Scientific Inquiry and Review (SIR)
Volume 8 Issue 3, 2024
ISSN$_{(P)}$: 2521-2427, ISSN$_{(E)}$: 2521-2435
Homepage: https://journals.umt.edu.pk/index.php/SIR

Article QR

Title:	**Predictive ARIMA Model with a Machine Learning (ML) Approach for COVID-19 Data in Pakistan**
Author (s):	Muhammad Ilyas[1], Shaheen Abbas[2], and Faisal Nawaz[3]
Affiliation (s):	[1]Department of Mathematics, Government College University Hyderabad, Pakistan [2]Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan [3]Dawood University of Engineering and Technology, Karachi, Pakistan
DOI:	https://doi.org/10.32350/sir.83.02
History:	Received: April 01, 2024, Revised: 29 July, 2024, Accepted: July 29, 2024, Published: September 26, 2024
Citation:	Ilyas M, Abbas S, Nawaz F. Predictive ARIMA model with a machine learning (ML) approach for COVID-19 data in Pakistan. *Sci Inq Rev.* 2024;8(3):25–57. https://doi.org/10.32350/sir.83.02
Copyright:	
Licensing:	
Conflict of Interest:	Author(s) declared no conflict of interest

A publication of
The School of Science
University of Management and Technology, Lahore, Pakistan

Predictive ARIMA Model with a Machine Learning (ML) Approach for COVID-19 Data in Pakistan

Muhammad Ilyas[1*], Shaheen Abbas[2], and Faisal Nawaz[3]

[1]Department of Mathematics, Government College University Hyderabad, Pakistan
[2]Mathematical Sciences Research Centre, Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan
[3]Department of mathematics, Dawood University of Engineering and Technology, Karachi, Pakistan

ABSTRACT

This study is based on the application of an ARIMA (p, d, q) based machine learning (ML) approach to evaluate the dynamics of COVID-19 pandemic. The focus is on estimating epidemic trends and performing diagnostic scrutiny with model fitting. The data including all four waves of the pandemic pertaining to Pakistan, covering all four provinces (Sindh, Punjab, Khyber Pakhtunkhwa, Balochistan, as well as Gilgit Baltistan, Azad Jammu Kashmir, and the capital city Islamabad**,** collected from February 26, 2020, to September 30, 2021, is analyzed. The ML algorithm is used to optimize the results of ADF, unit root test which ensures the minimum of ACF, and PACF graphs intention of the data series. The results employ the fitted ARIMA models (1, 1, 1) and (1, 1, 7) for the 1st to 4th waves, confirming daily infected cases across the entire dataset of Pakistan. The cumulative trained observations are from the 1st wave (February 26, 2020, to October 20, 2020), 2nd wave (October 21, 2020, to March 16, 2021), 3rd wave (March 17, 2021, to July 10, 2021), and 4th wave (July 11, 2021, to September 30, 2021), with a further 14-day forecast (from October 1 to October 14, 2021). The results show a strong correlation between the trained and predicted values, ranging from 0.8789 to 0.99236. To select predictive model parameters, the model that results in the minimum Bayesian Information Criterion (BIC) value and residuals from the datasets obtained after detaching the unnecessary errors and the 95^% CI for the forecasting error ($\bar{y}_0 \pm t_{crict}.s.e$) are calculated. These values would help to decide the best fitted predictive model.

*Corresponding Author: dr.m.ilyas@gcuh.edu.pk

Keywords: ARIMA (*p, d, q*), machine learning (ML), predictive model parameters, trained and validation data

1. INTRODUCTION

Coronavirus Disease 2019 (COVID-19), caused by the simple acute respiratory syndrome virus 2 (SARS-CoV-2), was initially detected in China in December 2019. In a very brief period of time, it spread worldwide. COVID-19 has affected more than 209 countries with over 120 million confirmed cases by March 2021. This is not the first outbreak of a coronavirus. Previously, a different type of the virus caused the acute respiratory syndrome known as Middle East Respiratory Syndrome (MERS), which emerged in 2003 and 2012 in Middle Eastern countries. The disease is caused by the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) [1–4]. The initial symptoms of the COVID-19 patients are minimal fever and dry cough which appear between 2 and 14 days after getting infected. Other symptoms may include severe nosebleed and shortness of breath, and these may even cause death [1–3]. The symptoms of COVID-19 are shown in Figure 1.

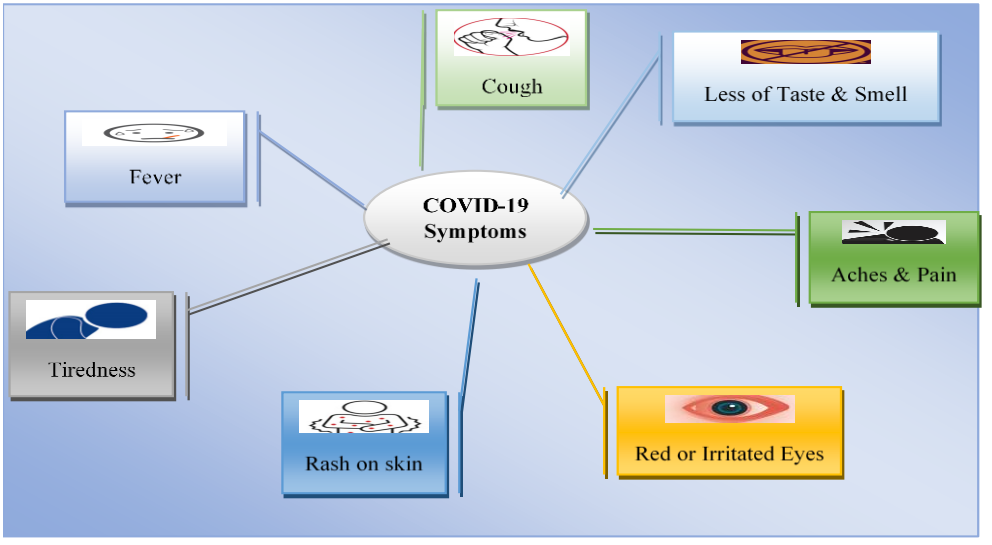


Figure 1. Symptoms of COVID-19

In late 2019, researchers in China identified a novel coronavirus responsible for a respiratory illness, therefore, initially termed Novel Coronavirus Pneumonia (NCP) due to its primary effects on the respiratory

system [1, 2, 5–7]. This virus was later officially named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) by the International Committee on Taxonomy of Viruses [6, 8]. The COVID-19 pandemic has been recognized as one of the most significant global health crises in history. To understand and manage the spread of the virus, various epidemiological models have been employed. These models, which include both classical and advanced approaches, are used to analyze transmission dynamics and predict future trends. Researchers have explored numerous hypothetical parameters to refine these models and improve their accuracy in forecasting disease progression (see Figure 2).

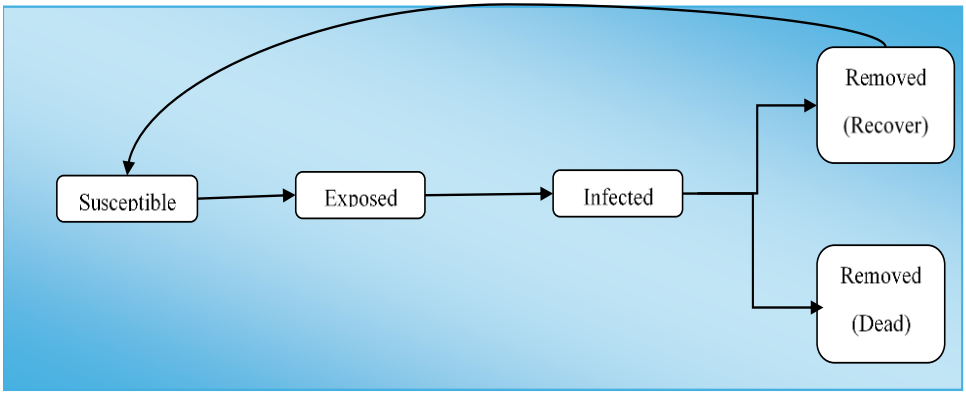


Figure 2. The Flowchart Shows Patient History

The first COVID-19 case in Pakistan was reported on February 26, 2020, in Karachi. By March 15, 2020, the virus had spread to all provinces of Pakistan (Punjab, Sindh, Khyber Pakhtunkhwa, Balochistan), as well as Gilgit-Baltistan, Azad Jammu and Kashmir, and the capital city, Islamabad. As of September 30, 2021, Pakistan has recorded over 1,246,538 cases, 1,170,590 recoveries, and 48,163 deaths. This study aims to forecast COVID-19 trends from February 26, 2020, to September 30, 2021, using stochastic time series analysis. The analysis relies on a single variable to predict COVID-19 infection rates and aims to develop a robust model for forecasting the pandemic's progression.

Early literature on COVID-19 dynamics has employed various methodologies [3, 8]. Many of these approaches use compartmental models, which divide the population into different compartments to describe the spread of infectious diseases. In this study, we propose Box-Jenkins Auto Regressive Integrated Moving Average (ARIMA) models, integrated with

machine learning techniques. The ARIMA (p, d, q) model is selected for its effectiveness in time series forecasting, combining autoregression, differencing, and moving averages to analyze and predict daily COVID-19 data.

Machine learning (ML) algorithms are used to solve many complex problems in various fields, such as business, climatology, healthcare, and robotics [4, 9]. Unlike conventional algorithms, ML algorithms are based on the trial-and-error method. It is generally used for forecasting and prediction [7, 9–15]. The regression and neural network models can predict a patient's future condition for specific diseases [10]. In this study, ML is used to predict COVID-19 data.

2.DATA SCATTERING AND METHODS

To investigate and measure accurate prediction of the COVID-19 pandemic across Pakistan, including all four provinces (Sindh, Punjab, Khyber Pakhtunkhwa, Balochistan), as well as Gilgit Baltistan, Azad Jammu and Kashmir and the capital city Islamabad. Data was collected from the WHO website on a daily basis. Data was collected for the period 26 February 2020 to 30 September 2021. Data was collected separately for the four (4) waves of the pandemic, starting from 26 February 2020 to 21 October 2020, 22 October 2020 to 16 March 2021, 17 March 2021 to 10 July 2021, and 11 July 2021 to 30 September 2021. We computed the stationary of the data set, the residual of the data panel, and detached the pointless errors from prediction models.

The ARIMA model, combined with machine learning techniques, is considered a superior predictive model compared to uncorrelated models and those based on neural networks, which are part of artificial intelligence (AI). These stochastic univariate methods, such as regression or support vector machines, to analyze and predict trends in time series data.

The ARIMA model, when combined with machine learning techniques, integrates the fundamental concepts of autoregression, differencing, and moving averages to accurately forecast future values and analyze time series data [1, 2, 6]. This approach effectively handles non-stationary data through differencing, improving accuracy in capturing trends and seasonality within the time series. Therefore, integrating ARIMA with machine learning algorithms enhances the model's predictive capabilities.

In this study, we utilize stochastic series within ARIMA models to analyze the observed datasets. We are applying various statistical tests, such as the Dickey-Fuller test to assess time series trends and the statistical properties of the data. The integration of Box-Jenkins's ARIMA model with machine learning algorithms for predictive analysis is illustrated in Figure 3. Keeping in view the primary focus on COVID-19 cases, ARIMA models are employed to examine the spread and impact of the virus. We conduct 14-day forecasts for each wave of data, specifically (1) October 21, 2020, to November 4, 2020; (2) March 17, 2021, to March 30, 2021; (3) July 11, 2021, to July 24, 2021; and (4) September 31, 2021, to October 13, 2021. This study demonstrates how integrating ARIMA models with machine learning algorithms enhances predictive capabilities for COVID-19, exploring additional variables and data sources from previous time intervals.

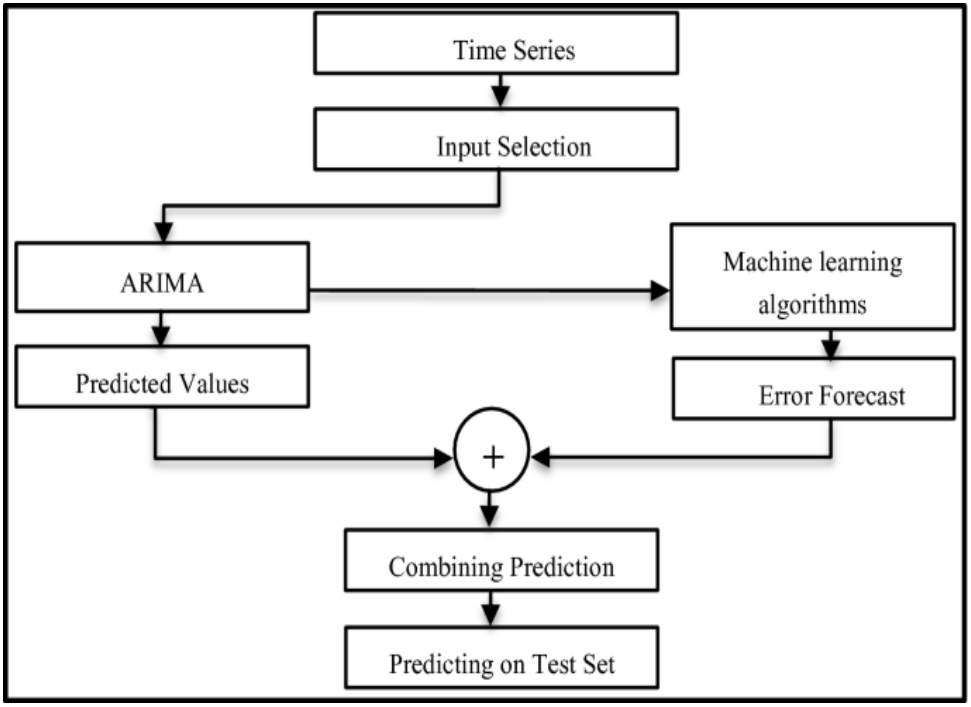


Figure 3. General Structure of ARIMA with ML Algorithm [3]

2.1. Predictive ARIMA Models

In this study, ARIMA models integrated with ML algorithms are used for future prediction and model validation for 14-days forecasting simulations. Many time series datasets exhibit irregularities, seasonality, cyclic patterns, and trends. These characteristics do not always conform to a specific pattern in every time series model. These components help to find suitable predictive approaches for short-term investigations [3, 8]. ARIMA models are used to provide explanations through a broad statistical methodology. The procedure is as follows.

Step 1: Examine the data for high frequency components, non-constant mean μ and variance σ2. Assess seasonality (e.g., daily, weekly) and test for stationarity.

Step 2: Model functions are formulated, and parameters are obtained based on the assumption that the data series remains stationary in Step 1.

Step 3: Diagnosis checking is applied to obtain the validation of the model assumptions of Step 1 and hypothesis confirms that residuals should be satisfied with zero mean μ, constant variance σ2 and normal distribution $(X \sim N\ (\mu, \sigma 2)$ for uncorrelated process $\rho\ (X, Y) = 0$. When both variables are independent, then their correlation would be 0.

Step 4: In the final step, the model is ready to make prediction. go to step 2 and compute the predicted data values.

These steps mainly analyze the prediction of time series data sets. The Augmented Dickey-Fuller test confirms the stationary components, such as constant mean, constant variance, and auto covariance which are not time dependent. The theory of ARIMA models is known as the Box-Jenkins method. An ARIMA model is a procedure of univariate regression analysis that predicts future values based on differences among values, rather than actual values. The ARIMA model integrates three main components to analyze time series data. The first component is Auto Regression (AR), which models how the current value of a variable is influenced by its previous values. It captures the relationship between a variable and its lagged values. The second component is Integration (I), which involves differencing the data to achieve stationarity. This process adjusts the data by subtracting previous values from current values to account for trends and make the time series stationary. The third component is Moving Average (MA), which models the relationship between a variable and the residual errors from a moving average of past observations. It smooths out

fluctuations in the data by averaging data points over a specified period. Together, these components enable the ARIMA model to effectively analyze and forecast time series data.

The ARIMA (*p*, *d*, *q*) is obtained as follows. Here, the theoretical *y* denotes the d^{th} difference of *Y*, in expressions of y.

When d = 0, then

$$y_t = Y_t \tag{1}$$

When d = 1, then

$$y_t = Y_{t-}Y_{t-1} \tag{2}$$

When d = 2, then

$$y_t = (Y_{t-}Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \tag{3}$$

In ARIMA models, y_t depends on its own lagged value. In MA (moving average), the value of y_t depends on lagged predictive errors. Thus, the predictive value depends on its previous (lagged) value. The number of iterations with p, q can be represented by

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-P} + \omega_t \tag{4}$$

$$y_t = \omega_t + \theta_1 y\omega_{t-1} + \theta_2 y\omega_{t-2} + \cdots + \theta_q y\omega_{t-q} \tag{5}$$

To detect the suitable ARIMA model for y_t accordingly, obtaining the differencing (*d*) order must be stationary of the series. If the stationary series is silent then it has auto correlated errors. It suggests that the prediction equality of various values of AR and MA expressions (p *and* $q \geq 1$) are required in the predicting procedure by AR and MA models. The syntax of the ARIMA model is as follows:

ARIMA (< Data interval _ Name > {Target value_ Label}, order = (p, *d*, q). (6)

In time-series analysis, there are various measurement tools for model selection. The best fitted predictive evaluation criteria are discussed below.

2.2. Model Predicting Diagnostic Tools

For the predictive analysis of observed intervals using ARIMA, descriptive statistics of the dataset were examined. To assess the performance of the ARIMA models, various metrics including test

statistics, p-values, lags, and the number of observations were considered in graphical representations. The results indicated that the data were not stationary. Subsequently, the model was used to estimate the trends, and the Dickey-Fuller test was employed to test for stationarity. Additionally, the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) were analyzed to further evaluate the dataset's characteristics, supported by graphical exhibitions to determine the value of Q and P. Furthermore, the AR and MA models were implemented which predicted the future, successfully [1, 3, 12]. Finally, we converted the dataset into the cumulative sum and plotted the diagrams.

The Mean Absolute Scaled Error (MASE) is the extent of the exactitude of the forecasted values and obtained through dividing the means absolute error of the predicted value by the mean absolute error of the original data [15], mathematically written as follows:

$$MASE = \frac{\frac{1}{J}\sum_j (e_j)}{\frac{1}{T}\sum_{t-2}^{T}(Y_t - Y_{t-1})} \quad (7)$$

The Symmetric Mean Absolute Percentage Error (SMAPE) is a measurement of accuracy based on percentage [1, 2]. The difference between A_t and F_t, the actual value and forecasted value is divided by the half of sum of these two values and calculated by

$$SMAPE = \frac{100\%}{n}\sum_{t-1}^{n}\frac{(F_t - A_t)}{(F_t + A_t)/_2} \quad (8)$$

The Mean Absolute Error (MAE) measures the errors in the observations. It is stated as follows:

$$MAE = \frac{\sum_{i=1}^{n}(e_i)}{n}, \quad (9)$$

where e_i is the absolute error which is the difference of predicted value Yi and true value Xi. The Root Mean Square Error (RMSE) is the sum of error and variance based on the difference between predicted values Y_i and true value $X_i i$. It is calculated by

$$RMSE = \sqrt{[\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)]^2} \quad (10)$$

The BIC is normally considered to select models in the linear exponential regression [1, 3, 6].

$$BIC = ln\hat{\sigma}^2 + \frac{(p+q)\ln(\ln(n))}{n} \quad (11)$$

The BIC function is directly associated to Akaike Information Criterion (AIC) [1, 4, 8].

The value of predictive ARIMA is computed and provided in the Results and Discussion section. The optimal values of the fitted 14-day predictive model is evaluated by 95% CI for the predictive value.

2.3. The 14-day Prediction Testing

The 95% CI for the predicted values of the novel COVID-19 virus in Pakistan is calculated by using the following equations:

$\bar{y} \pm t_{crict}.s.e$

and

$$s.e = Sy_x\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$

The results show the 14-day forecasted values of the novel COVID-19 pandemic across four (4) waves in Pakistan. The forecasted values can be calculated through the equation

$\bar{y}_0 \pm t_{crict}.s.e,$

where the standard error of the predicted values is calculated through

$$s.e = Sy_x\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_x}}.$$

2.4. Computational Structure for Model Selection by ML Algorithm

In this section, we select the model with appropriate parameters (*p*, *d*, *q*) of the ARIMA models using ACF and PACF graphs. For this purpose, we applied the ML algorithm for non-stationary exhibition of data series by using correlograms ACF and PACF. Time series plots (1(a-d)) were performed. Autocorrelation slightly decrease as the number of delays increases. To manage the data sets in a proper manner for the selection of the appropriate model, the unit root test was used at different levels. The ARIMA (p, d, q) models require an average processing time of only 7 seconds to perform each simulation on the computer.

The datasets derived from the four (4) waves of COVID-19 is considered as trained data (observed values) and validation data (forecasted values) of daily infected cases in Pakistan. By utilizing the trained datasets, suitable models were recognized, and validation datasets were used to find the projecting act of the model. ML tools were observed by UCL and LCL graphical values for all waves data sets (Figure 7 (a-d)). Model forecasting of each data set was investigated on RMSE, MAPE, and MASE. For data sets, MAE and BIC were found to have the least value to estimate the dynamics of the COVID-19 outbreak across the country in all waves. The estimates made by the ARIMA models were compared with the actual infected dataset observed for 14-day prediction of each wave separately, with the standard error 95% CI. The values of the parameter's alpha, beta, and gamma were confirmed for 14 days prediction.

3. RESULT AND DISCUSSION

This research study presents an optimized ARIMA-ML-based stochastic Box-Jenkins predictive model used to forecast COVID-19 daily infected cases in Pakistan, as described in Section 2. The data covers the period from the 1st wave (26 February 2020) to the 4th wave (30 September 2021), including the entire dataset of Pakistan along with other selected regions.

Firstly, the unit root test results for stationarity (ADF at 1%, 5%, and 10%) are presented in Table 1(a-h), representing data from Pakistan including the provinces of Sindh, Punjab, Khyber Pakhtunkhwa, and Balochistan, as well as Gilgit-Baltistan, Azad Jammu and Kashmir, and the capital city Islamabad. The p-values for all datasets are less than 0.05 ($p < 0.05$), indicating statistical significance. Time series data were computed as stationary with an integrated order I(1), as AR and MA aspects in an ARIMA model applied only to stationary datasets. Initially, parameters p and q were identified with exponential decay starting at lag 1, and no further correlations were found within additional lags, as depicted in the ACF and PACF plots for the period February 2020 to September 2021 in Pakistan, shown in Figure 4(a-d).

Table 1. (a) Results of ADF and ERS Test for Unit Root of All Four Waves of COVID-19 in Pakistan

	LEVEL	1st wave		2nd Wave		3rd wave		4th wave	
		t-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability
Test at Initial Level	ADF	-1.4318	0.5663	-1.8674	0.3468	-1.4102	0.575	-1.8551	0.3517
	1%	-3.4581		-3.4781		-3.488		-3.5133	
	5%	-2.8736		-2.8824		-2.8867		-2.8976	
	10%	-2.5732		-2.5779		-2.5802		-2.5861	
Test at 1st Level	ADF	-9.2143	0.00	-4.003	0.00	-12.896	0.00	-8.2671	0.00
	1%	-3.4581		-3.4781		-3.488		-3.5133	
	5%	-2.8736		-2.8824		-2.8867		-2.8976	
	10%	-2.5732		-2.5779		-2.5802		-2.5861	

Table 1. (b) Results of ADF and ERS Test for Unit Root of All Four Waves of COVID-19 in Punjab

	LEVEL	1st wave		2nd Wave		3rd wave		4th wave	
		t-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability
Test at Initial Level	ADF	-1.5013	0.5314	-1.3386	0.6105	-0.7412	0.8311	-1.6594	0.4478
	1%	-3.4579		-3.4761		-3.4885		-3.5144	
	5%	-2.8735		-2.8815		-2.8869		-2.8981	
	10%	-2.5732		-2.5775		-2.5804		-2.5863	
Test at 1st Level	ADF	-19.344	0.00	-22.67	0.00	-14.969	0.00	-13.00	0.00
	1%	-3.4579		-3.4761		-3.4885		-3.5144	
	5%	-2.8735		-2.8815		-2.8869		-2.8981	
	10%	-2.5732		-2.5775		-2.5804		-2.5863	

Table 1. (c) Results of ADF and ERS Test for Unit Root of All Four Waves of COVID-19 in Sindh

	LEVEL	1st wave		2nd Wave		3rd wave		4th wave	
		t-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability
Test at Initial Level	ADF	-1.9436	0.3119	-1.5951	0.4825	-1.9768	0.2967	-1.7311	0.4119
	1%	-3.4578		-3.4761		-3.4885		-3.5133	
	5%	-2.8735		-2.8815		-2.8869		-2.897	
	10%	-2.5732		-2.5775		-2.5804		-2.5861	
Test at 1st Level	ADF	-22.692	0.00	-21.779	0.00	-18.949	0.00	-9.7559	0.00
	1%	-3.4578		-3.4761		-3.4885		-3.5133	
	5%	-2.8735		-2.8815		-2.8869		-2.897	
	10%	-2.5732		-2.5775		-2.5804		-2.5861	

Table 1. (d) Results of ADF and ERS Test for Unit Root of All Four Waves of COVID-19 in Khyber Pakhtoon Khuwa

	LEVEL	1st wave		2nd Wave		3rd wave		4th wave	
		t-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability
Test at Initial Level	ADF	-1.4904	0.5369	-3.6718	0.0055	-0.4001	0.9042	-3.1675	0.0256
	1%	-3.4579		-3.4761		-3.4919		-3.5133	
	5%	-2.8735		-2.8815		-2.8884		-2.8976	
	10%	-2.5732		-2.5775		-2.5811		-2.5861	
Test at 1st Level	ADF	-19.543	0.00	-8.8854	0.00	-4.5756	0.00	-12.404	0.00
	1%	-3.4579		-3.4778		-3.4919		-3.5144	
	5%	-2.8735		-2.8823		-2.8884		-2.8981	
	10%	-2.5732		-2.5779		-2.5811		-2.5863	

Table 1. (e) Results of ADF and ERS Test for Unit Root of All Four Waves of COVID-19 in Balochistan

	LEVEL	1st wave		2nd Wave		3rd wave		4th wave	
		t-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability
Test at Initial Level	ADF	-2.3052	0.1712	-1.5054	0.5281	-5.2439	0.00	-2.8621	0.0544
	1%	-3.4581		-3.4771		-3.488		-3.5144	
	5%	-2.8736		-2.8819		-2.8867		-2.8981	
	10%	-2.5732		-2.5777		-2.5802		-2.2586	
Test at 1st Level	ADF	-15.036	0.00	-10.705	0.00	-10.896	0.00	-14.417	0.00
	1%	-3.4581		-3.4771		-3.4891		-3.5144	
	5%	-2.8736		-2.8819		-2.8871		-2.8981	
	10%	-2.5732		-2.5777		-2.5805		-2.2586	

Table 1. (f) Results of ADF and ERS Test for Unit Root of All Four Waves of COVID-19 in Gilgit Baltistan.

	LEVEL	1st wave		2nd Wave		3rd wave		4th wave	
		t-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability
Test at Initial Level	ADF	-3.3146	0.0254	-2.5432	0.1075	0.6392	0.9902	-1.1196	0.7045
	1%	-3.4581		-3.4764		-3.4896		-3.5155	
	5%	-2.8736		-2.8816		-2.8874		-2.8986	
	10%	-2.2573		-2.5775		-2.5806		-2.5866	
Test at 1st Level	ADF	-15.561	0.00	-12.704	0.00	-10.814	0.00	-12.301	0.00
	1%	-3.4581		-3.4764		-3.4896		-3.5155	
	5%	-2.8736		-2.8816		-2.8874		-2.8986	
	10%	-2.2573		-2.5775		-2.5806		-2.5866	

Table 1. (g) Results of ADF and ERS Test for Unit Root of All Four Waves of COVID-19 in Azad Jammu and Kashmir

	LEVEL	1st wave		2nd Wave		3rd wave		4th wave	
		t-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability
Test at Initial Level	ADF	-0.5934	0.0254	-2.5432	0.1075	-1.115	0.7079	-0.9853	0.755
	1%	-3.4584		-3.4768		-3.4919		-3.5155	
	5%	-2.8738		-2.8818		-2.8884		-2.8986	
	10%	-2.5733		-2.5776		-2.5811		-2.5866	
Test at 1st Level	ADF	-10.124	0.00	-12.362	0.00	-4.6106	0.00	-13.412	0.00
	1%	-3.4584		-3.4768		-3.4919		-3.5155	
	5%	-2.8738		-2.8818		-2.8884		-2.8986	
	10%	-2.5733		-2.5776		-2.5811		-2.5866	

Table 1. (h) Results of ADF and ERS Test for Unit Root of All Four Waves of COVID-19 in Islamabad

	LEVEL	1st wave		2nd Wave		3rd wave		4th wave	
		t-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability	*t*-Statistic	Probability
Test at Initial Level	ADF	-1.465	0.5497	-1.4614	0.5504	-1.3648	0.5971	-2.4209	0.1393
	1%	-3.4583		-3.4761		-3.4891		-3.5144	
	5%	-2.8737		-2.8815		-2.8871		-2.8981	
	10%	-2.5733		-2.5775		-2.5805		-2.5863	
Test at 1st Level	ADF	-10.753	0.00	-18.572	0.00	-12.821	0.00	-13.897	0.00
	1%	-3.4583		-3.4761		-3.4891		-3.5144	
	5%	-2.8737		-2.8815		-2.8871		-2.8981	
	10%	-2.5733		-2.5775		-2.5805		-2.5863	

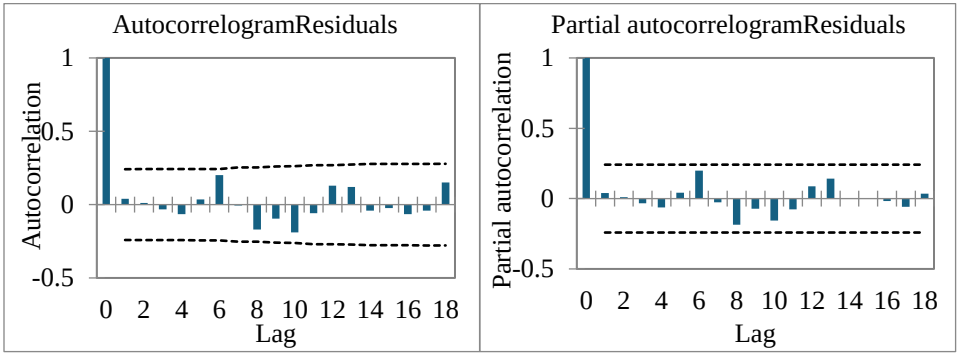


Figure 4 (a). Correlogram Plots During 1st wave (26 Feb 2020 to 21 October 2020).

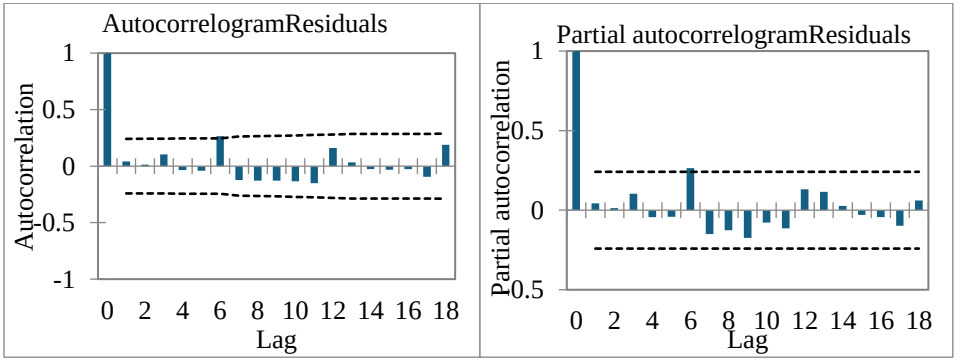


Figure 4 (b). Correlogram for the 2nd wave (22 October 2020 to 16 March 2021)

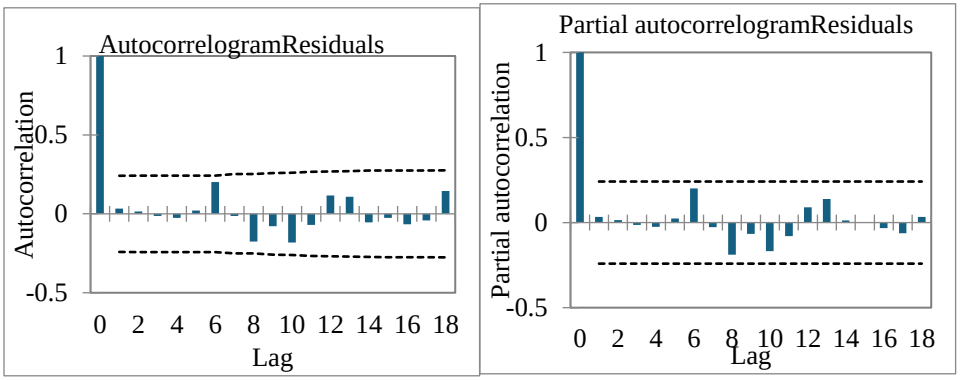


Figure 4. (c). Correlogram for the 3rd wave (17 March 2021 to 10 July 2021)

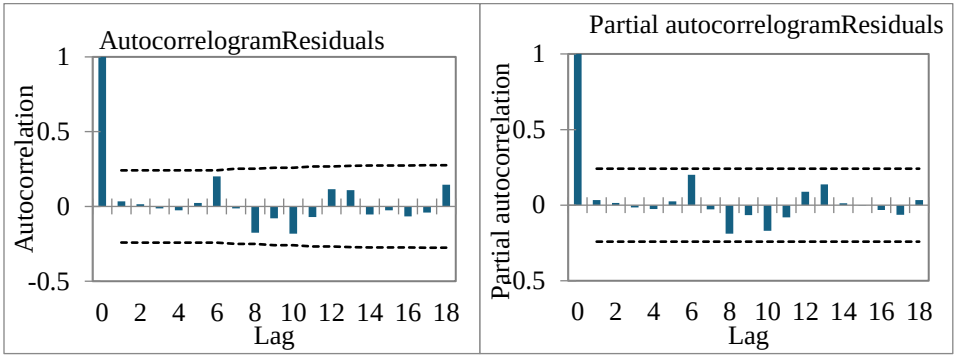


Figure 4. (d). Correlogram for the 4th wave (11 July 2021 to 30 September 2021).

To compute the cumulative trained observations, we utilized the datasets for training the fitted ARIMA (1, 1, 1) model for the first wave (26 February 2020 to 21 October 2020), and the fitted ARIMA (1, 1, 7) model for the second to fourth waves (22 October 2020 to 16 March 2021, 17 March 2021 to 10 July 2021, and 11 July 2021 to 30 September 2021). These models were statistically appropriate, with R^2 values of 0.1266, 0.2549, 0.11, and 0.782, respectively. The minimum values for AIC and BIC were 14.56, 14.61, 14.327, 14.409, 15.02, 15.11, 15.18, and 15.29, confirming the model's accuracy across Pakistan.

The model's validity was further confirmed using diagnostic tools such as RMSE, MAPE, Max APE, MAE, the Ljung-Box Q test, and BIC. Various error parameters (mean, sum of error, minimum, maximum) and forecasted values are shown in Table 2 and Figure 5(a-d). These gradient and residual graphs for the fitted ARIMA (1,1,1) and ARIMA (1,1,7) models demonstrate that the data points deviate slightly from the fitted model, highlighting the model's robustness and accuracy.

Table 2. Different Errors and Parameters of Forecasted Values of Four Different Waves of COVID-19 in Pakistan

Fit Statistic		Stationary *R*-squared	*R*-squared	RMSE	MAPE	Max APE	MAE	Max AE	Normalized BIC
1st wave	Mean	0.307	0.775	11.5944	5.3554	7.3415	6.9381	5.5449	8.305
	SE	0.096	0.183	11.9025	2.4563	6.1286	7.354	5.1272	2.688
	Minimum	0.166	0.413	9.886	25.979	2.5811	6.379	3.1181	4.605
	Maximum	0.429	0.954	3.3919	8.475	20.812	2.0998	14.7162	11.737
2nd wave	Mean	0.371	0.712	9.599	5.166	3.5807	7.0424	3.6952	7.683
	SE	0.174	0.139	10.028	6.281	8.96851	7.4673	3.6953	3.123
	Minimum	0.244	0.514	2.932	1.126	4.587	2.108	9.758	2.185
	Maximum	0.787	0.844	28.392	19.933	25.7385	2.135	10.6719	11.399
3rd wave	Mean	0.367	0.786	14.25	28.326	20.464	9.8199	5.0844	8.856
	SE	0.243	0.188	14.032	17.939	15.0284	9.955	4.3243	2.558
	Minimum	0.13	0.373	9.862	1.1291	6.888	6.003	5.9558	4.618
	Maximum	0.753	0.944	4.1694	6.744	5.3616	30.186	11.356	12.188
4th wave	Mean	0.293	0.72	14.643	28.294	2.1464	11. 755	4.3838	9.191
	SE	0.204	0.115	14.282	15.985	1.7162	10.932	3.6872	2.093
	Minimum	0.011	0.584	18.417	10.74	4.4294	13.364	5.6467	5.935
	Maximum	0.688	0.893	4.4203	5.7083	6.1247	33.4754	11.4344	12.292

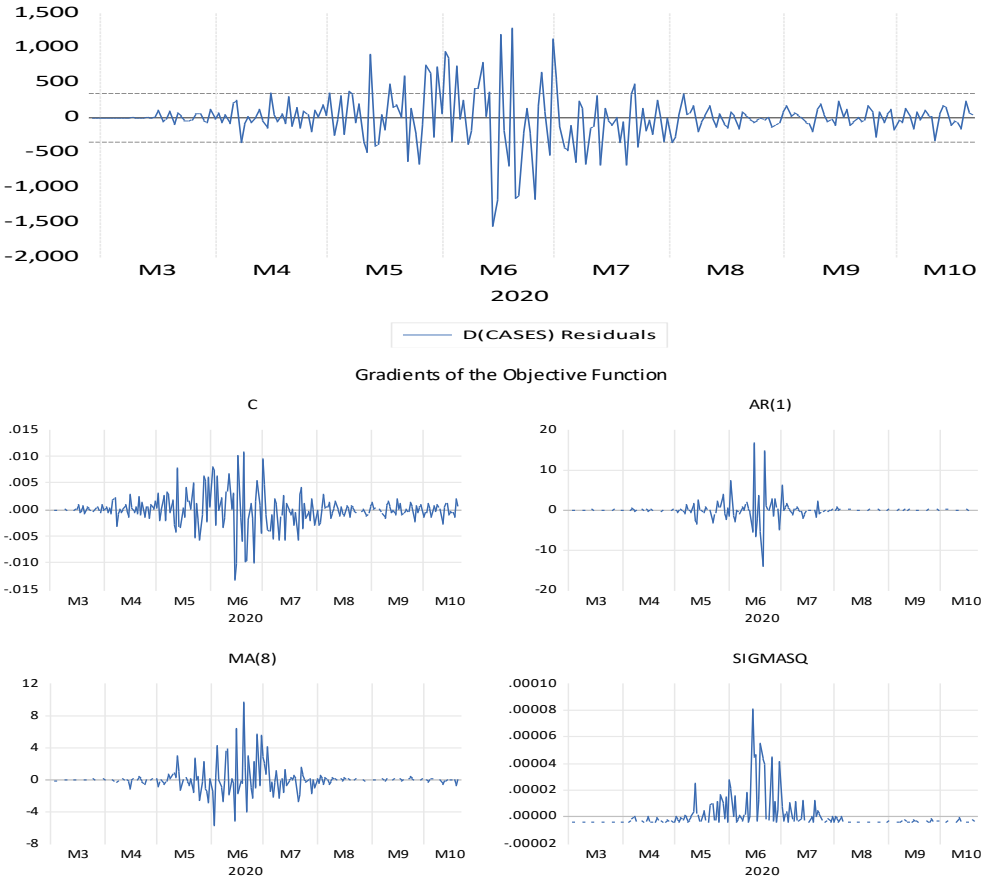
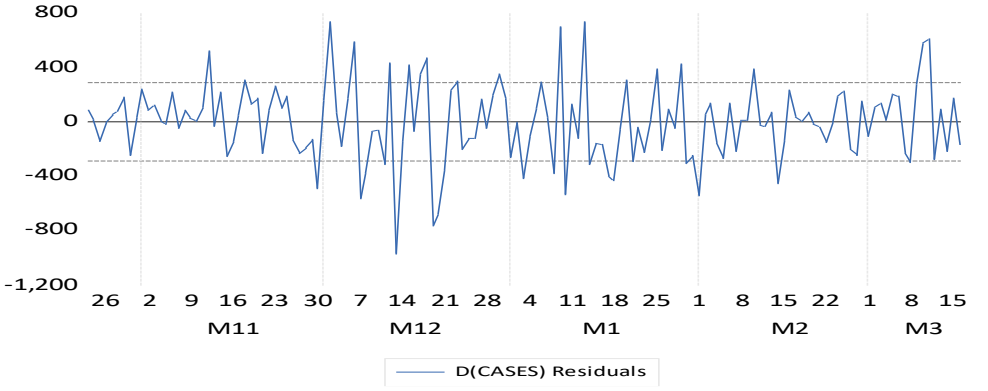


Figure 5 (a). Gradient and Residual Graph for Fitted ARIMA (1, 1, 1) for the First Wave of COVID-19 in Pakistan



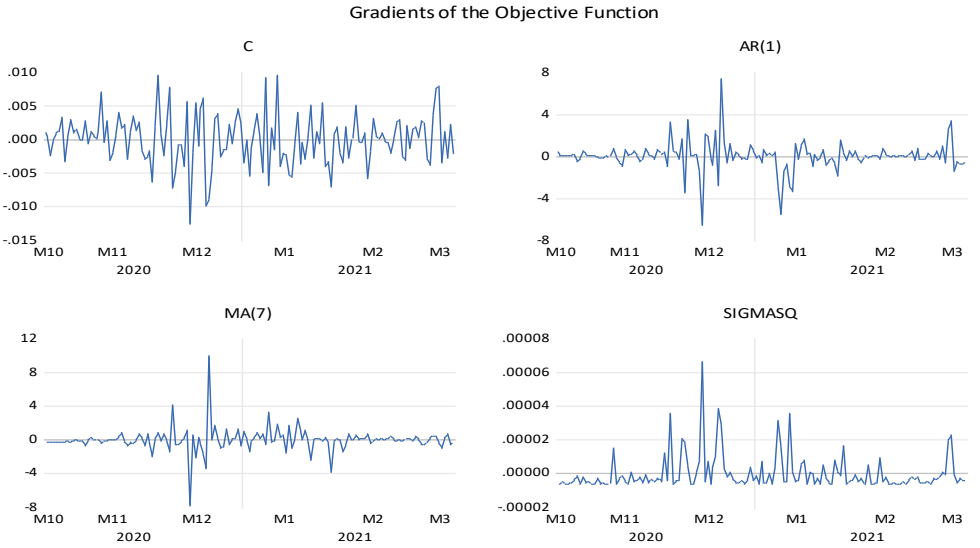
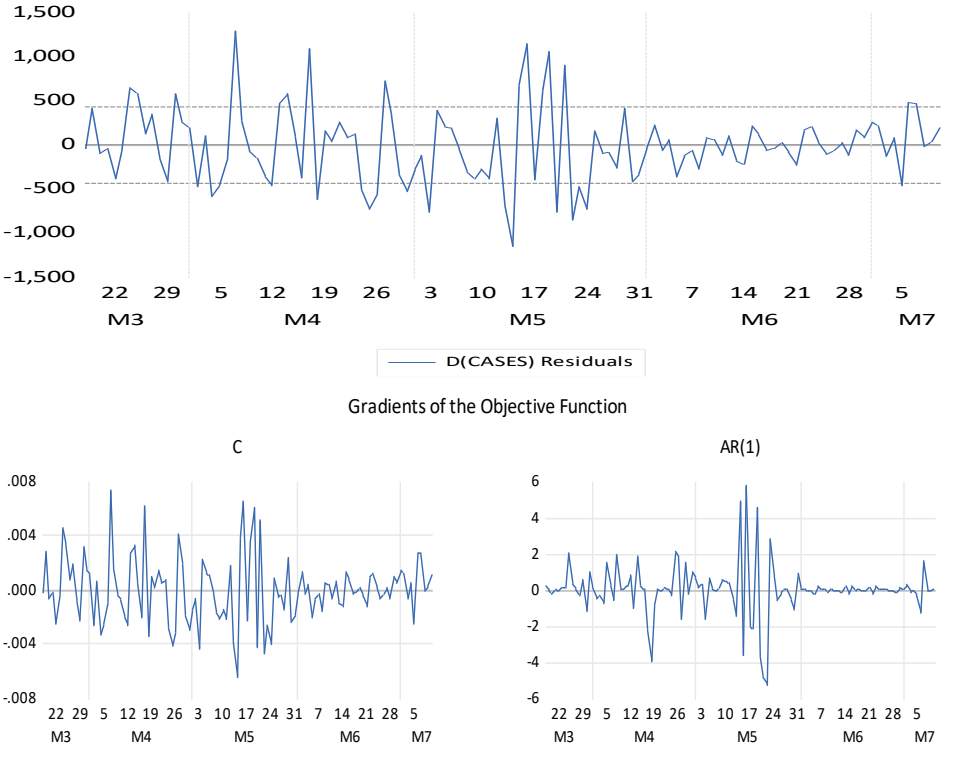


Figure 5 (b). Gradient and Residual Graph for Fitted ARIMA (1,1,7) for the Second Wave of COVID-19 in Pakistan



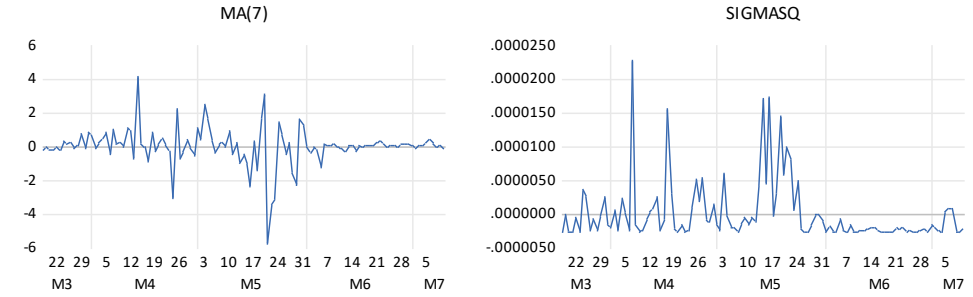


Figure 5 (c). Gradient and Residual Graph for Fitted ARIMA (1,1,7) for the Third Wave of COVID-19 in Pakistan

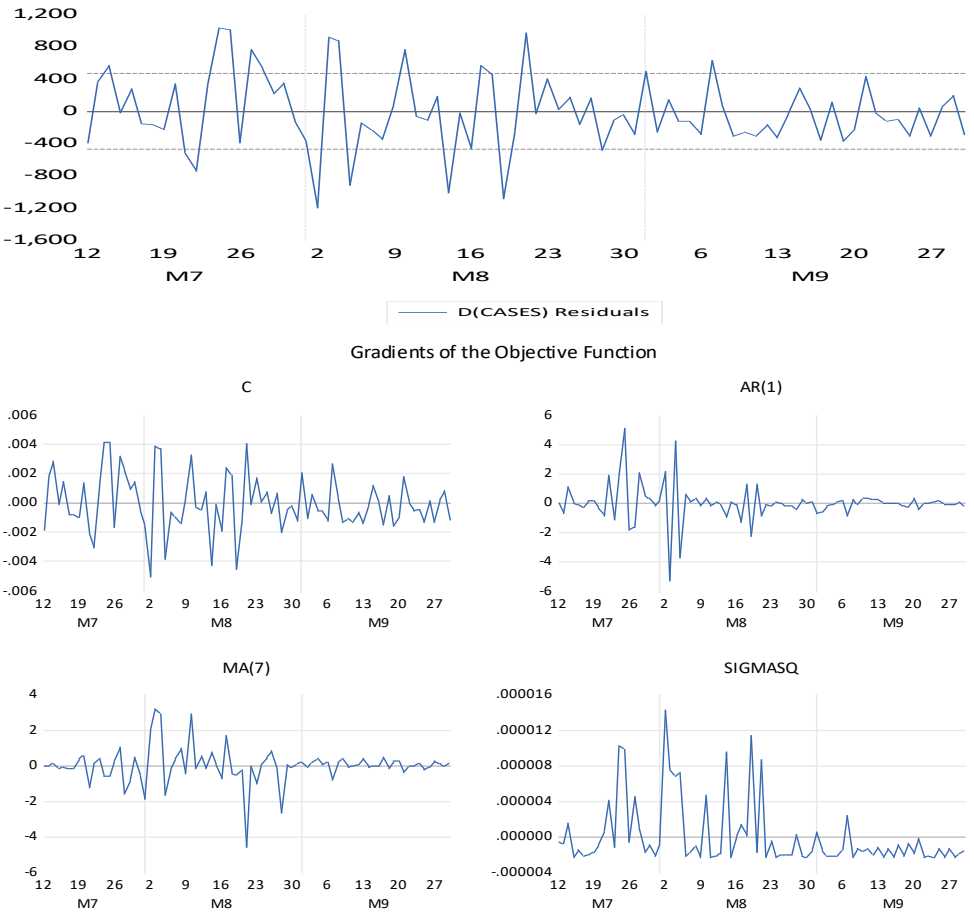


Figure 5 (d). Gradient and Residual Graph for Fitted ARIMA (1,1,7) for the Fourth Wave of COVID-19 in Pakistan

Table 3. Basic Statistics of the Daily Cases of COVID-19 in Pakistan Including its four Provinces, GB, AJK and Islamabad City

	ARIMA Model	Model Fit statistics								Ljung-Box Q (18)		
		Stationary R^2	R^2	RMSE	MAPE	MAE	Max APE	Max AE	Normalized BIC	Statistics	DF	Sig.
First Wave	Pakistan	.198	.954	3.33919	2.5979	2.0998	3.4013	14.7162	11.737	30.814	13	.004
	Punjab	.362	.886	1.9805	4.1390	1.1722	3.7152	9.4913	10.671	25.688	14	.028
	Sindh	.166	.894	2.1798	3.4668	1.2435	4.2965	9.4166	10.785	34.756	17	.007
	KPK	.315	.878	6.4231	3.5759	3.9800	258.911	4.2322	8.394	27.111	15	.028
	Balochistan	.276	.629	4.8244	8.4576	2.8146	20.8312	2.8005	7.775	38.208	17	.002
	GB	.422	.413	9.886	8.4750	7.289	10.2616	3.1181	4.605	13.943	17	.671
	AJK	.429	.694	1.064	7.8051	6.379	9.1813	3.9319	4.733	21.291	13	.067
	Islamabad	.288	.853	4.5805	4.3261	2.279	4.5041	2.9149	7.741	32.610	14	.003
Second Wave	Pakistan	.270	.843	2.83692	11.226	2.1435	45.187	10.67319	11.399	23.411	15	.076
	Punjab	.787	.746	1.2462	15.727	9.203	99.830	5.29304	9.682	31.724	17	.016
	Sindh	.295	.834	1.9796	1.987	13.985	12.3476	6.5228	10.605	47.411	17	.000
	KPK	.338	.514	7.9003	1.9933	5.7828	27.389	3.5001	8.773	29.647	17	.029
	Balochistan	.309	.568	1.1439	5.456	8.377	4.6763	4.072	4.908	20.689	17	.241
	GB	.244	.776	2.932	5.196	2.108	4.1515	9.758	2.185	19.149	17	.320
	AJK	.332	.571	1.9923	4.4.947	1.417	16.1785	7.608	6.018	23.433	17	.136
	Islamabad	.392	.844	4.9144	1.957	3.539	13.2267	2.086	7.892	11.179	15	.740
Third Wave	Pakistan	.163	.936	4.1694	1.1291	3.0186	6.8888	11.3060	12.188	22.724	15	.090
	Punjab	.130	.944	2.5254	1.6591	1.6167	1.0410	8.6072	11.103	19.642	17	.293
	Sindh	.281	.732	1.9438	1.7271	1.2764	1.1967	9.9236	10.582	18.765	17	.342
	KPK	.326	.880	1.3085	2.0584	9.449	1.0875	4.0684	9.824	17.595	16	.348
	Balochistan	.171	.373	3.2398	3.9251	2.371	2.7611	1.3657	6.997	18.076	17	.384
	GB	.709	.755	9.862	6.7044	6.003	5.3216	5.958	4.618	15.464	17	.562
	AJK	.400	.760	2.533	3.0606	1.887	2.2497	6.736	6.777	8.060	11	.708
	Islamabad	.753	.909	7.836	2.3967	5.278	1.9945	3.6252	8.755	31.267	17	.019
Fourth Wave	Pakistan	.150	.826	4.423	1.040	3.3454	5.0512	1.1434	12.292	23.448	16	.102
	Punjab	.688	.893	1.537	1.5859	1.1911	1.7635	4.8134	10.121	10.256	17	.893
	Sindh	.011	.814	2.637	1.588	2.020	4.494	7.4783	11.201	21.837	17	.191
	KPK	.191	.658	1.188	2.384	9.088	1.4097	4.5612	9.605	19.579	17	.296
	Balochistan	.256	.584	3.289	5.783	2.195	6.1147	1.4184	7.040	12.468	17	.771
	GB	.423	.666	1.817	4.515	1.364	2.8671	5.6467	5.935	8.523	16	.932
	AJK	.368	.729	6.082	3.125	4.142	2.0754	1.998	8.364	18.435	15	.240
	Islamabad	.256	.591	8.668	2.757	6.158	2.0402	2.7718	8.966	26.381	17	.068

The provinces of Punjab and Sindh showed the fitted ARIMA models (1, 1, 2), (1, 1, 1), (1, 1, 7), (1, 1, 6), and (1, 1, 5) for the first to fourth waves of the pandemic. Specifically, the ARIMA (1, 1, 1) model was suitable for the first, second, and fourth waves, while the ARIMA (1, 1, 7) model was useful for the third wave. In Khyber Pakhtunkhwa (KPK), the ARIMA (1, 1, 2) and ARIMA (1, 1, 1) models were found appropriate for the first, second, third, and fourth waves. Similarly, in Balochistan, the ARIMA (1, 1, 1) model found best fitted for all four waves. In Gilgit-Baltistan, ARIMA (1, 1, 1) was found to be the best fit for all four waves. For Azad Jammu & Kashmir, ARIMA (1, 1, 2) worked best for the first and second waves, while ARIMA (1, 1, 7) and ARIMA (1, 1, 2) were best for the third and fourth waves. For Islamabad, the ARIMA (1, 1, 4) model was suitable for the first wave, while ARIMA (1, 1, 1) was for the remaining three waves of the pandemic.

Table 3 shows the fitted models for each dataset of COVID-19. The diagnostic tools revealed that MAE had the smallest value compared to RMSE and MAPE. Figure 5(a-d) shows residuals and gradient plots of all waves across the selected datasets. In this analysis, the smallest parameters $(p, q \leq 2)$ and greater degrees of freedom confirmed the suitability of ARIMA (1, 0, 1). For most of the selected datasets, ARIMA (1, 1, 1) and ARIMA (1, 1, 7) were found to be the best-fitting models for each wave. Various time series required differencing until they became stationary, with "d" which represents the degree of differencing to achieve stationarity.

The purpose of this research is to predict all four waves of the COVID-19 pandemic using the appropriate ARIMA models and a 14-day forecast for each dataset. The next 14-day predictions were made for each wave from 21 October 2020 to 4 November 2020, 16 March 2021 to 30 March 2021, 10 July 2021 to 24 July 2021, and 30 September 2021 to 14 October 2021 in Pakistan, with standard errors computed at a 95% confidence interval. The assessment of trained (observed) and predicted values for each wave of the Pakistan datasets is presented in Table 4. The performance of observed and predicted data is depicted in Figure 6 (a-d). The blue line represents the actual data values, the orange line shows the predicted data, and the two red lines indicate the upper and lower limits of the confidence interval (C.I.) for the standard deviation in all waves across Pakistan. These three lines illustrate the estimated standard deviation and variance of the predicted data for COVID-19 across all waves in Pakistan.

Table 4. Predicted and Observed Values of Four Different Waves of COVID-19 in Pakistan

Date	Forecasted Value	Original Value	Lower Bound	Upper Bound
1st wave of COVID-19 in Pakistan (21 Oct 2020 – 4 Nov 2020)				
21-Oct	736	736	736.00	736.00
22-Oct	736.29823	736	45.20	1427.40
23-Oct	736.59645	847	-127.69	1600.89
24-Oct	736.89468	832	-271.61	1745.40
25-Oct	737.1929	707	-397.67	1872.05
26-Oct	737.49113	773	-511.28	1986.27
27-Oct	737.78936	825	-615.61	2091.19
28-Oct	738.08758	908	-712.66	2188.83
29-Oct	738.38581	1078	-803.81	2280.58
30-Oct	738.68404	807	-890.05	2367.42
31-Oct	738.98226	977	-972.14	2450.10
1-Nov	739.28049	1123	-1050.64	2529.20
2-Nov	739.57871	1167	-1126.01	2605.16
3-Nov	739.87694	1313	-1198.61	2678.36
4-Nov	740.17517	1302	-1268.75	2749.10
2nd wave of COVID-19 in Pakistan (16 March 2021 – 30 March 2021)				
16-Mar	2351	2351	2351.00	2351.00
17-Mar	2347.9244	3495	1693.83	3002.02
18-Mar	2344.8489	3449	1687.43	3002.27
19-Mar	2341.7733	3874	1680.97	3002.57
20-Mar	2338.6977	3667	1674.47	3002.92
21-Mar	2335.6221	3669	1667.92	3003.32
22-Mar	2332.5466	3270	1661.32	3003.77
23-Mar	2329.471	3301	1654.68	3004.26
24-Mar	2326.3954	3946	1647.98	3004.81
25-Mar	2323.3198	4368	1641.24	3005.40
26-Mar	2320.2443	4468	1634.45	3006.04
27-Mar	2317.1687	4767	1627.61	3006.73
28-Mar	2314.0931	4525	1620.72	3007.47
29-Mar	2311.0175	4084	1613.78	3008.25
30-Mar	2307.942	4757	1606.80	3009.08
3rd wave of COVID-19 in Pakistan (10 July 2021 – 24 July 2021)				
10-Jul	1980	1980	1980.00	1980.00
11-Jul	1939.5712	1993	969.59	2909.55
12-Jul	1899.1424	1590	528.07	3270.21
13-Jul	1858.7136	1980	179.22	3538.21

Date	Forecasted Value	Original Value	Lower Bound	Upper Bound
14-Jul	1818.2848	2545	-121.67	3758.24
15-Jul	1777.856	2493	-391.95	3947.66
16-Jul	1737.4272	2783	-640.51	4115.36
17-Jul	1696.9984	2607	-872.63	4266.62
18-Jul	1656.5696	2452	-1091.75	4404.89
19-Jul	1616.1408	2145	-1300.27	4532.55
20-Jul	1575.712	2579	-1499.92	4651.35
21-Jul	1535.2831	2158	-1692.02	4762.59
22-Jul	1494.8543	1425	-1877.59	4867.30
23-Jul	1454.4255	1841	-2057.43	4966.28
24-Jul	1413.9967	2819	-2232.21	5060.20
4^{th} wave of COVID-19 in Pakistan (30 Sep 2021 – 14 Oct 2021)				
30-Sep	1411	1411	1411.00	1411.00
1-Oct	1400.4771	1664	455.31	2345.65
2-Oct	1389.9543	1656	159.02	2620.89
3-Oct	1379.4314	1490	-82.95	2841.82
4-Oct	1368.9086	1308	-293.45	3031.26
5-Oct	1358.3857	1212	-482.75	3199.52
6-Oct	1347.8629	1453	-656.54	3352.26
7-Oct	1337.34	912	-818.34	3493.02
8-Oct	1326.8172	955	-970.53	3624.16
9-Oct	1316.2943	767	-1114.78	3747.37
10-Oct	1305.7715	1004	-1252.35	3863.89
11-Oct	1295.2486	689	-1384.18	3974.68
12-Oct	1284.7258	1021	-1511.02	4080.48
13-Oct	1274.2029	1016	-1633.47	4181.88
14-Oct	1263.6801	1089	-1752.02	4279.38

Moreover, the highest correlation for each wave in Pakistan is greater than 0.8 between predicted and observed data, proving the model's sufficiency and consistency. This is depicted in Table 5 and Figure 7 (a-d), where the red line represents observed values, the blue line represents fitted values, and the dotted lines show the upper and lower control limits (UCL and LCL) for model predictions for the first to fourth waves of data across Pakistan, including Sindh, Punjab, Khyber Pakhtunkhwa, Balochistan, Gilgit-Baltistan, Azad Jammu & Kashmir, and Islamabad. The results indicate that the ARIMA-ML technique is highly effective for epidemic modelling.

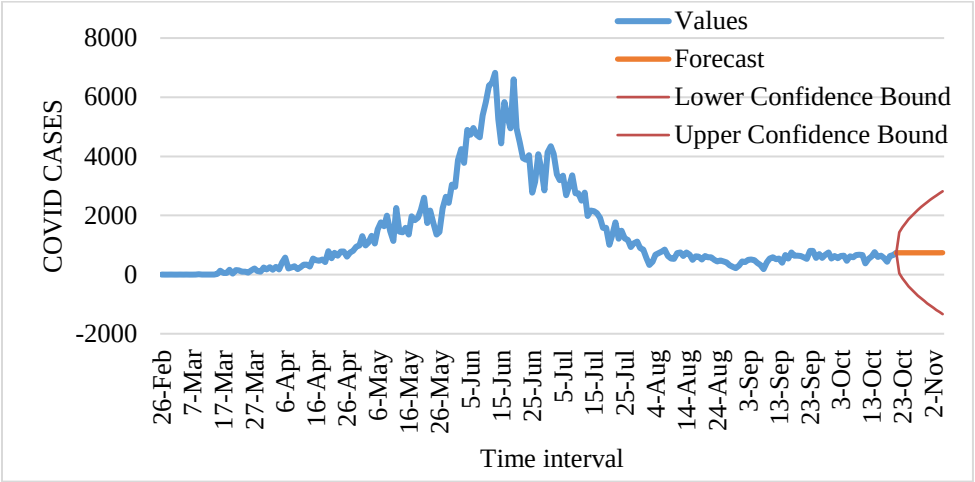


Figure 6 (a). Actual and forecasted values of the 1st wave of the pandemic in Pakistan ranging from 26 Feb 2020 to 20 Oct 2020 and 21 Oct 2020 to 4 Nov 2020

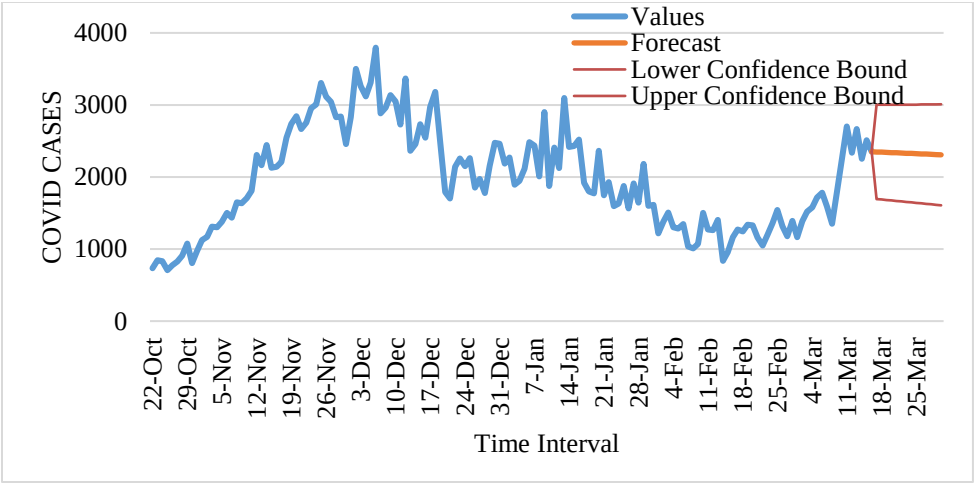


Figure. 6 (b). Actual and Forecasted Values of the 2nd wave of the Pandemic in Pakistan Ranging from 22 Oct 2020 to 16 March 2021 and 17 March 2021 to 30 March 2021

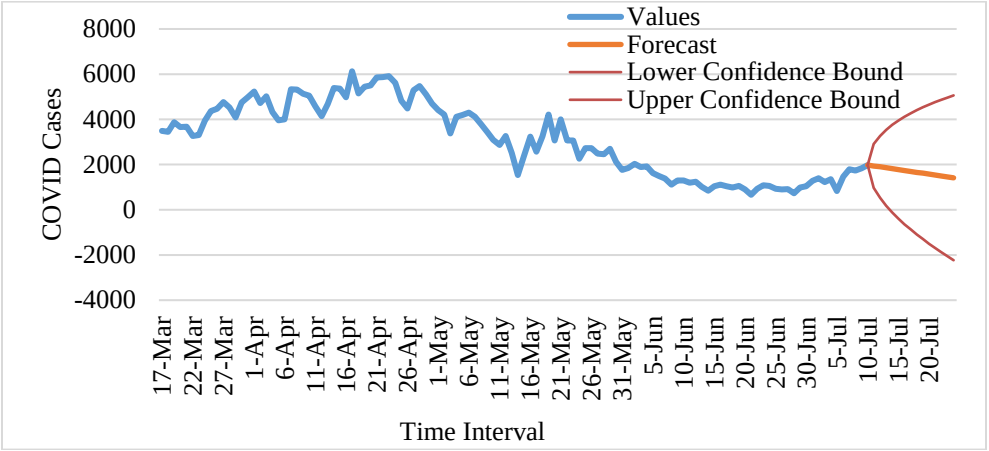


Figure. 6(c) Actual and Forecasted Values of 3rd Wave of the Pandemic in Pakistan Ranging from 17 March 2021 to 10 July 2021 and 11 July 2021 to 24 July 2021

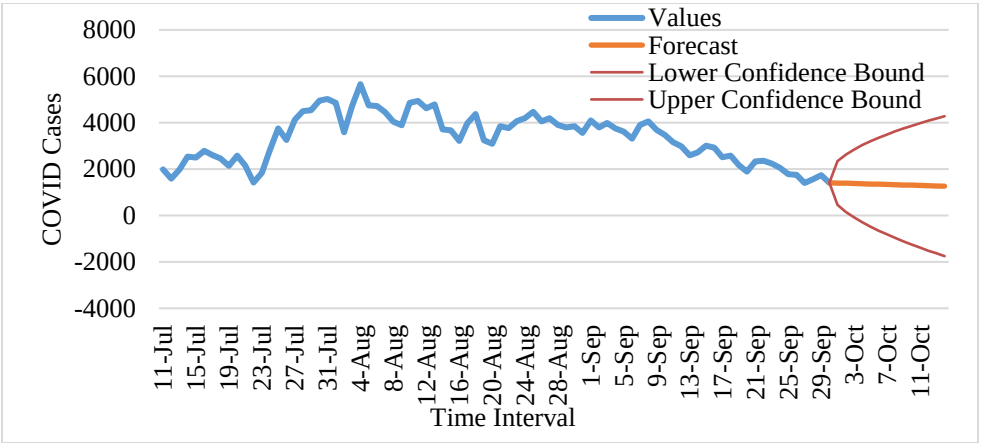


Figure. 6(d) Actual and Forecasted Values of the Fourth Wave of the Pandemic in Pakistan Ranging from 11 July 2021 to 30 Sep 2021 and 1st Oct 2021 to 14 Oct 2021.

Table 5. Correlation between Trained and Predictive Data Sets

Correlation	Observed 1st wave	Observed 2nd wave	Observed 3rd wave	Observed 4th wave
Predictive 1st wave	0.9564	-	-	-
Predictive 2nd wave	-	0.99236	-	-
Predictive 3rd wave	-	-	0.8789	-
Predictive 4th wave	-	-	-	0.8906

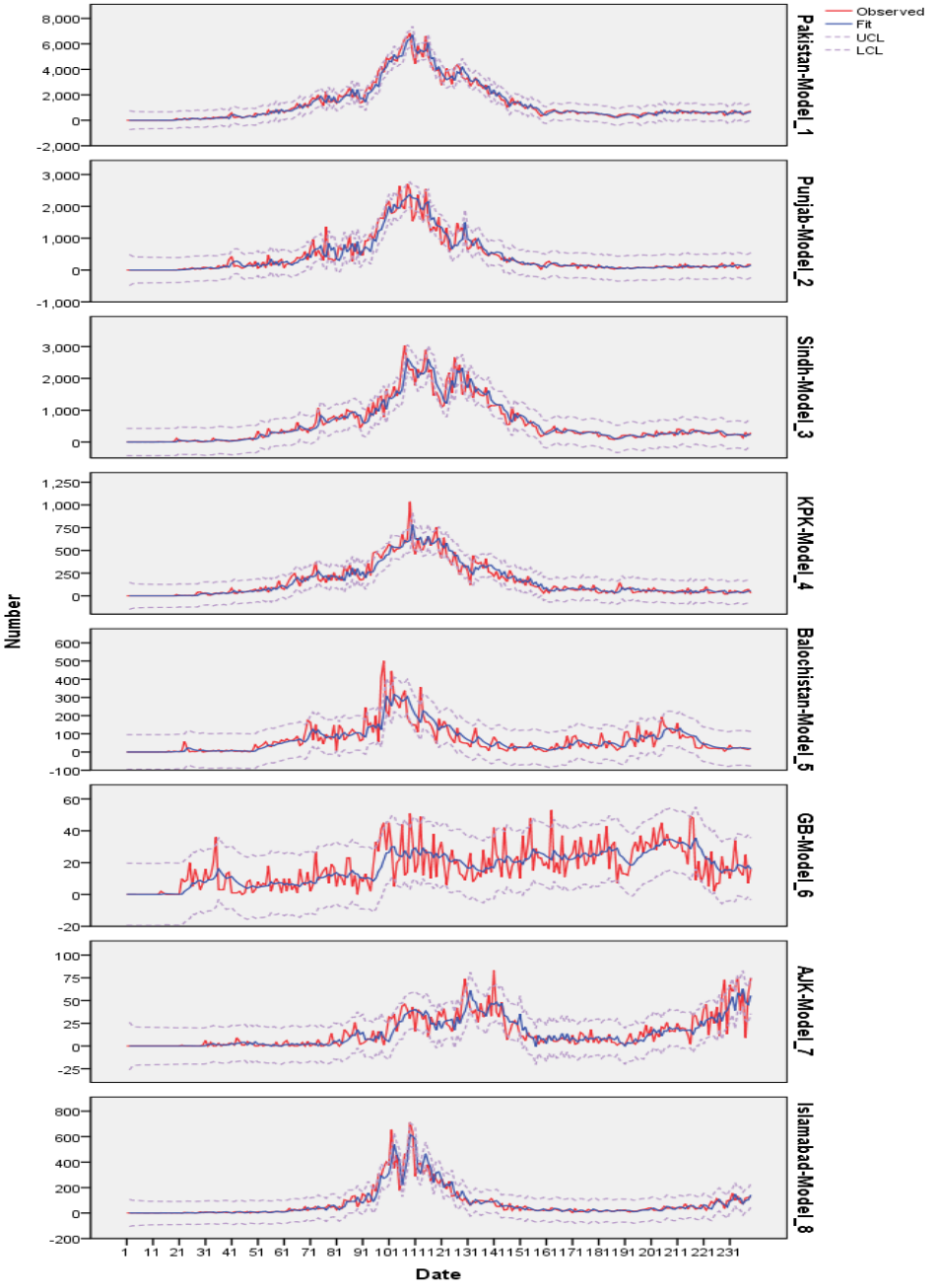


Figure 7 (a). Graph of Observed, Fit, UCL, and LCL Values of COVID-19 (from 26 Feb 2020 to 21 Oct 2020).

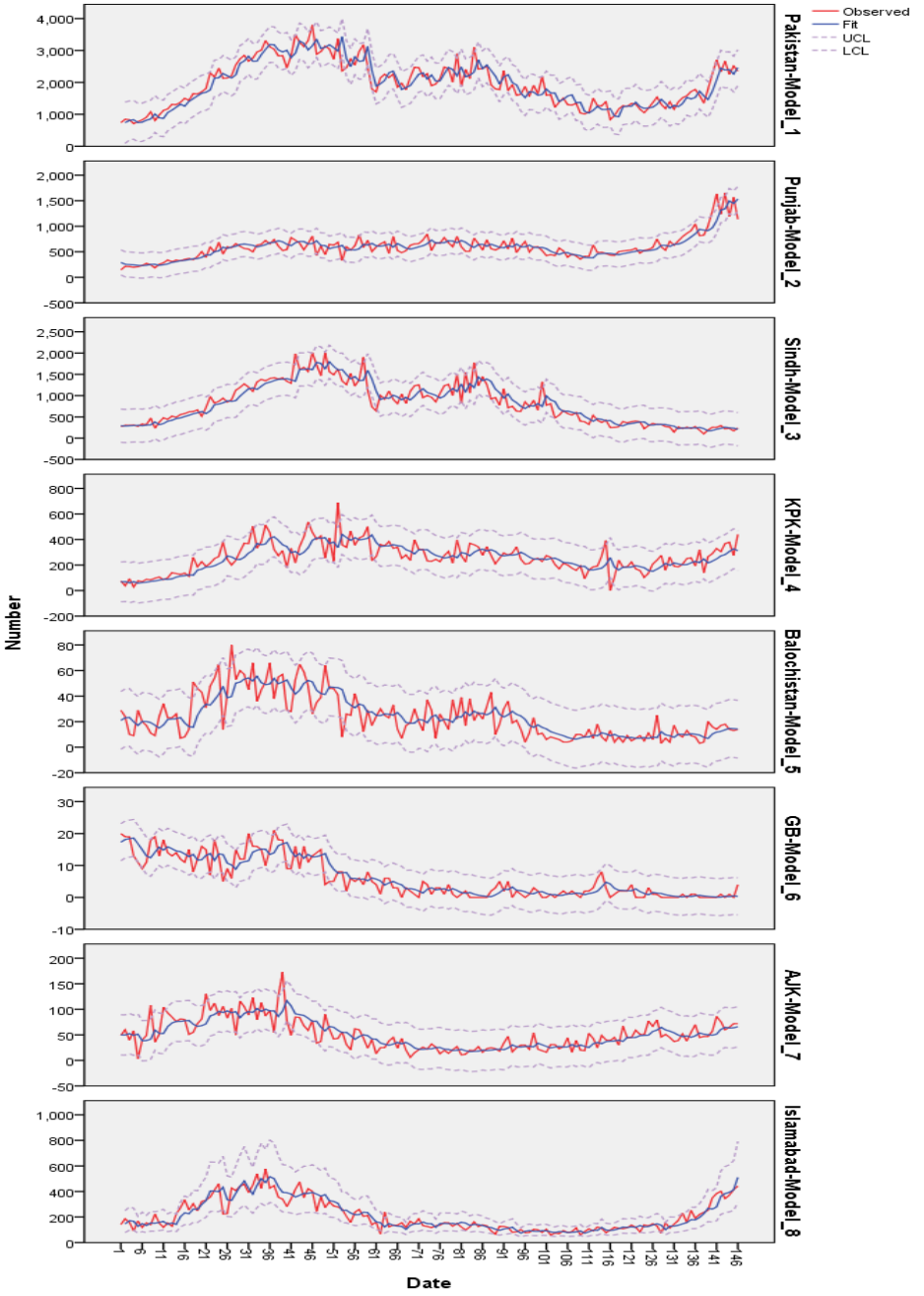


Figure 7 (b). Graph of Observed, Fit, UCL, and LCL Values of COVID-19 (from 21 Oct 2020 to 16 March 2021).

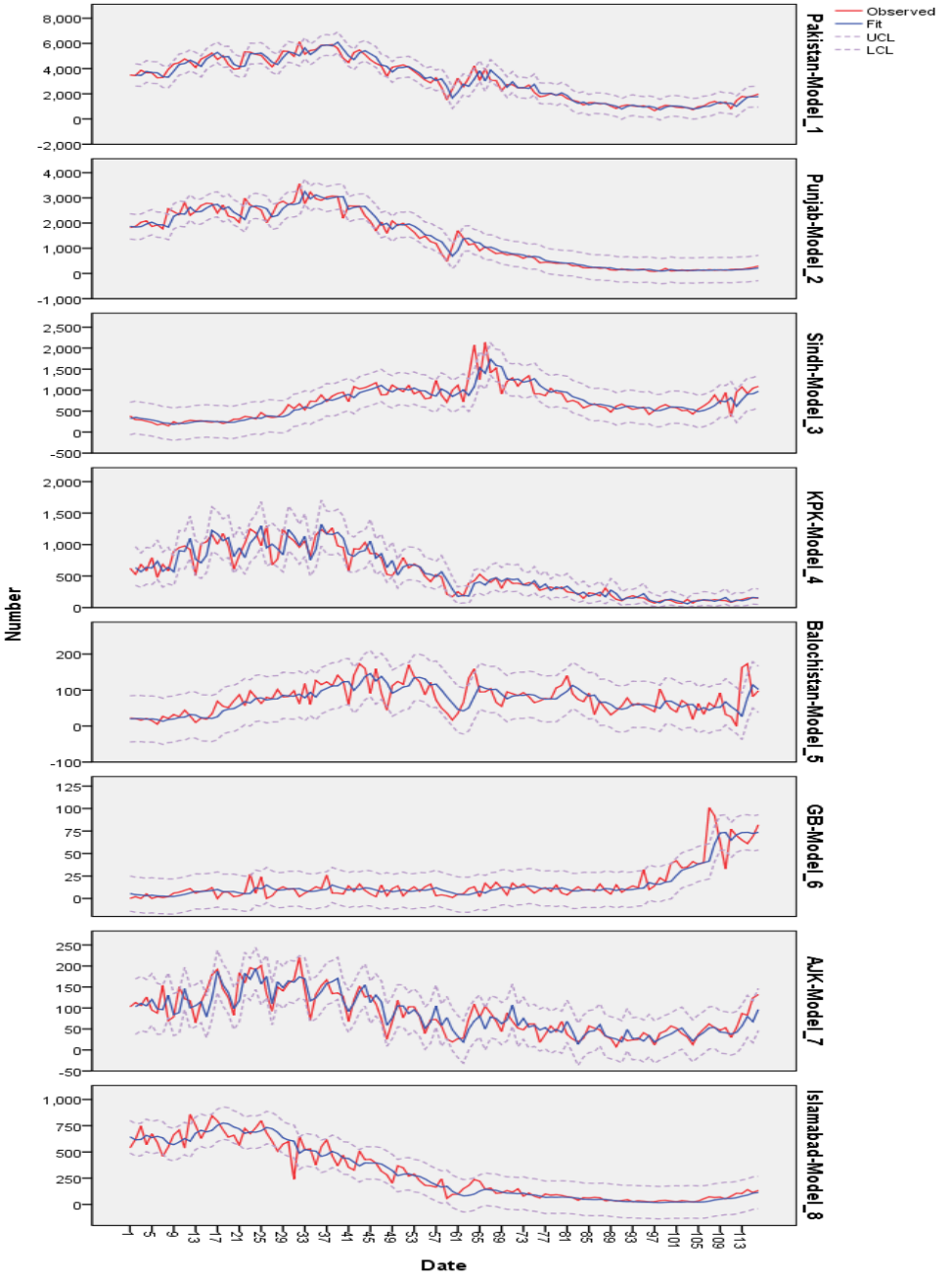


Figure 7 (c). Graph of Observed, Fit, UCL, and LCL Values of COVID-19 (from 17 March 2021 to 10 July 2021).

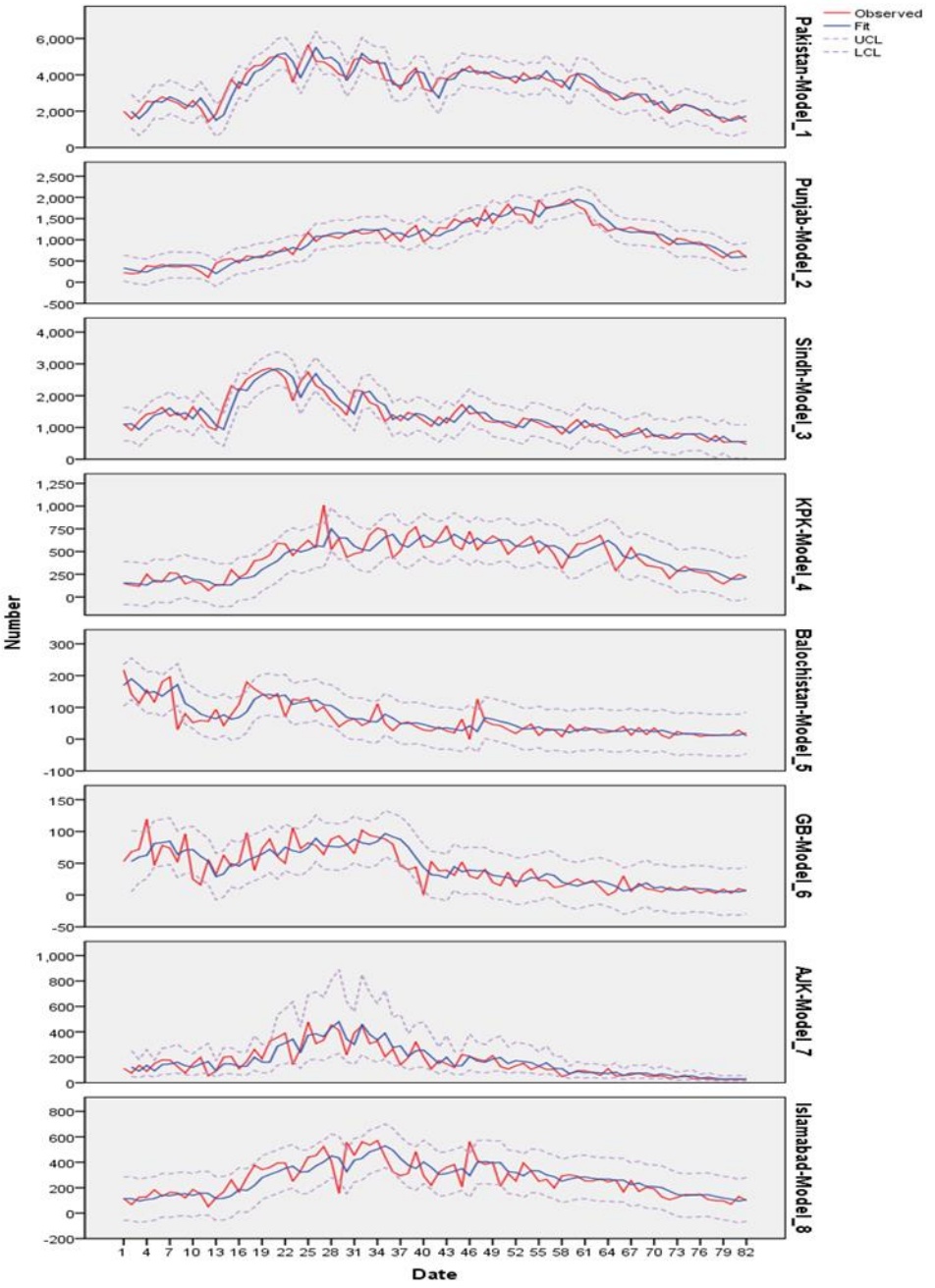


Figure 7 (d). Graph of Observed, Fit, UCL, and LCL Values of COVID-19 (from 11 July 2021 to 30 Sep 2021).

4.1. Conclusion and Future Work

The ARIMA-based ML model is a suitable method for exploring and modeling the exponential dynamics of COVID-19. ARIMA models have shown excellent accuracy in predictive analysis. In conclusion, these ARIMA-based ML techniques are not only useful for understanding COVID-19 but can also be applied to future pandemics and the prediction of other infectious diseases.

The proposed methodology was applied to univariate stochastic data series but extending it to multivariate data series presents an exciting area for future research. The ARIMA-based machine learning approaches have demonstrated promising results and can be seen as valuable additions to the existing literature on stochastic data analysis and prediction.

Looking ahead, researchers could explore the application of various other predictive models, such as Bayesian networks, Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs), for analyzing pandemic-related data and disease outbreaks.

CONFLICT OF INTEREST

The authors of the manuscript have no financial or non-financial conflict of interest in the subject matter or materials discussed in this manuscript.

DATA AVALIABILITY STATEMENT

Data availability is not applicable as no new data was created.

FUNDING DETAILS

No funding has been received for this research.

REFERENCES

1. Chyon FA, Suman MNH, Fahim MRI, Ahmmed MS. Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. *J Virol Methods*. 2022;301:e114433. https://doi.org/10.1016/j.jviromet.2021.114433
2. Feng Y, Hao W, Li H, Cui N, Gong D, Gao L. Machine learning models to quantify and map daily global solar radiation and photovoltaic power. *Renewable Sustain Energy Rev*. 2020;118:e109393. https://doi.org/10.1016/j.rser.2019.109393

3. Ilu SY, Prasad R. Time series analysis and prediction of COVID-19 patients using discrete wavelet transform and auto-regressive integrated moving average model. *Multimed Tools Appl*. 2024;83:72391–72409. https://doi.org/10.1007/s11042-024-18528-x
4. Roosa K, Chowell G. Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. *Theor Biol Med Model*. 2019;161:e1. https://doi.org/10.1186/s12976-018-0097-6
5. Contreras J, ARIMA models to predict next-day electricity process. *IEEE Trans Power Syst*. 2004;19(1):366–374. https://doi.org/10.1109/TPWRS.2002.804943
6. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes Metab Syndr Clin Res Rev*. 2020;14(4):337–339. https://doi.org/10.1016/j.dsx.2020.04.012
7. Chimmula VKR, Zhang L. Time series forecasting of COVID19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals*. 2020;134:e109864. https://doi.org/10.1016/j.chaos.2020.109864
8. Alboaneen D, Pranggono B, Alshammari D, Alqahtani N, Alyaffer R. Predicting the epidemiological outbreak of the coronavirus disease 2019 (COVID-19) in Saudi Arabia. *Int J Environ Res Public Health*. 2020;17(12):e4568. https://doi.org/10.3390/ijerph17124568
9. Sardar I, Akbar MA, Leiva V, Alsanad A, Mishra P. Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: ethodology, evaluation, and case study in SAARC countries. *Stoch Environ Res Risk Assess*. 2023;37:345–359. https://doi.org/10.1007/s00477-022-02307-x
10. Alkady W, ElBahnasy K, Leiva V, Gad W. Classifying COVID-19 based on amino acids encoding with machine learning algorithms. *Chemom Intell Lab Syst*. 2022;224:e104535. https://doi.org/10.1016/j.chemolab.2022.104535
11. Paparoditis E, Politis DN. The asymptotic size and power of the augmented Dickey–Fuller test for a unit root. *Econom Rev*. 2018;37(9):955–973. https://doi.org/10.1080/00927872.2016.1178887
12. Sujath RA, Chatterjee JM, Hassanien AE. A machine learning forecasting model for COVID-19 pandemic in India. *Stoch Environ Res Risk Assess*. 2020;34:959–972. https://doi.org/10.1007/s00477-020-01827-8

13. Kim S, Kim H. A new metric of absolute percentage error for intermittent demand forecasts. *Int J Forecast.* 2016;32(3):669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003
14. Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E. Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak. *Int J Interact Multimed Artif Intell.* 2020;6(1):132–140. https://doi.org/10.9781/ijimai.2020.02.002
15. Zhan C, Tse CK, Lai Z, Hao T, Su J. Prediction of COVID-19 spreading profiles in South Korea, Italy and Iran by data-driven coding. *PloS One*. 2020;15(7):e0234763. https://doi.org/10.1371/journal.pone.0234763