Regressive Integrated Moving Average (ARIMA) can help in estimating the development of this prevalent disease. The ARIMA model, known also as the Box-Jenkins methodology ²⁴, is employed for forecasting and analysis ²⁵⁻²⁷. It has been proven as an effective model in the health field as well as in numerous fields due to its capability to clarify the data set, rapid development, and construction simplicity ²⁸. ARIMA models were used effectively to forecast the rate and frequency of influenza mortality. Moreover, ARIMA modeling is broadly used for the estimation of epidemic diseases like dengue fever, tuberculosis, and hemorrhagic fever with renal syndrome.

ARIMA models are effective in simulating the spatial correlation structure of the time series due to varying patterns, periodic seasonals, and irregular deformities in the time series. The end-user has a relatively simple explanation since there is little statistics or mathematics in the ARIMA methods. The end-user can, thus, have an insight into the development of the estimation method and can depend more on this model in strategic decisions.

LSTM and NARANN models increase the computing complexity as compared to the ARIMA model because there are more hyperparameters that are required to be tuned. This makes the computational effort required by the ARIMA model reach high quality results with less effort than that exerted by the LSTM and NARANN models for the same results. Moreover, LSTM are prone to problems such as exploding and gradient vanishing. LSTM and NARANN works better in dealing with huge amount of data while the ARIMA model is more efficient with small datasets²⁹⁻³¹. On the other hand, ARIMA is simple to implement since it requires fewer parameters to tune (only three).

The Genetic Algorithm (GA) is a total optimization technique that is always employed to conquer precise estimation based on the ARIMA model. As the coronavirus has become a serious problem among many countries, therefore, it is necessary to assess the situation of the COVID-19 pandemic and forecast the evolution of prevalence rates and mortality rates, especially in the countries most affected by the virus. Consequently, the aim of this study is to develop an improved prediction model that can be applied to predict mortality rates (total deaths per 1 million population time series of COVID 19) in the USA, Brazil, Russia, United Kingdom, Spain and Italy which have the highest incident cases of COVID 19 based on the WHO regions. For this purpose, a hybrid algorithm (GA-ARIMA) was proposed to attain a high efficiency of forecasting. Daily confirmed COVID-19 death rates for the six largest countries were recorded from January 1, 2020 to May 25, 2020 as reported by the World Health Organization and the Worldometer website³² which are the main data sources for this study. To evaluate the proposed model, several diagnostic tests of residual errors were applied. Additionally, the performance was compared based on previous static and dynamic forecasting. After that, the available data were used to forecast the future mortality rates for each country one month later from 1 October 2020 to 1 November 2020. Moreover, the proposed hybrid model was employed to predict the total approved cases of COVID-19 in Spain, France, and Italy, which were compared to the real and forecasted values presented by¹⁹.

2. Contribution

This study contributed to the development of an advanced (GA- ARIMA) model for the forecasting of the daily Total Deaths per million Population time series of the COVID-

19. The study was applied in six of the top most affected identified countries according to the WHO. The proposed model was then employed to estimate the expected future death rates in the six countries for three months later. The results revealed the high accuracy of prediction of the proposed forecasting model as compared to the classical ARIMA model and it can be a reliable method in predicting future forecasts of expected mortality rates as well as confirmed cases of COVID-19. Therefore, it was concluded that a hybrid method is more accurate than a distinct method and GA can be an alternate method in calculating orders and parameters for the ARIMA model instead of using the traditional statistical tools employed in previous works ¹⁹. These traditional methods take much time and effort in comparing the performance of various ARIMA models in order to select the best fit model. No doubt that this advanced hybrid model could help the healthcare authority to deliver quality services for future patients.

3. Methodology

3.1 Dataset

Total daily deaths due to COVID-19 were recorded for the top six countries that reported confirmed cases during the period from January 1, 2020 to May 25, 2020 using data collected from both the World Health Organization and the Worldometer websites ³².

3.2 Autoregressive Integrated Moving Average (ARIMA) model

Autoregressive Integrated Moving Average (ARIMA) model is a time series model, which depends on a sequence of time-ordered data points ^{33 34}. The study of the time series helps to expose accurate and relevant data and uses them to forecast future series values ^{35–37}. The major components of a Time Series data set include trends, seasonal and irregular, or random irregularity (Noise).

Box and Jenkins presented the ARIMA model in 1970 ²⁴. ARIMA is one of the most commonly employed time series models, which takes into account changing patterns, periodic variations, and random fluctuations in the time series model. ARIMA is appropriate for all data types, as well as seasonality, trends, and cyclicity. Moreover, it is also robust and efficient when analyzing the spatial dependency structure of a time series.

The main components of the ARIMA models are Auto- Regression (AR), Integration (I), and Moving Average (MA). AR is used to forecast potential values using a linear combination of past values. To be able to use ARIMA on the data for the prediction process, data has to be shifted from non-stationary to stationary by applying difference transformation method. Finally, MA takes advantage of predicting errors in earlier stages to improve the overall forecasting process. AR, Integration, and MA are typically defined using p, d, and q which are three integers. That is why the ARIMA model can be written as ARIMA (p; d; and q) in short form. AR (p) and MA (q) models can be specified as ³⁸:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t \quad (1)$$

$$y_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_p \epsilon_{t-p} + \epsilon_t \quad (2)$$

The terms in the equations are identified as follows:

c	The time series (data) mean value
ϕ_1	1st AR-lag coefficient
θ_1	1st MA-lag coefficient
y_{t-1}	The variable y First Lag
ϵ_{t-1}	The prediction errors first lag
p	AR-Lags order
q	MA-Lags order
ϵ_t	White noise

Consider the difference transformation (I) order is one, ($d = 1$), then y_t' can be written as:

$$y_t' = y_t - y_{t-1} \quad (3)$$

For model AR (1), ϕ_1 would meet the limitation ($-1 < \phi_1 < 1$), also θ_1 in model MA(1) would meet the limitation ($-1 < \theta_1 < 1$). Eventually, equations (1, 2 and 3) can be merged in order to obtain the general ARIMA equation (p, d, and q) model as follows:

$$y_t' = c + \phi_1 y_{t-1}' + \dots + \phi_p y_{t-p}' + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (4)$$

Consider the difference transformation (I) order is one, ($d = 1$), then y_t' can be written as:

$$y_t' = y_t - y_{t-1} \quad (3)$$

For model AR (1), ϕ_1 would meet the limitation ($-1 < \phi_1 < 1$), also θ_1 in model MA(1) would meet the limitation ($-1 < \theta_1 < 1$). Eventually, equations (1, 2 and 3) can be merged in order to obtain the general ARIMA equation (p, d, and q) model as follows:

$$y_t' = c + \phi_1 y_{t-1}' + \dots + \phi_p y_{t-p}' + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (4)$$

3.3 Genetic Algorithm (GA)

GA is a multi-objective optimization technique based on the theory of survival of the fittest and then applies this optimization to a variety of population forecasting models. The GA produces chromosome populations as potential options for predicting optimal forecasting³⁹. This algorithm is stable, versatile and simple to adjust, as it can be divided

into several phases: initialization, evaluation, selection, crossover, mutation, update, and implementation. The evaluation (fitness function) phase generates the source for a new chromosomes population. The new population is shaped by means of certain Genetic operators, as for examples crossover and mutation^{40,41}. The GA fitness function is derived from the ARIMA predicting model. An automated optimization GA technique evades the weakness of the ARIMA model by estimating the optimal parameters for forecasting⁴². The GA optimization can be employed to attain precise forecasting through automated performance improvements of the ARIMA model. Moreover, the GA can assess the forecasting precision using various fitness functions based on statistical models.

3.4 The proposed method: hybrid algorithm (GA-ARIMA) approach :

The philosophy of GA is founded on the best individual's survival. Over generations, the fitness function increases until the optimum solution is finally achieved. Initially, the leading population is produced randomly. Every component of the population is encoded as a real number in the range of [0-1]. After choosing additional appropriate individuals according to the fitness function, evaluation for the mating pool, crossover and mutation operators are applied respectively to produce new offspring. The same process is repeated again to a predefined number of iterations. figure (1) shows the Flow chart of the proposed (GA-ARIMA) model. The procedures taken to implement the forecasting model are presented in the following steps:

Step 1: Check Stationarity:

Time-series data with trends or with seasonality are not stationary because it will affect the value of the data at different times. The stationarity of a time series is essential due to the ease of getting precise estimates⁴³. Hence, a very important step before using the ARIMA model is inspecting the stationarity of the data by applying a stationary test. This test result defines the d order to be zero ($d = 0$), if the data is stationary. On the other case when the data is non-stationary, the first difference is taken and the resultant differenced data is verified once more. This process is repeated until the data is converted to be stationary and d is set to the number of times the differencing transformation is applied. The optimum order of d is the minimum differencing transformations needed to become stationary time series (TS)^{38 44}. There are many stationarity tests that can be employed like: Phillips Schmidt, Kwiatkowski and Shin (KPSS) test, Auto-Correlation Function (ACF), and Augmented Dickey Fuller (ADF) test.

Dickey-Fuller test⁴⁵ was applied in this work to check if the data are non-stationary or stationary. The data needs to be stationary in order to be applied on the ARIMA model, i.e. the null hypothesis of stationarity should not be rejected. After applying the Dickey-Fuller Test (DFT) in (5), if the data are non-stationary then $p = 1$, and if the data are stationary then $p < 1$.

$$DFT(x_t) = \alpha + px_{t-k} + \varepsilon_t \quad (5)$$

Where ;

- α the estimated constant value of the TS data
- p the hypothesis is either $p = 1$ or $p < 1$
- t time $\{1, \dots, k\}$
- ε the white noise of the TS data

Then the Autocorrelation (ACF) graphs for the time series were plotted to visualize the data. The ACF graph defines if the previous values are related to the next values in the series.

Step 2: Transformation from Non-Stationary to Stationary Data:

After checking the seasonality, we found that all the time series data had an obvious seasonality pattern (increasing trend). Consequently, the first-order difference was taken to transform the data from non-stationary to stationary. The equation of successive differences at the d_{th} difference in X_t is expressed in equations (6) and (7):

$$\Delta^d X_t = (1 - B)^d X_t, \quad (6)$$

where d is the different order, and B is the backshift operator. The consecutive difference at one-time lag equals:

$$\Delta^1 X_t = (1 - B)X_t = X_t - X_{t-1} \quad (7)$$

Step 3: Encoding

The parameters AR (p) and MA (q) were represent as a chromosome and the dimension of each chromosome L equals $(p + q)$. Each chromosome consists of real value coding and the values between (0 and 1). For example, an ARIMA (3, 2) chromosome may be shown as:

AR (3)			MA (2)	
0.261	0.451	0.632	0.840	0.123

Step 4: Initial population

The initial population is randomly nominated. Population size is one of the significant parameters in the GA which is the number of chromosomes in each generation; this parameter increases the efficiency of the algorithm. $p = 50$ was randomly selected as there are no standards to specify the size.

Step 5: Fitness function

The fitness function is based on the ARIMA model for predicting the time series (TS) as shown in equation (4). The fitness function $f(x)$ was computed for each chromosome

produced in the initialization step. It was aimed to perform predictions based on ARIMA model with minimum error.

Step 6: Selection

Selection is a major part of the GA to find the best chromosomes. The fittest chromosomes have a higher probability to be selected for the next generation. The roulette wheel is used for probabilistic selection, using the formula:

$$P(x_i) = \frac{f(x_i)}{\sum_{j=1}^N f(x_j)} \quad (8)$$

Step 7: Crossover and Mutation

The crossover drive generates the next generation new offspring by sharing information among two-parent chromosomes, chosen at random. The mutation drive is deployed to set new genetic information with gene modification of a selected chromosome having a crossover probability P_c and a mutation probability P_m . It was randomly assumed $P_c = 80\%$ and $P_m = 2\%$.

Step 8. Stopping criteria

Two different stopping criteria were used; the first criterion is the predefined maximum number of iterations. The other termination criterion is decided due to the enhancement of the fitness function. After a specified number of iterations, if there is no change in the fitness property of the last improved solution, then the algorithm is stopped. The maximum number of iterations was randomly selected to be equal to 2000.

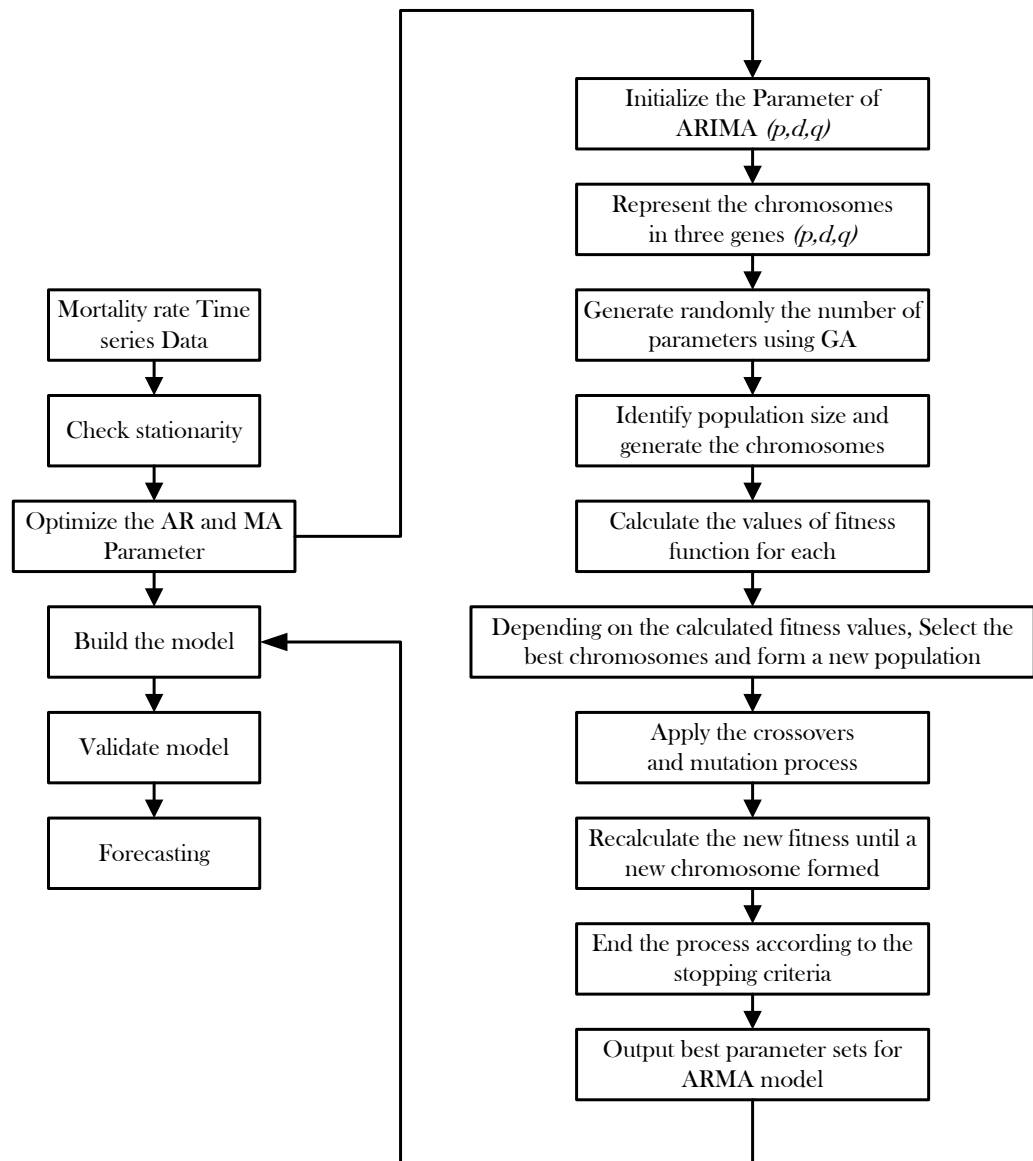


Fig. 1. Flow Chart of GA-ARIMA forecasting model

3.5 Model Evaluation

After fitting the GA-ARIMA model, several diagnostic tests were performed for the remaining errors. Density plots and Quantile-Quantile (QQ) plots for the residuals were explored to check the normality of the distribution of residual errors of the proposed predictor. The predictive accuracy of the model was evaluated by choosing the minimum value of Root Mean Square Error (RMSE) for the time series of all countries.

$$RMSE = \sqrt{\frac{1}{n} e_t^2} \quad (9)$$

Where $e_t = y_t - f_t$ Forecast Errors

y_t Actual Values

f_t Forecast Values

n Number of Forecasts

3.6 Implementation

The proposed forecasting algorithm (GA-ARIMA) was implemented using Jupyter Notebook version 6.0.0 along with Python version 3.4.0. Also, the analysis of the time series was performed with Minitab 18 Statistical software.

4. Results and Discussion

The total Deaths per 1 million population time series of COVID-19 for the six largest countries were compared in figure (2). By examining the time-series graphs, the trend is seen. It was clearly seen that there is an upward trend in the death rates for all countries except Russia. Besides, there were higher death rates in Spain and Brazil.

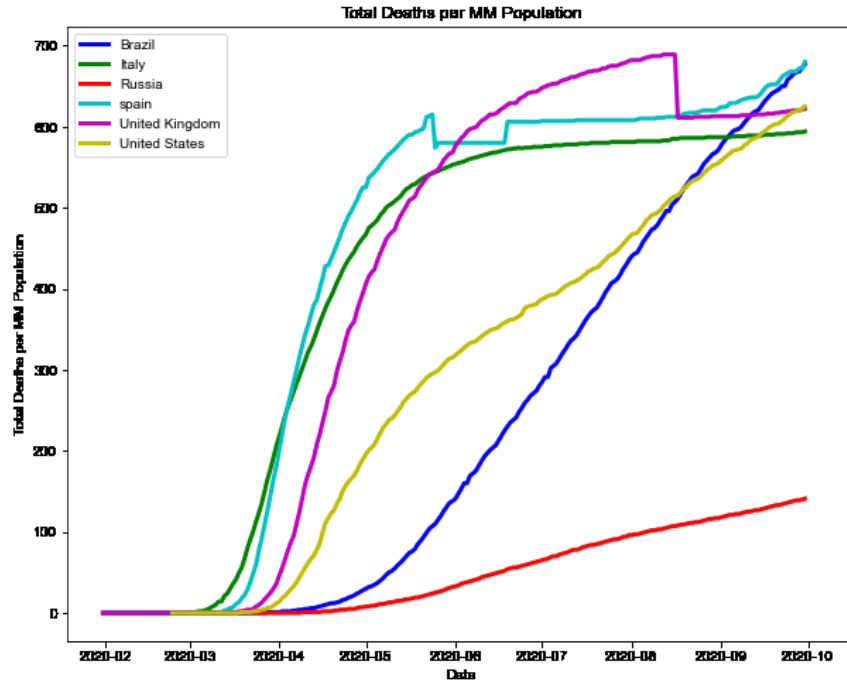


Fig. 2. Total Deaths per 1 million Population time series of COVID-19

Furthermore, the Autocorrelation (ACF) graphs shown in figure (3) were performed to visualize the data more clearly and check the stationarity of the time series. The autocorrelation plot has a 95% confidence band which allows the detection of the non-zero correlation. The extended lines in the plot show statistically crucial autocorrelations for the COVID-19 time series data for the six countries. The ACF graphs in figure (3) revealed that the series was not stationary since many delays exceed the confidence limits and the autocorrelations were decreased very slightly. This result confirms the trend observed in the time series plot in figure (2).

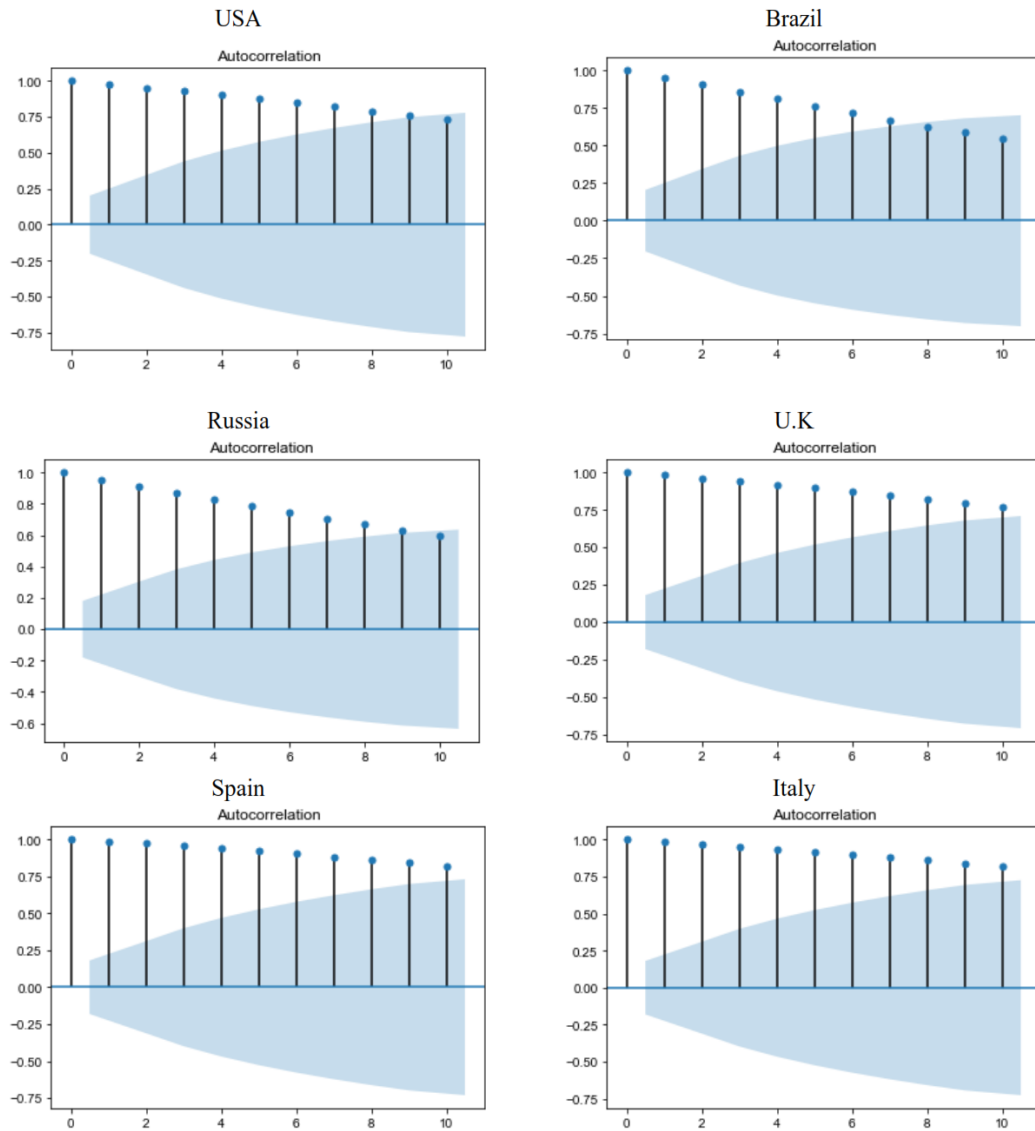


Fig. 3. Autocorrelation coefficients before adjustment (x-axis denotes the time lags).

In this case, the transformation step was applied to smoothen the data by eliminating variations in the level of the TS and, therefore, removing (or decreasing) the trend and seasonality. The ACF graphs after the transformation step were shown in figure (4). It was noted that all the time series turned to stationary after performing the transformation step

(second difference step). The second-order difference was applied because the trends of all series were not removed after the first order difference step.

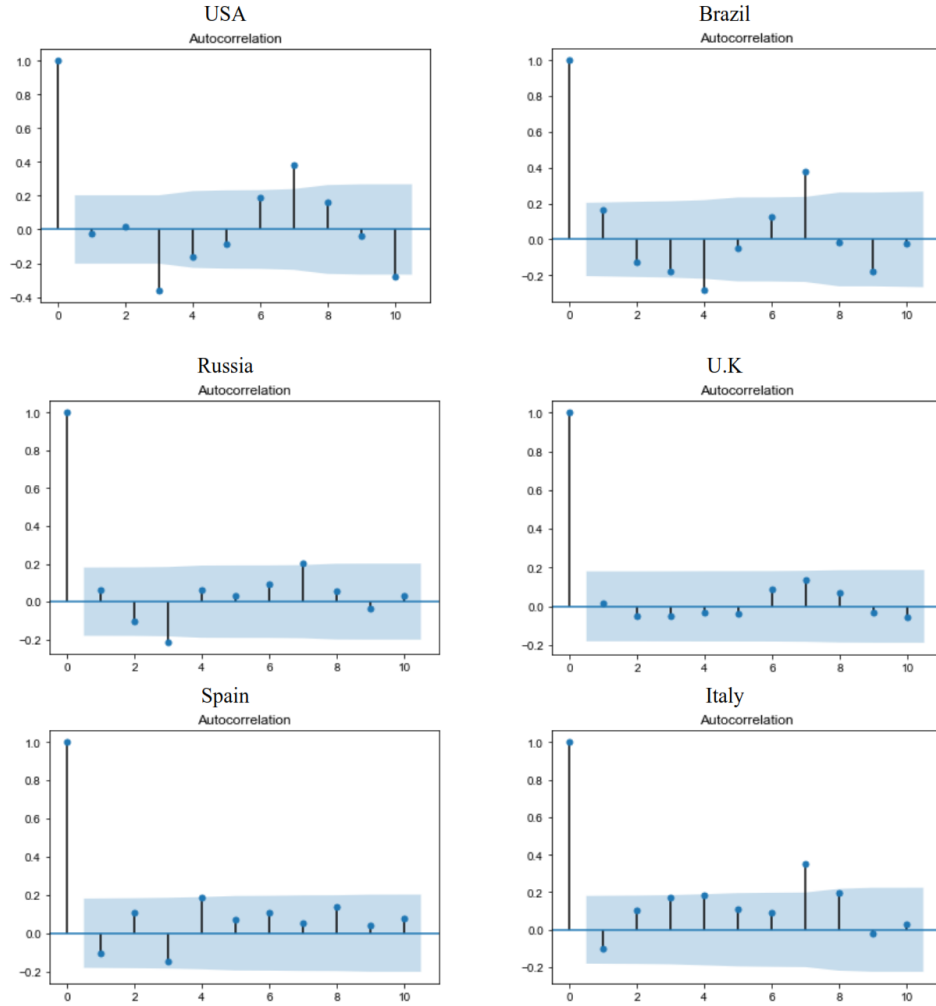


Fig. 4. Autocorrelation coefficients after seasonal adjustment (x-axis denotes the time lags).

The next step was the evaluation of the proposed (GA-AMIRA) model to check the normality of the distribution of residual errors. The residual error is the difference between the predicted and real values. figure (5) shows the Quantile-Quantile (QQ) plot of the residual errors of the (GA-AMIRA) model. The QQ plot Ordered distribution of the residual errors were plotted and compared to samples taken from a standard normal

distribution $N(0, 1)$ with a mean of 0 and standard deviation of 1. It was clear from the figure that the data are closely clustered around the red line (Normal distribution line) and a slight deviation of residuals from the straight line can be observed. This indicates that the errors are somewhat near to normal with a few outliers.

To further assess the normality of the residual errors of the forecasting model, Histograms and Probability Density Functions Plot estimated via Kernel Density Estimation (KDE) were constructed and compared with the Normal Gaussian distribution $N(0, 1)$. From figure (6), it can be observed that the KDE curve closely tracks the $N(0, 1)$ curve and the distribution does have a Gaussian shape which is a good indication that the residual errors of the proposed forecasting model are normally distributed, also the residual histogram backs up this result.

Table 1. The goodness of fit criteria and coefficients for the time series GA-ARIMA model

<i>Country</i>	<i>ARMA Model</i>	<i>Standard Error</i>	<i>p-value</i>
United States	AR(1)	0.053	0.000
	MA(1)	0.057	0.585
Brazil	AR(1)	0.057	0.000
	MA(1)	0.136	0.000
Russia	AR(1)	0.005	0.000
	MA(1)	0.064	0.000
United Kingdom	AR(1)	0.057	0.000
	MA(1)	0.136	0.000
Spain	AR(1)	0.081	0.000
	MA(1)	0.139	0.000
Italy	AR(1)	0.020	0.000
	MA(1)	0.104	0.000

To evaluate the goodness of the best fit model which was specified as ARIMA (1,2,1), the p values, and the Standard Errors for the parameters were summarized in table (1). It was clear that the AR (1) and MA (1) parameters had a p value of 0.000 in all countries except in the United States where MA (1) had a p value of 0.585. This implies that the parameters of the model are statistically significant because p values are less than 0.05.

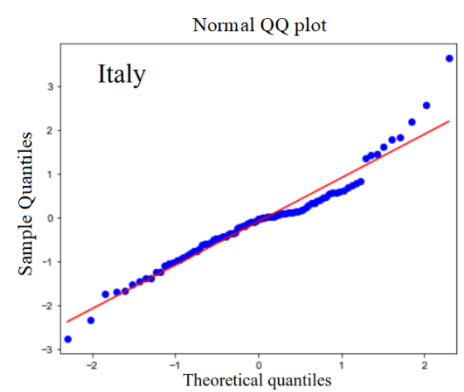
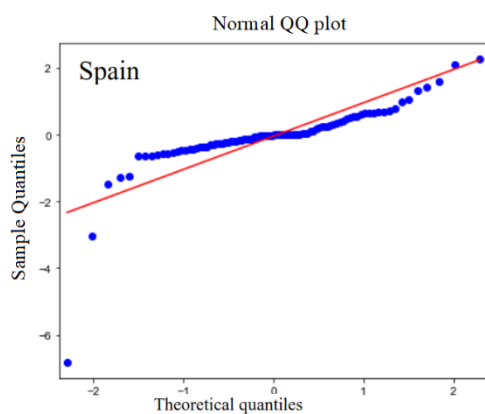
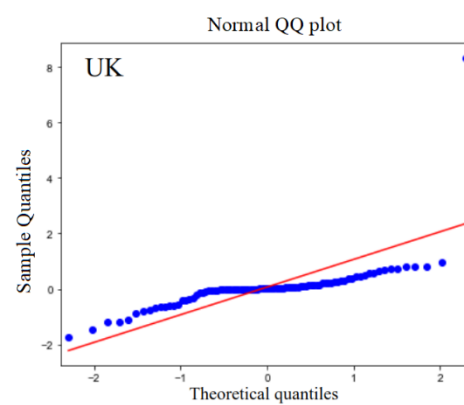
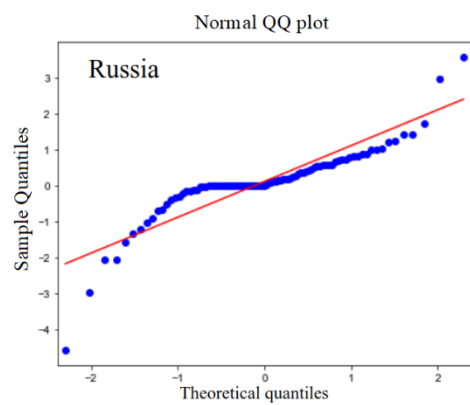
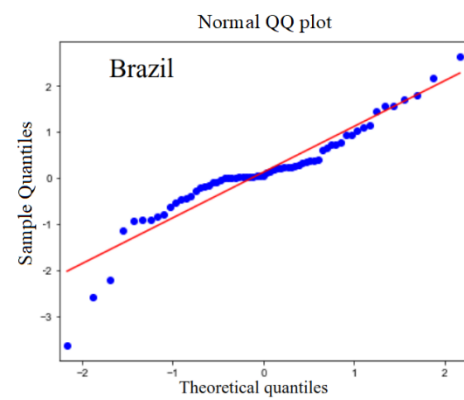
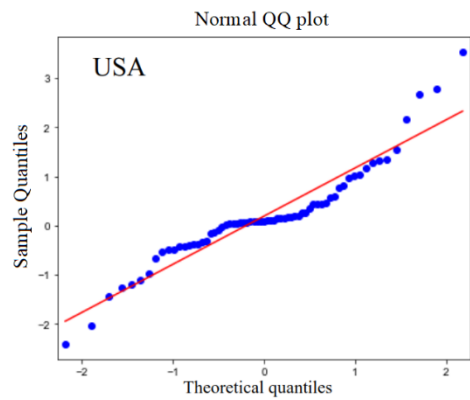


Fig.5. Quantile-Quantile plot of the residual errors of the (GA-ARIMA) Model

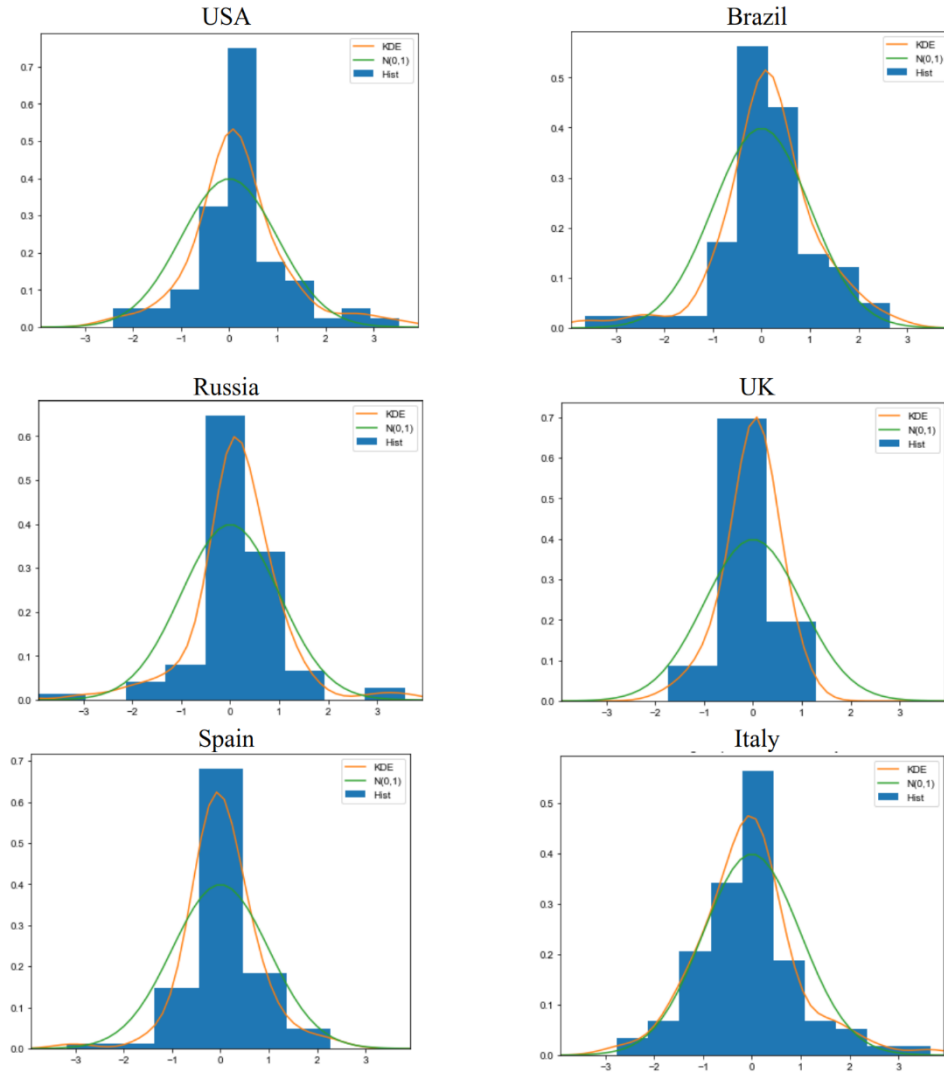


Fig. 6. Histogram with Probability Density Function Overlay compared with the Normal Distribution

The validation step of the (GA-ARIMA) model is very essential to test the forecasting ability of the model. Firstly, the prediction model was compared with other counterpart models such as the traditional (ADAM-ARIMA) model which employs the ADAM optimization method and (PSO-ARIMA) model which depends on the practical swarm optimizer. The RMSE was considered as an evaluation metric. Table (2) summarized the forecasting accuracy results of the proposed model and the selected benchmark models (Adam -ARIMA) and (PSO -ARIMA). The RMSE values for all models were calculated for prediction of mortality rates starting from 25 April 2020 to 25 May 2020 for all

countries. It was noticed that the RMSE values for the (GA-ARIMA) model were less than the counterpart models for all countries which reveals the high accuracy and better forecasting ability of the proposed model. These results were attributed to the GA optimizer search behavior. The GA helps in solving complex data and intended to discover the best forecasting solution using automated optimization of the ARIMA model parameters. This implies that GA is a powerful mathematical technique to improve the predicting accuracy of the predictor model. In addition, since the COVID-19 pandemic may occur at discrete intervals of time all over the world, hence GA is the most suitable optimizer as it has proven to be superior to PSO in discrete optimization^{46 47 48}.

Secondly, to check the robustness of the forecast a comparison between the Dynamic and Static forecasting performance of the (GA-ARIMA) model was done. It is worth noting that the dynamic forecasting depends on the previous forecasting data to predict the future while the Static (one step ahead) depends on the previous actual data to predict the future. The results of the comparison as shown in table (3) indicated the lower values of RMSE of Static forecasting as compared to the Dynamic for all countries and there was a significant difference between the values.

Table 2: Forecasts accuracy comparison between the proposed and benchmark model.

	Adam –ARIMA	PSO –ARIMA	GA -ARIMA
	RMSE	RMSE	RMSE
United States	5108	5254	3644
Brazil	5352	5143	3888
Russia	5596	5987	4132
United Kingdom	5840	5944	4376
Spain	6084	5843	4620
Italy	6328	5612	4864

Figure (7) showed another comparative study between Static and Dynamic forecasting using the (GA-ARIMA) algorithm to predict the time series for the death rates with 95% confidence limits from 25 April 2020 to 25 May 2020. The observed differences between the two performances for all countries were clearly distinguished. It was noticed that the static forecasting line almost matches the real value line which indicates its accurate forecasting ability and also it fits well within the confidence intervals of the proposed GA-ARIMA algorithm forecast. These results confirm the previous results of the better performance of Static forecasting. However, it was found that the overall forecasts for the time series of the USA and Russia were accurate even when the dynamic forecasting was applied, that is because the curves of the forecasted values of the time series fit well with the real value line. Therefore, the Static Forecast was used to predict further forecasts in this study.

Table 3: Summary of Comparative Static and dynamic forecasting performance

	<i>Static</i>	<i>Dynamic</i>
	RMSE	RMSE
United States	1.69	8.52
Brazil	0.97	17.38
Russia	0.14	0.99
United Kingdom	10.79	28.31
Spain	8.39	74.24
Italy	1.45	87.53

In further investigation of the prevalence and death rates of the COVID 19, the proposed (GA-ARIMA) model was employed to predict the mortality rates for all countries for 3 months later, starting from 1 October 2020 to 1 November 2020. Figure (8) showed the predicted values of the death rates in the future. The model predicted a significant rise in the mortality rates of COVID 19 in United States, Brazil and Russia while a slight increase in the rates was observed in Italy. However, it was noticed that the mortality rates for United Kingdom and Spain remained constant for a certain interval of time and then it increased again with time. In this study, it was very important to concentrate on the mortality rates of COVID 19 because the mortality rates determine if a highly infectious disease becomes a public concern. The rise in the number of death amongst the studied countries is a sign that the number of infected people will increase, hence, governments will face disastrous problems, since increased number of patients will be a burden on health care systems

To compare the actual and forecasted values of the (GA-ARIMA) model, the model was employed to estimate the total confirmed cases of COVID-19 within the period of 16 April to 25 April in Italy, Spain, and France and compared it to a previous work ¹⁹ in which the author had previously predicted these values depending on the classical ARIMA model. The forecasting ARIMA model presented by Z. Ceylan ¹⁹ was taken as the reference model in the comparative study. The predicted forecasts of the two models against the real numbers of confirmed cases were compared in figure (9). It was observed that the predicted values of the hybrid proposed model were closer to the real data than the reference model in all countries, which reveals the high precision of the (GA-ARIMA) model in forecasting. These results are attributed to the genetic algorithm which is a powerful mathematical technique in improving the forecasting accuracy of the ARIMA model. It has the capability of finding the best forecasting by choosing the optimal ARIMA model parameters. Also, it provides effective solutions faster than other classical approaches. The comparison between the total numbers of confirmed cases of COVID-19 predicted by the two models (the hybrid model and the reference model) and the actual values during the identified period was further demonstrated in table (4).

Table (4): Prediction of the total number of confirmed cases of COVID-19 comparison between the proposed model and reference model

<i>Date</i>	<i>Spain</i>			<i>France</i>			<i>Italy</i>		
	Real	ARIMA_GA Model	Reference Model	Real	ARIMA_GA Model	Reference Model	Real	ARIMA_GA Model	Reference Model

16-Apr	184948	181432	175866	147091	147280	106312	168941	168031	165891
17-Apr	190839	189667	179009	149130	150970	110090	172434	171965	169294
18-Apr	191726	195094	182270	149149	153636	113869	175925	176591	172698
19-Apr	198674	195329	185455	154097	155341	117648	178972	179420	176101
20-Apr	200210	203217	188689	156480	157340	121427	181228	182232	179504
21-Apr	204178	203576	191892	159297	160086	125205	183957	183870	182907
22-Apr	208389	207720	195115	157125	163553	128984	187327	186568	186311
23-Apr	213024	211886	198324	159460	159878	132763	189973	190381	189714
24-Apr	219764	216411	201542	159952	161630	136541	192994	192973	193117
25-Apr	223759	224093	204755	161644	162268	140320	195351	196186	196520

To support the previous findings, the RMSE values were compared for both models in table (5) which showed that the RMSE values of the Hybrid (GA-ARIMA) model were much lower than that of the reference model in all countries. This result confirms the findings of the high prediction ability of the presented hybrid model and leads to the conclusion that the (GA-ARIMA) model renders better and precise prediction as compared to the traditional ARIMA models and it can be a reliable method in predicting future forecasts of expected mortality rates as well as confirmed cases of COVID-19.

Table (5): RMSE Comparison between the proposed model and the reference model for prediction of total confirmed cases of COVID-19.

Country	ARIMA- GA Model	Reference Model
Spain	2424	13623
France	2668	32648
Italy	644	2107

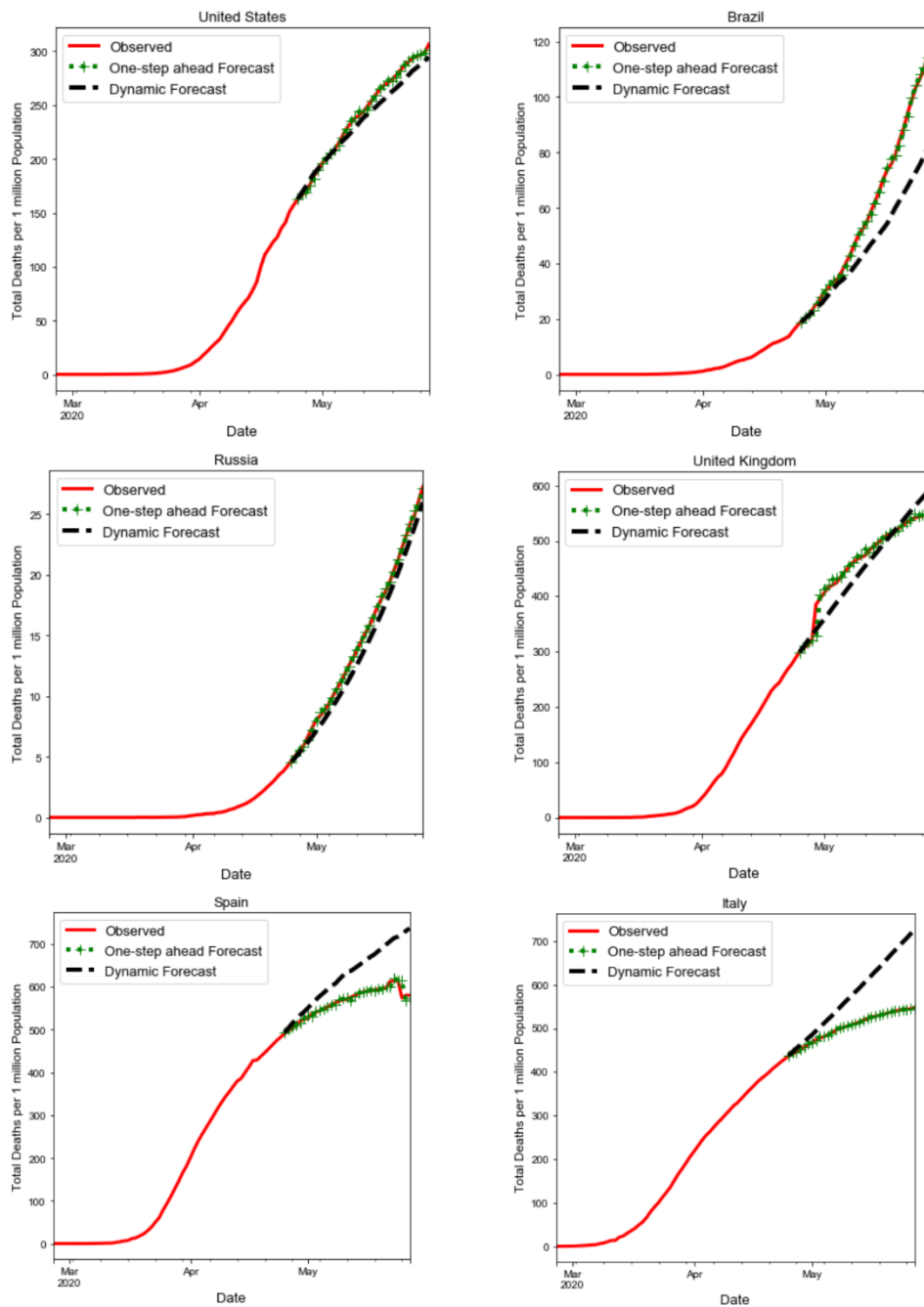
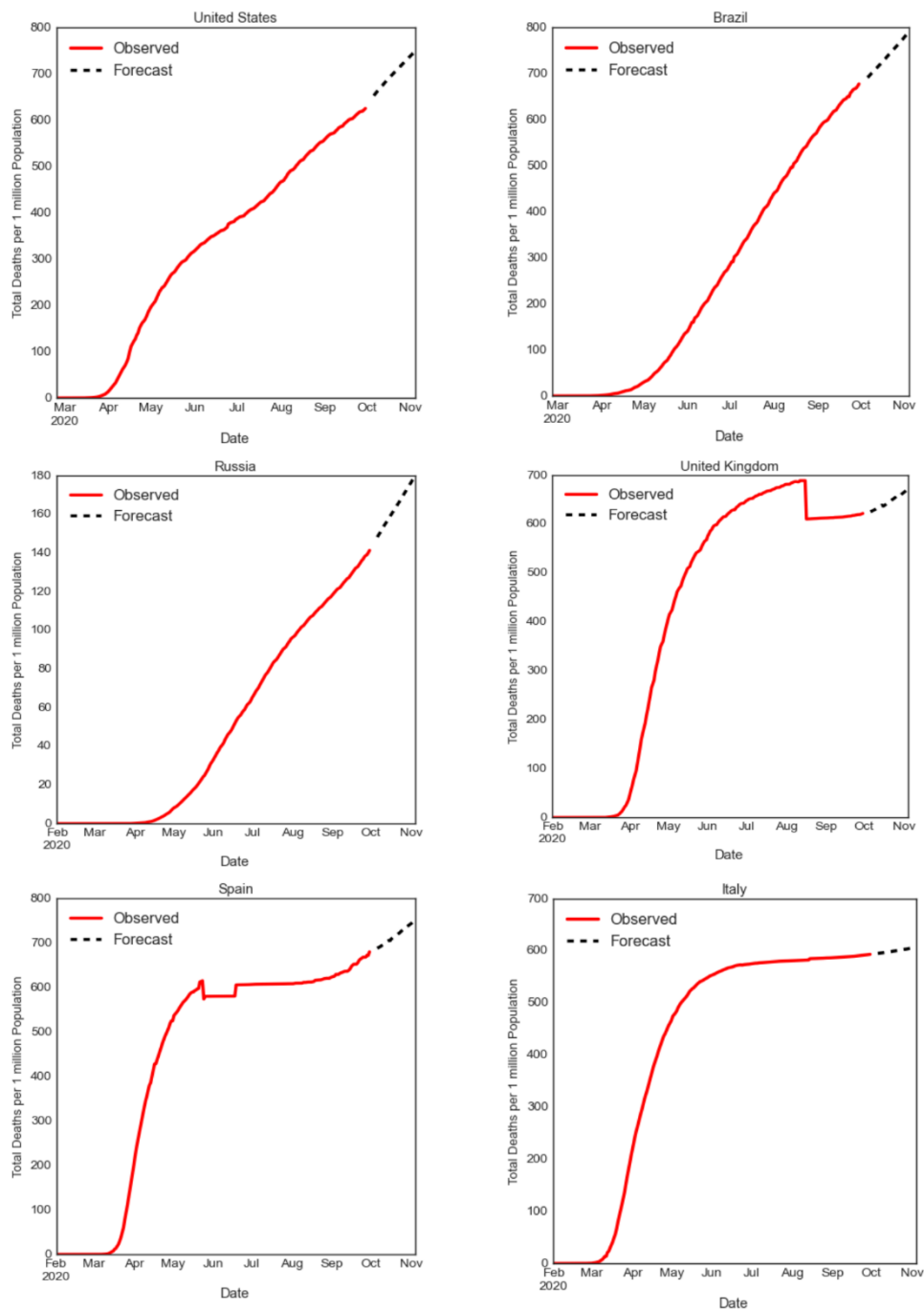


Fig. 7. Static and Dynamic forecasts compared to the observed values for Total Deaths per 1 million Population with 95% confidence limits.



. Fig. 8. Future Forecasts for Total Deaths per 1 million Population with 95% confidence limits

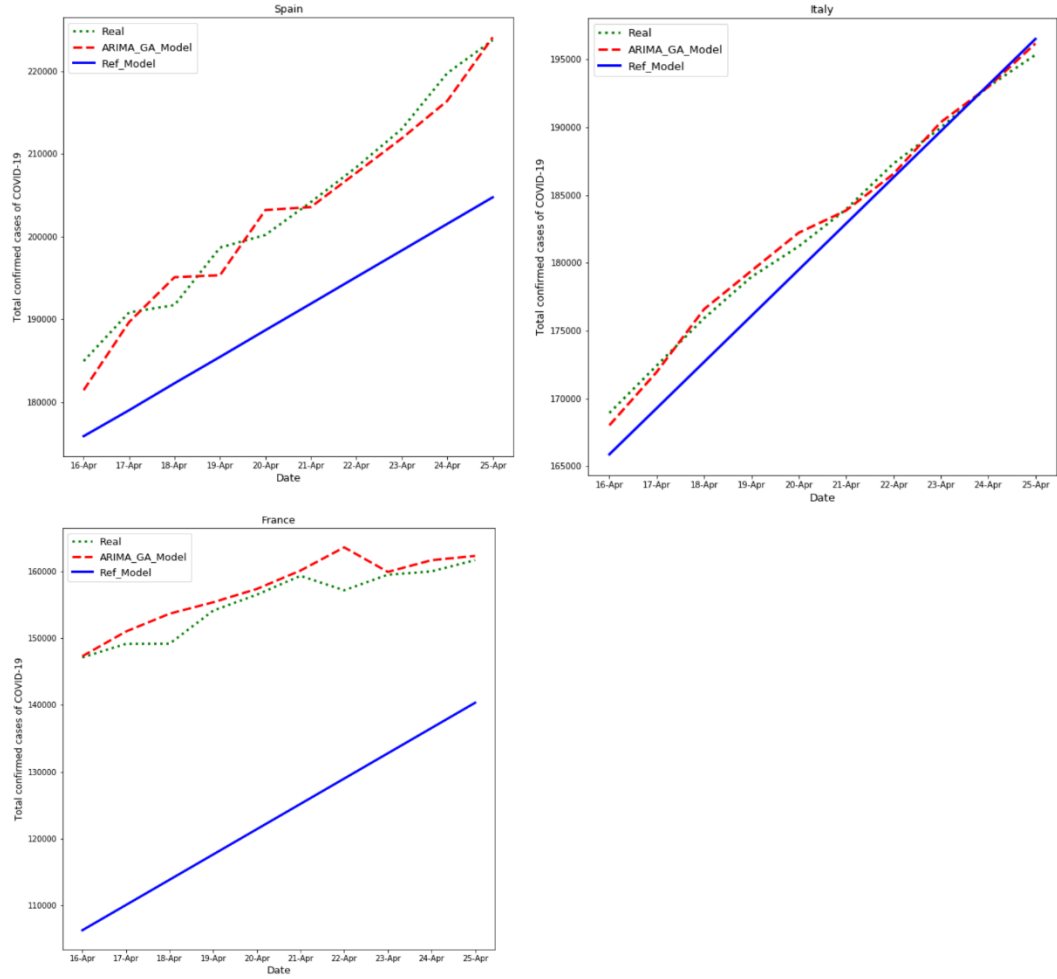


Fig. 9. Total confirmed cases prediction for COVID-19 comparison between the proposed model and reference model with 95% confidence limits

5. Conclusion

In this paper, a hybrid Genetic Algorithm based on Autoregressive Integrated Moving Average (GA-ARIMA) model was proposed to forecast the mortality rates of COVID-19 in six of the top most affected countries. The aim of this study was to develop an advanced reliable predicting model that provides future forecasts of possible confirmed cases and mortality rates that could assist the public health authorities in future plans to efficiently face the pandemic threat. After evaluating the best fit model, the model was further employed to forecast the future mortality rates for each country one month later from 1 October 2020 to 1 November 2020. The model predicted a rise in the mortality rates of

COVID-19 for all countries except Spain where the death rates were expected to decline as of June. The number of confirmed cases of COVID-19 in Italy, Spain, and France predicted by the proposed hybrid model and that of a previous study based on the classical ARIMA model was compared to real values. The comparative study showed that the predicted values of the hybrid proposed model were closer to the real data than the reference model in all countries which reveals the high precision of the (GA-ARIMA) in forecasting. The RMSE values of the proposed model were observed to be much lower than that of the reference model which supported these findings of the high forecasting ability of the presented hybrid model. This leads to the conclusion that the (GA-ARIMA) model renders better and precise prediction as compared to the traditional ARIMA models and it can be a reliable method in predicting future forecasts of expected mortality rates as well as confirmed cases of COVID-19.

References

1. European Centre for Disease Prevention and Control. COVID-19 situation update worldwide. www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
2. Jung S, Akhmetzhanov AR, Hayashi K, et al. Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: inference using exported cases. *J Clin Med.* 2020;9(2):523.
3. Nishiura H, Kobayashi T, Yang Y, et al. The rate of underascertainment of novel coronavirus (2019-nCoV) infection: estimation using Japanese passengers data on evacuation flights. Published online 2020.
4. Ala'raj M, Majdalawieh M, Nizamuddin N. Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections. *Infect Dis Model.* 2021;6:98-111.
5. Wang L, Li J, Guo S, et al. Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. *Sci Total Environ.* Published online 2020:138394.
6. World Health Organization. Coronavirus disease (COVID-19) outbreak situation. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>
7. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med.* Published online 2020.
8. Malki Z, Atlam E-S, Ewis A, et al. ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Comput Appl.* 2021;33(7):2929-2948.
9. Chen T-M, Rui J, Wang Q-P, Zhao Z-Y, Cui J-A, Yin L. A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infect Dis poverty.* 2020;9(1):1-8.
10. ArunKumar KE, Kalaga D V, Kumar CMS, Chilkoor G, Kawaji M, Brenza TM. Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA). *Appl Soft Comput.* 2021;103:107161.
11. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS One.* 2020;15(3):e0230405.
12. Sato RC. Disease management with ARIMA model in time series. *Einstein.* 2013;11(1):128.
13. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and Random Forest

- time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics*. 2014;15(1):276.
14. Liu Q, Li Z, Ji Y, et al. Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu Province of China using advanced statistical time-series analyses. *Infect Drug Resist*. 2019;12:2311.
 15. Ren H, Li J, Yuan Z-A, Hu J-Y, Yu Y, Lu Y-H. The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China. *BMC Infect Dis*. 2013;13(1):421.
 16. Kucharski AJ, Russell TW, Diamond C, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *lancet Infect Dis*. Published online 2020.
 17. Gábor A, Banga JR. Robust and efficient parameter estimation in dynamic models of biological systems. *BMC Syst Biol*. 2015;9(1):1-25.
 18. Ferrández MR, Ivorra B, Redondo JL, Ramos AM, Ortigosa PM. A multi-objective approach to estimate parameters of compartmental epidemiological models. *Appl to Ebola Virus Dis epidemics Res net*. Published online 2020:1-49.
 19. Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. *Sci Total Environ*. Published online 2020:138817.
 20. Prado F, Minutolo MC, Kristjanpoller W. Forecasting based on an ensemble Autoregressive Moving Average-Adaptive neuro-Fuzzy inference system--Neural network-Genetic Algorithm Framework. *Energy*. 2020;197:117159.
 21. Wiczeorek Michał and Siłka J, Woźniak M. Neural network powered COVID-19 spread forecasting model. *Chaos, Solitons & Fractals*. 2020;140:110203.
 22. Hawas M. Generated time-series prediction data of COVID-19' s daily infections in Brazil by using recurrent neural networks. *Data Br*. 2020;32:106175.
 23. Saba AI, Elsheikh AH. Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. *Process Saf Environ Prot*. Published online 2020.
 24. Box GEP, Jenkins GM, Reinsel GC. *Time Series Analysis: Forecasting and Control*. Vol 734. John Wiley & Sons; 2011.
 25. Sun H, Koch M. Case study: analysis and forecasting of salinity in Apalachicola Bay, Florida, using Box-Jenkins ARIMA models. *J Hydraul Eng*. 2001;127(9):718-727.
 26. Adebisi AA, Adewumi AO, Ayo CK. Comparison of ARIMA and artificial neural networks models for stock price prediction. *J Appl Math*. 2014;2014.
 27. Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea. *Symmetry (Basel)*. 2019;11(2):240.
 28. Cao L, Liu H, Li J, Yin X, Duan Y, Wang J. Relationship of meteorological factors and human brucellosis in Hebei province, China. *Sci Total Environ*. 2020;703:135491.
 29. Gonzalez J, Yu W. Non-linear system modeling using LSTM neural networks. *IFAC-PapersOnLine*. 2018;51(13):485-489.
 30. Zhao Y, Yang R, Chevalier G, Xu X, Zhang Z. Deep residual bidir-LSTM for human activity recognition using wearable sensors. *Math Probl Eng*. 2018;2018.
 31. Basnayake W, Attygalle D, Hansen LL, Nandala13 KDW. Grid search based parameter tuning of dynamic neural network to forecast daily reservoir inflow. In: *Proceedings of the International Conference on Computational Modeling and Simulation (ICCMS 2017)*. ; 2017:291-295.
 32. Worldometer. 2020. Covid-19 Coronavirus Pandemic. Published 2020. <https://www.worldometers.info/coronavirus>
 33. Fanoodi B, Malmir B, Jahantigh FF. Reducing demand uncertainty in the platelet supply chain through artificial neural networks and ARIMA models. *Comput Biol Med*. 2019;113:103415.
 34. Roy S, Bhunia GS, Shit PK. Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. *Model earth Syst Environ*. 2021;7(2):1385-1391.
 35. Liu Q, Liu X, Jiang B, Yang W. Forecasting incidence of hemorrhagic fever with renal

- syndrome in China using ARIMA model. *BMC Infect Dis.* 2011;11(1):218.
36. He Z, Tao H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. *Int J Infect Dis.* 2018;74:61-70.
 37. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data Br.* Published online 2020:105340.
 38. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. OTexts; 2018.
 39. Ruiz AB, Saborido R, Luque M. A preference-based evolutionary algorithm for multiobjective optimization: the weighting achievement scalarizing function genetic algorithm. *J Glob Optim.* 2015;62(1):101-129.
 40. Cordon O, Herrera F, Gomide F, Hoffmann F, Magdalena L. Ten years of genetic fuzzy systems: current framework and new trends. In: *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*. Vol 3. ; 2001:1241-1246.
 41. Shi C, Cai Y, Fu D, Dong Y, Wu B. A link clustering based overlapping community detection algorithm. *Data Knowl Eng.* 2013;87:394-404.
 42. Hyndman RJ, Khandakar Y, others. *Automatic Time Series for Forecasting: The Forecast Package for R*. Monash University, Department of Econometrics and Business Statistics~...; 2007.
 43. Erevli S, Uzgören N, Bingöl D, Erevli B. Drinking water quality control: control charts for turbidity and pH. *J Water, Sanit Hyg Dev.* 2016;6(4):511-518.
 44. Li T, Kou G, Peng Y. Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods. *Inf Syst.* 2020;91:101494.
 45. Leybourne SJ, Mills TC, Newbold P. Spurious rejections by Dickey--Fuller tests in the presence of a break under the null. *J Econom.* 1998;87(1):191-203.
 46. Sreedhar M, Reddy SAN, Chakra SA, Kumar TS, Reddy SS, Kumar BV. A Review on Advanced Optimization Algorithms in Multidisciplinary Applications. In: *Recent Trends in Mechanical Engineering*. Springer; 2020:745-755.
 47. Kou G, Xiao H, Cao M, Lee LH. Optimal computing budget allocation for the vector evaluated genetic algorithm in multi-objective simulation optimization. *Automatica.* 2021;129:109599.
 48. Li T, Kou G, Peng Y, Shi Y. Classifying with adaptive hyper-spheres: An incremental classifier based on competitive learning. *IEEE Trans Syst man, Cybern Syst.* 2017;50(4):1218-1229.