Int J Syst Assur Eng Manag (December 2023) 14(6):2626–2641
https://doi.org/10.1007/s13198-023-02127-4

ORIGINAL ARTICLE

SARS-CoV-2 forecasting using regression and ARIMA

Chaman Verma[1] · Purushottam Sharma[2] · Sanjay Singla[3] · Abhishek Srivastava[2] · Ruchi Sharma[4]

Received: 7 July 2022 / Revised: 19 July 2023 / Accepted: 9 August 2023 / Published online: 10 September 2023

Abstract Respiratory syndrome coronavirus 2 (SARS-CoV-2) is a pandemic coronavirus that is spreading quickly throughout the world. Every person experiences fear due to the unexpected pandemic (Covid-19), which is spreading quickly and affecting people. The affected patients' daily growth rate is accelerating. A predictive analysis was done to determine the potential number of deaths brought on by this pandemic. An official dataset of 35 states/UTs of India was investigated with predictive analysis (regression modeling). To predict the patient's deceased, both recovered and active cases have been impacted. This paper estimated future deceased counts based on active and recovered cases individually and jointly. The high positive linear correlation proved that the active and recovered case affected the patient's deceased rate. Regression models explored high aspects of the deceased ahead. Multiple linear regression has predicted a deceased patient based on active and recovered patients with significant $R^2 = 0.89$. Further, temporal dynamics of Covid-19 timing analyzed with Auto-Regressive Integrated Moving Average forecasted confirmed, active, recovered, and deceased cases for the next 40 days. According to the results, a high cure is still required for active and recovered patients, and the government should follow in obligatory footsteps to avoid more deceased predicted with the models.

Keywords Covid-19 · Correlation · Deceased prediction · Regression · ARIMA

✉ Purushottam Sharma
puru.mit2002@gmail.com

Chaman Verma
chaman@inf.elte.hu

Sanjay Singla
dr.ssinglacs@gmail.com

Abhishek Srivastava
abhishek.sri13@gmail.com

Ruchi Sharma
rs.sharma184@gmail.com

1 Department of Media and Educational Informatics, Eotvos Lorand University, Budapest, Hungary

2 Department of Information Technology, ASET, Amity University Uttar Pradesh, Noida, India

3 Department of Computer Science and Engineering, UIE, Chandigarh University, Mohali, India

4 Department of MIS and Analytics, International Management Institute, Kolkata, West Bengal, India

1 Introduction

Since the end of 2019, the coronavirus, which is causing the pandemic of the century, has been spreading extremely quickly. After being first discovered in December 2019, it infected a Chinese person in Wuhan, China, on January 30, 2020 (Almendros-Jimenez et al. 2021). On March 11th, the World Health Organisation (WHO) declared this new pneumonia outbreak to be a "global pandemic" and gave it a new name, Covid-19 (Hao and Park 2021; Stoecklein et al. 2022). The term is officially known as "severe acute respiratory syndrome coronavirus 2" (SARS-CoV-2) by the International Committee on Taxonomy of Viruses. Covid-19 has been transferred from animal to human; presently, it is scattered across all continents. The lack of any preventive vaccine proved dangerous to human life.

This global pandemic hardly impacts all sectors, such as education, hospitality, transportation, trading, etc., and many more. As of 13th June 2020, out of a total of 7,553,182 confirmed cases, 423,349 deaths have been confirmed in the whole world. Several scientists from all over the world have

been attempting to forecast Covid-19 cases. Utilising both forward prediction and backward inference, epidemic development trends in South Korea, Italy, and Iran were predicted (Chitra and N., R. Shanmathi, and Dr. R. Rajesh. 2015). Handling curve fitting and recurrent neural network future Covid-19 positive cases and confirmed cases were identified in India (Gao et al. 2020).

The statistical forecasting models are helpful in controlling and predicting this global pandemic (Kumar et al. 2023). In this study, ARIMA was used to predict the Covid-19 trend. The Box-Jenkins approach, also referred to as the ARIMA model (Li et al. 2020) engaged in forecasting and analysis (Mohler 1990; N et al.. 2020; Rishabh et al. 2019). We found the ARIMA model applied appropriate forecasting in Covid-19 (Kotlyar et al. 2019; Garcia-Flores et al. 2022; Izquierdo-Pujol et al. 2022; Male 2022) cases in IRAN (Sharma et al. 2021, 2022a, 2016).

2 Related work

Table 1 compares the current research plan with the existing research, which used the ARIMA model to forecast Covid-19.

India has 28 states and 8 Union Territories (UT), and Covid-19 began to affect those areas on January 30, 2020, when the first case of Chinese origin was noted. Up to 13th June, 35 states and UTs infected in India. Figure 1 shows

Table 1 The extant research with the present study

Parameter	Moftakhar et al	Moftakhar and Seif	T.T. Tran et al	Present Study
Database	41 days (19th February-30th March)	32 days (19th February-21st March)	75 days (February 20-May 04)	13 days (1st June-13 June)
Forecasting	30 days (31st March -29th April)	30 days (22nd March -20th April)	20 days (5th May-24th May)	40 days (13th June 23rd July)
Model's parameter	(ARIMA (0,1,0))	(ARIMA (0,1,0))	(ARIMA (0,2,0)), (ARIMA (0,1,0)), (ARIMA (1,1,0))	(ARIMA (0,1,0))
Variables	total confirmed, deaths cases with a growth rate	total confirmed, deaths cases with a growth rate	total confirmed, deaths cases with a growth rate	total confirmed, deaths, recovered, and active cases with trends
Regression (Linear and Multilinear)	No	No	No	impact of active and recovered cases on deceased, Correlations of deceased with active and recovered cases
Country	Iran	Iran	Iran	India

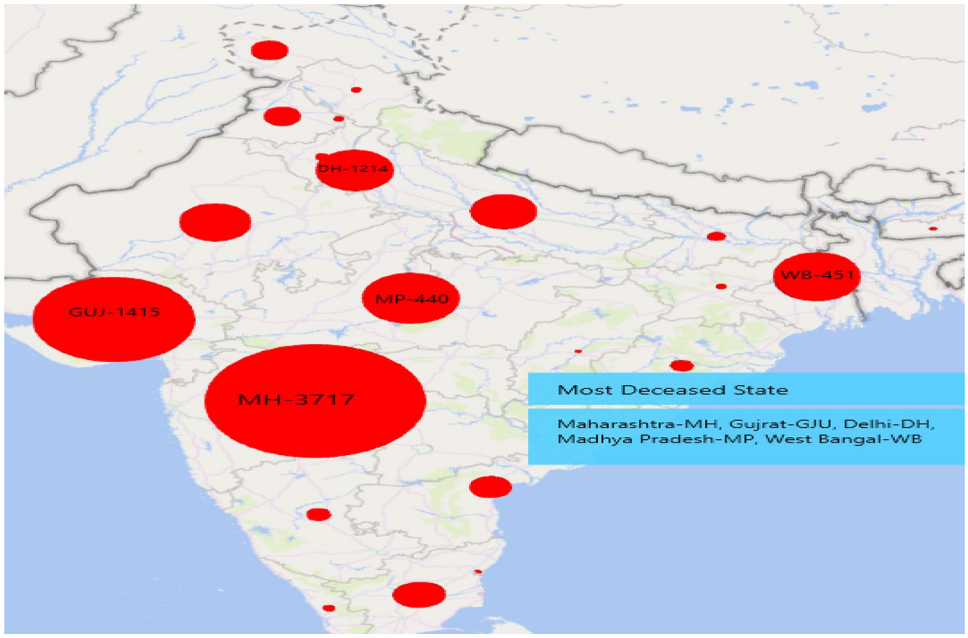

Fig. 1 Most infected states on map as of 13 June 2020

the red-hot spot on the Indian map due to the maximum deceased. The top 5 states are identified based on the overall death tolls. Due to Covid-19, the highest harmed state is Maharashtra (MH), with 3717 deceased as shown in Fig. 1 (Kumar et al. 2023). The second place of Gujrat (GUJ) state also suffered from a significant loss of human beings with 1415 cases. The capital of India, Delhi (DH) comes in the third position in loss of humans about 1214. Further, Madhya Pradesh (MP) and West Bengal (WB) also have 440 and 451 deceased cases, respectively.

Figure 2 depicts an increasing number of confirmed, active, recovered, and deceased cases. Until 13th June, there were 301,009 confirmed cases in total; active cases are 137,795, and recovered cases are 154,330 noted. Gradually increasing Covid-19 cases (Rizzo et al. 2022; Tatura 2022; Bungaro et al. 2022; Boudry et al. 2022; Rujen et al. 2023; Sharma et al. 2022b) became matters of anxiety, not only for the government but also for the average human. A total of 8102 deceased cases were reported caused by COVID-19 (Tomar and Gupta 2020).

Figure 3 visualizes the current scenario of Covid-19 patients having confirmed, active, recovered, and Fig. 3 shows the deceased cases in 35 states/UTs. The highest number of deceased are reported, 3717 in Maharashtra state, which seems to be a red-zone area highly. The second-largest number of deceased, 1415 reported in Gujrat (GJ). The third-largest death count, 1214 cases reported in the capital of India.

The primary concepts of this paper are: to discover the impact of active cases on the deceased, identification of active, recovered, confirmed, and deceased cases, state-wise decease predictions based on active and recovered cases, and association of deceased with active and recovered cases.

There are five sections in this paper. Section 1 discussed the Covid-19 introduction theory by concentrating on a recent effect on Indian provinces. Section 2 outlines our contribution to this work. Section 3 focused on objectives with hypotheses, designs conceptual schema with the methodology used. Section 4 is dedicated to experiments performed with discussion. Section 5 discusses the significant limitations of the paper Sect. 6 concludes the study's primary essence, including future work.

3 Contribution

This paper is written to help government officials and policymakers become aware of early detection of cases of Covid-19 in different provinces. They might use these results to prepare future cure and prevention mechanisms to defend against this pandemic. With the online deployment of this model, early detection of deceased, active, recovered and

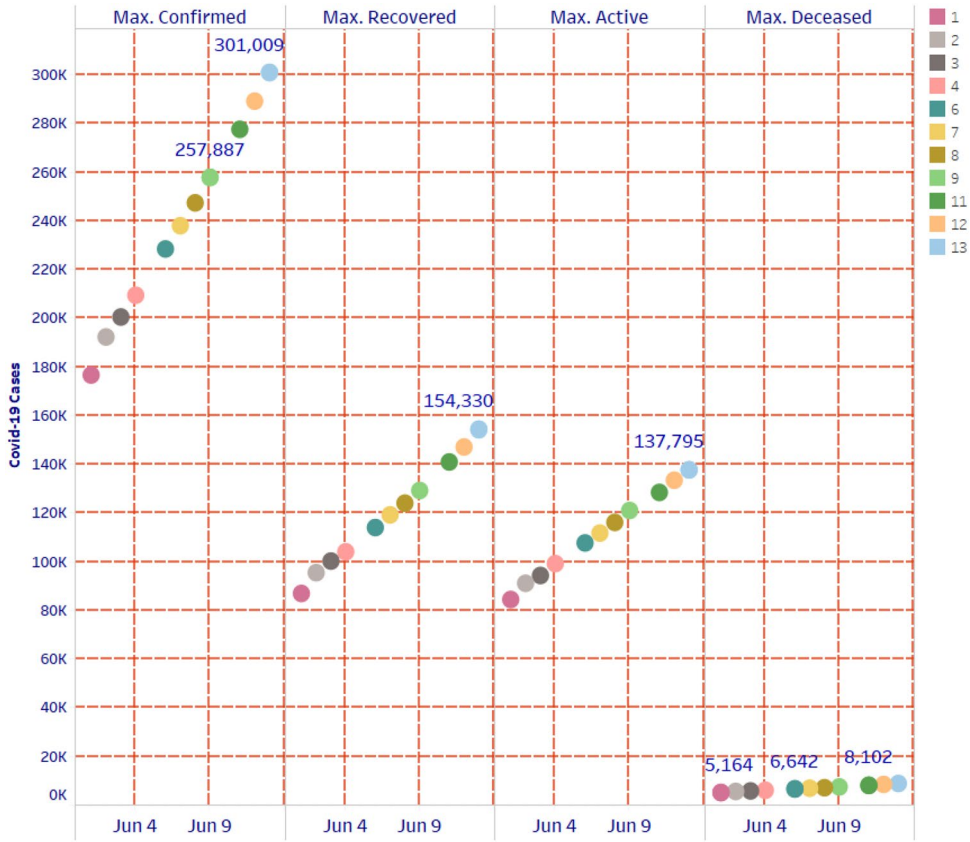

Fig. 2 Cumulative trend of Covid-19 (From1st June to 13th June)

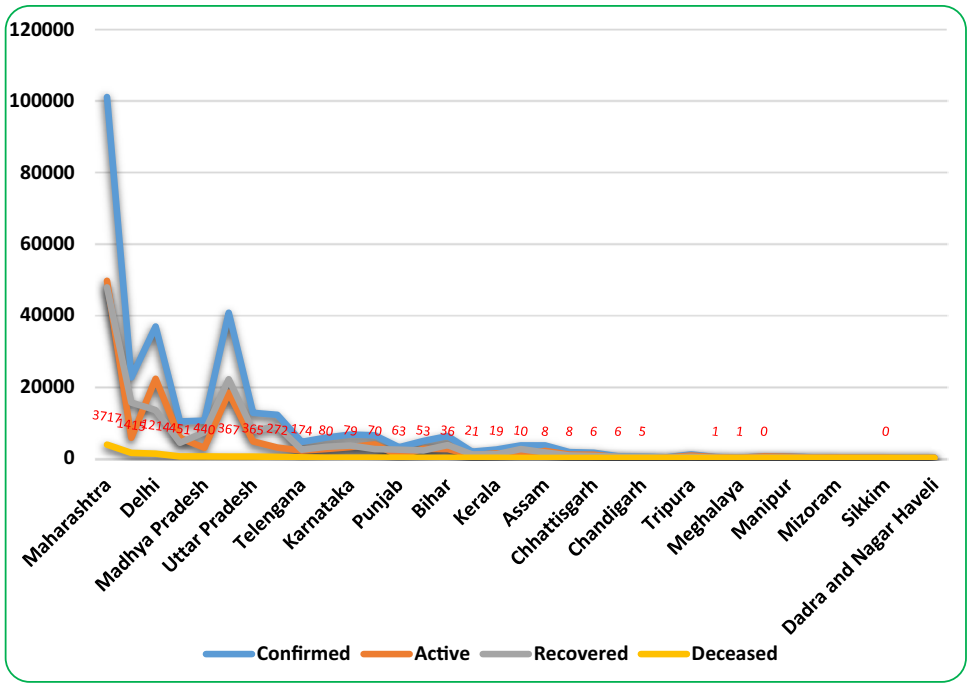

Fig. 3 State-wise Covid-19 scenario (1st June to 13th June 2020)

confirmed cases might be estimated. Hence, we need to propose an optimistic model. Using regression, we found useful information that the active and recovered cases positively impacted the deceased rates in each province. For this, the MLR obtained the highest accuracy of 89% in the early detection of the deceased. With a significant R-value of 0.927, we discovered a positive linear association between deceased patients and active cases that demonstrates the acceleration of deaths based on the sharp rise in active cases. We explored that there was a significant linear relationship between deceased patients with recovered cases and active cases. Additionally, we presented regression models that predicted deceased cases ($p < 0.05$), and we also applied the ARIMA model that identified deceased case cases more accurately than regression. We also demonstrated that the ARIMA model is superior to regression methods for time series data. Additionally, the temporal dynamics behavior of Covid-19 was analyzed with the ARIMA model (Sharma et al. 2016), which forecasted (Sharma et al. 2022a) the 40 days of Covid-19 cases.

4 Research design and methodology

4.1 Objectives

To discover the association of deceased cases with recovered and active cases.

(a)RH0: $\rho XY = 0$ {No association between deceased patient and recovered cases.}

(b)RHa: $\rho XY \neq 0${An association between deceased patient and recovered cases.}

(c)AH0: $\rho XY = 0$ {No association between deceased patient and active cases.}

(d)AHa: $\rho XY \neq 0${An association between deceased patient and active cases.}

To explore the impact of recovered cases on the deceased.

(e)ERH0: {No effect of recovered cases on the deceased.}

(f)ERHa: {An effect of recovered cases on the deceased.}

To examine the effects of active cases on the deceased.

(g)EAH0: {No effect of active cases on the deceased.}

(h)EAHa: {An effect of active cases on the deceased.}

To estimate the total number of confirmed, active, recovered, and deceased cases over the course of the following 40 days.

To estimate the deceased prediction based on active cases and recovered cases.

4.2 Conceptual design

We visualized the present research design in Fig. 4, which presents the schematic diagram with the conceptual idea of the study. This paper analyzed the impact of active and recovered on the deceased. Also, the association of the deceased has been found with the same variable. Regression modelling was used to forecast and examine the covid-19 in this. Based on active and recovered cases, regression analysis (LR and MLR) is applied after the fulfillment of assumptions. Three objectives (Impact, relationship, and prediction) need to be accomplished using regression. We also used the ARIMA forecasting model to predict the trend

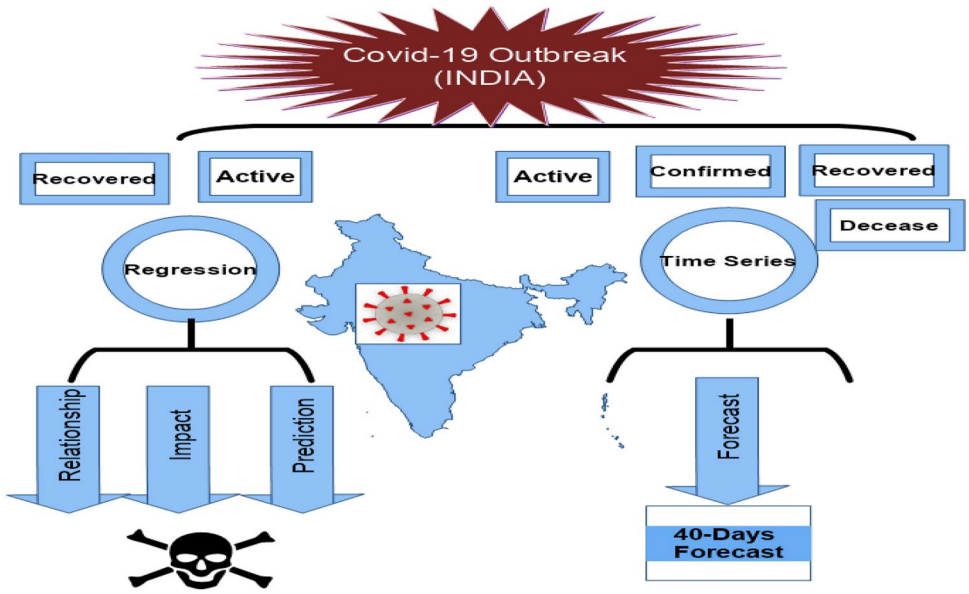

Fig. 4 Covid-19 out-break detection schema

Table 2 Date-wise cases

	Confirmed	Active	Recovered	Deceased
1-Jun	176,652	84,504	86,984	5164
2-Jun	192,292	91,167	95,527	5598
3-Jun	200,492	94,374	100,303	5815
4-Jun	209,436	99,254	104,107	6075
5-Jun	219,160	103,350	109,462	6348
6-Jun	228,465	107,750	114,073	6642
7-Jun	238,023	111,801	119,293	6929
8-Jun	247,422	116,192	124,095	7135
9-Jun	257,887	121,102	129,314	7471
11-Jun	277,648	128,517	141,029	8102
12-Jun	289,220	133,527	147,195	8498
13-Jun	301,009	137,795	154,330	8884

of all four cases (confirmed, active, recovered, and deceased) for the following 40 days while taking into account the time-series analyses.

4.3 Dataset description

The present study continually used standard and official data (Tomar and Gupta 2020) from 30th January to 13th June 2020. The five important variables in the dataset are state, confirmed, recovered, active, and deceased. All variables are scale types, with the exception of the state variable. At the end of each day, the data set is updated with the most recent information for 33 Indian states and UTs. Using the Cronbach alpha test, the reliability of data samples is calculated as 0.841. The reliability of data samples is calculated by 0.841 using the Cronbach alpha test. Table 2 shows the recent thirteen days' data of cases reported, from 1st June to 13th June mid-night. Table 3 stores the cases reported from the 35 states/UTs.

4.4 Statistical characteristics of variables

The significant dataset's statistical characteristics are necessary before the analyses. It shows the mean or average (μ) in Eq. (1), dispersion in Eq. (2) properties of data samples.

$$Sum = \sum_{i=1}^{n} Xi \tag{1}$$

$$Std.Deviation = \sqrt{\frac{|x-u|^2}{N}} \tag{2}$$

There are N data points in the population, where x is one of the values in the data set, and is the mean (μ) of the data set.

Figure 5 shows the essential statistical properties of confirmed, active, recovered, and deceased variables. 101,141 confirmed cases are the most, with an 8600 deviation value. A less deviation value of 253.8 can be seen in the highest amount of 3717 deceased cases. The standardized values of the four variables are used (Z-score). The term "standard score" is usually used for normal populations; the term "Z score" should only be used for normal distributions. We transformed all variables into the standardized form using Eq. (3):

$$Z = (X - \mu)/\sigma \tag{3}$$

Table 3 State-wise cases

State/UTs	Confirmed	Active	Recovered	Deceased
Maharashtra	101,141	49,628	47,796	3717
Gujarat	22,527	5619	15,493	1415
Delhi	36,824	22,212	13,398	1214
West Bengal	10,244	5587	4206	451
Madhya Pradesh	10,443	2802	7201	440
Tamil Nadu	40,698	18,284	22,047	367
Uttar Pradesh	12,616	4642	7609	365
Rajasthan	12,068	2898	8898	272
Telengana	4484	2032	2278	174
Andhra Pradesh	5680	2495	3105	80
Karnataka	6516	2997	3440	79
Haryana	6334	3789	2475	70
Punjab	2986	641	2282	63
Jammu and Kashmir	4730	2591	2086	53
Bihar	6103	2480	3587	36
Uttarakhand	1724	756	947	21
Kerala	2322	1303	1000	19
Odisha	3498	1014	2474	10
Assam	3498	1953	1537	8
Jharkhand	1617	937	672	8
Chhattisgarh	1429	873	550	6
Himachal Pradesh	486	183	297	6
Chandigarh	334	43	286	5
Puducherry	157	88	67	2
Tripura	961	682	278	1
Ladakh	239	176	62	1
Meghalaya	44	21	22	1
Goa	463	394	69	0
Manipur	385	308	77	0
Nagaland	156	107	49	0
Mizoram	104	103	1	0
Arunachal Pradesh	67	63	4	0
Sikkim	63	61	2	0
Andaman and Nicobar	38	5	33	0
Dadra and Nagar Haveli	30	28	2	0

We checked the multicollinearity problem among independent variables with Tolerance (T) (Tran et al. 2020). It is calculated with 1-R^2, and the maximum value of T depicts the lowest collinearity. Also, the Variance Inflation Factor (VIF) is calculated by inverting the T. The maximum value justifies low collinearity. Table 4 stores two critical metrics for the multicollinearity problem. For both independent variables, $T = 0.18$ and VIF = 5 instruct to accept moderate collinearity.

4.5 Regression and correlation

We used the regression methods LR and MLR in the prediction task, which explored influence after modeling. We constructed three predictive models (LR-1, LR-2, and MLR). Below Eq. (5) shows the general equations of LR. During model LR-1, Y is deceased, X is recovered cases, coefficients (a, b) of predictor recovered cases to explain the model, and ε is the error term.

$$Y = \mathrm{b} + \mathrm{a.X} + \varepsilon \tag{4}$$

$$Y = \mathrm{b} + \mathrm{a.X} + \varepsilon \tag{5}$$

$$b = \frac{\sum(X - \bar{X})\,(Y - \bar{Y})}{\sum(X - \bar{X})^2} \tag{6}$$

$$a = \bar{Y} - b.\bar{X} \tag{7}$$

$$\varepsilon = \sum(Y_1 - Y_{fit}) \tag{8}$$

In the model LR-2, we set Y = deceased, X is active cases, coefficients (a, b) of active predictor cases to explain the model, and ε is the error term. In Eqs. (6) and (7), the regression coefficient indicates the amount by which change in independent variable X must be multiplied to provide an average update in Y. Also, the amount of Y alters for a unit increase in X forces changes in slope. In Eq. (8), The difference between the predicted value and fit value of the dependent variables is used to calculate the total prediction error. The standard error of the slope SE(b) depicted in Eq. (9) and residual standard deviation S_{res} is shown in Eq. (10).

$$SE_{(b)} = \frac{S_{res}}{\sqrt{\sum X - \bar{X}}} \tag{9}$$

$$S_{res} = \frac{\sum(Y - Y_{fit})^2}{\sum n - 2} \tag{10}$$

In Eq. (11), we fit values for the MLR, where $\hat{Y}$ is deceased, X_1 is active cases, X_2 is recovered cases, coefficients (b_0, b_1, b_2) of predictors active and recovered cases to predict the deceased model, and ε is the error term.

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \varepsilon \tag{11}$$

Pearson Correlation in Eq. (12) is used to discover the association of deceased cases with recovered and active cases. Where SP is the total deviation score of the deceased, recovered, and later for active cases, and R is the correlation. The sum of the squared deviations for recovered cases

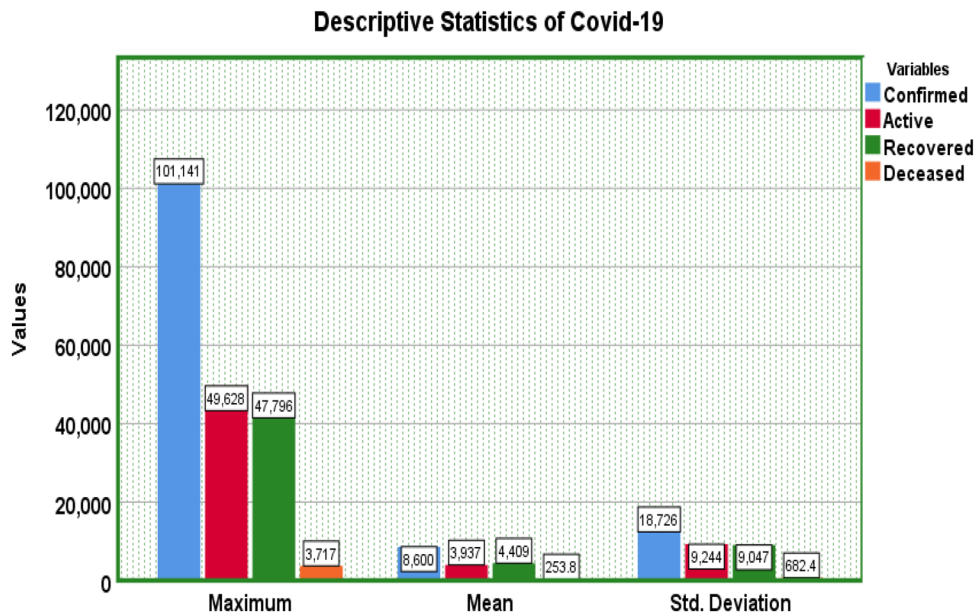

Fig. 5 Statistical characteristics of dataset

Table 4 Multicollinearity

	T	VIF
Recovered	.18	5
Active	.18	5

is SSy, and the sum of the squared deviations for deceased cases is SSx.

$$R = \frac{SP}{\left(\sqrt{SSx}\right)\left(\sqrt{SSy}\right)} \tag{12}$$

4.6 ARIMA model

For early detection of covid-19 cases, we used the time-series forecasting ARIMA model in IBM SPSS statistics 25. This model gains information from the dependable variables itself to estimate the trends. A time series, or collection of observations obtained by repeatedly measuring a single variable over time, was used in this model. The ARIMA model predicts future covid-19 cases based on previously known time-series values in the covid-19 dataset (Sharma et al. 2016). The common ARIMA forecasting equation is shown below in Eq. (13).

$$\mathrm{Y} := \mathrm{ARIMA}(\mathrm{p}, \mathrm{d}, \mathrm{q}) \tag{13}$$

- p is no. of lags of the dependent variable,
- d is no. of differences to become a stationarity variable, and
- q is no. of lags of the error term.

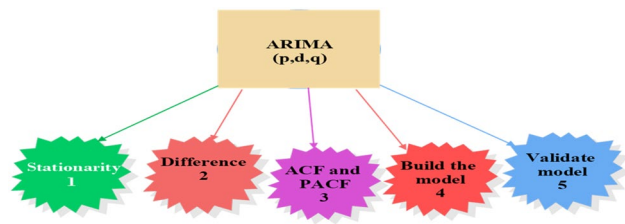

Fig. 6 ARIMA process

The base equation of ARIMA is shown in Eq. (14), where moving average parameters (θ's),

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q} \tag{14}$$

We made the following Eq. (15) of the ARIMA model where Y is a confirmed dependent variable and ŷ1, ŷ2…… ŷ40 are days to be identified with forecasting series.

$$\mathrm{Y} : \hat{y}_1, \hat{y}_2 \ldots\ldots \hat{y}_{40} = \mathrm{ARIMA}(\mathrm{p},\ \mathrm{d},\ \mathrm{q}) \tag{15}$$

Our model provides the forecasting model for all four variables, and Eq. (16) depicts the value of $p = 0$ describes no autoregressive, d = 1 shows difference, and $q = 0$ states no seasonal moving average parameter,

$$\mathrm{Y} : \hat{y}_1, \hat{y}_2 \ldots\ldots \hat{y}_{40} = \mathrm{ARIMA}(0, 1, 0) \tag{16}$$

Figure 6 illustrates the five crucial steps that were taken to build ARIMA and forecast Covid-19 cases.

We have tested the component's seasonality and found the stationary data to use ARIMA to forecast. ARIMA model is also used to make data stationaries through

differencing in lack of stationaries. Further, using the correlograms, we tested the autocorrelation using Auto Correlation Factor (ACF) and Partial Auto Correlation Factor (PCF). Later, the ARM model was built and validated using appropriate metrics. Figure 7 shows the correlograms with ACF and Partial Auto Correlation Factor PCF of (a) Confirmed (c) Active (e) Recovered (g) Deceased and PCF of (b) Confirmed (d) Active (f) Recovered (c) Deceased against various lags at difference 1. An ACF calculates and displays the average correlation between data points in a four-variable time series and earlier series values calculated with various lag lengths. In contrast to the ACF, the PACF uses correlation to account for any correlation between observations made at shorter lags. The four variables are found stationary because series autocorrelation lies near zero below the lines and insignificant

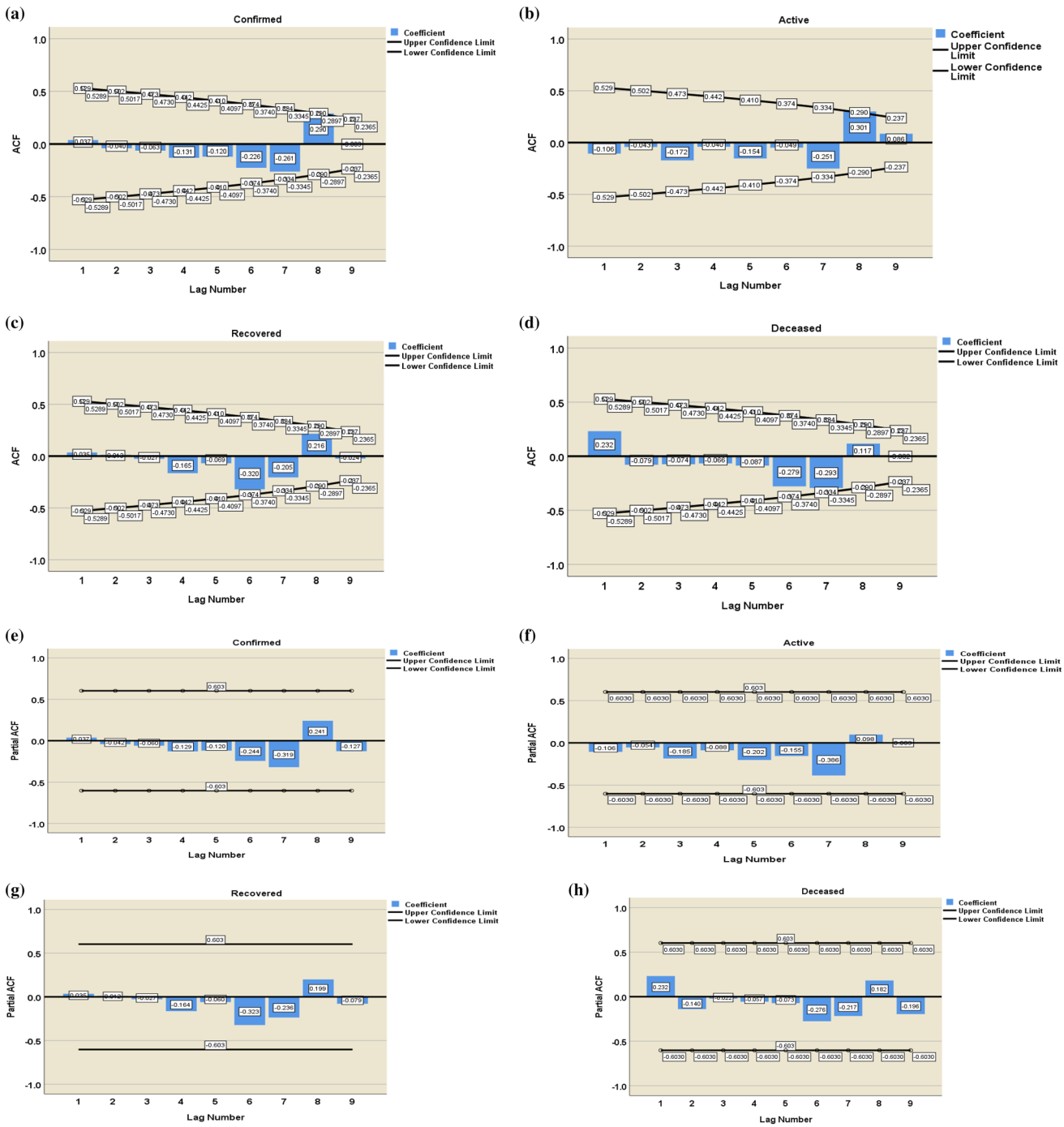

Fig. 7 ACF **a** confirmed **b** active **c** recovered **d** deceased, and PACF **e** confirmed **f** active **g** recovered **h** deceased

relationships (Wang et al. 2020; Zhang et al. 2020; Yang et al. 2021).

5 Results and discussion

This section discussed the experimental findings with validation metrics after using the processed dataset to implement regression models in IBM SPSS statistics 25. To discover the relationship between deceased patients with recovered and active cases, 2–tailed Pearson correlation is applied at a 0.01 level of significance. Figure 8 displays the positive linear correlation between deceased patients and active cases, with a significant R-value of 0.927 reflecting the death enhancement based on the rapid increase of active cases.

Due to the highest $R = 0.988$, active cases are increasing with the growth of confirmed. Also, the confirmed and active cases found positively correlated with recovered cases 0.988, and 0.954, respectively. Also, deceased cases are related to active, confirmed, and recovered cases. Enhancement in the deceases can be seen based on recovery caused by $R = 0.936$. Further, we observed that if cases are increasing actively still, there is a significant recovery of patients ($R = 0.954$). Thus, the null hypothesis RH0: $\rho XY = 0$ "No association between deceased patient and recovered cases" is failed to accept. Therefore, the alternative hypothesis RHa: $\rho XY \neq 0$ "An association between deceased patient and recovered cases" is accepted. Thus, a significant linear relationship between deceased patients and recovered cases is observed. To test the null hypothesis AH0: $\rho XY = 0$ "No association between deceased patient and active cases", a high positive correlation is found $R = 0.927$, and the cause failed to accept. Its alternative hypothesis AHa: $\rho XY \neq 0$ "An association between deceased patient and active cases" is accepted. Hence, a significant linear correlation was explored between deceased patients and active cases.

Further, the effect (individual and combined) of both active and recovered cases on the deceased rate is explored. We built three regression models to the standardized value of variables. One side, the LR-1 model's findings signify the impact of active cases on patient deceased, and contrasted with, the LR-2 model explains the power of recovered cases to predict the deceased.

Table 5 compares the critical parameters of both LR models. We observed the highest correlation R and goodness of fitting by LR-2 model ($0.927 < 0.937$) ($0.859 < 0.877$). Therefore, the LR-2 model predicted the deceased higher than the LR-1. The maximum coefficient of determination of LR-2 also proved the predictive strength of the deceased patient model. A significant t-values ($P < 0.005$) might be useful in hypothesis testing. These metrics demonstrated that active cases predicted the deceased of patients more accurately.

Table 6 equates the ANOVA results of LR-1 and LR-2. The residual error of LR-2 lowered as compared to LR-1 ($4.1 < 4.8$). Both model's F-values (200.7, 235.5) were found significant ($P < 0.005$). The LR-2 model significantly reduced the residual error and proved its explanatory power.

Table 5 Individual impact of active and recovered cases on the deceased

	R	R2	Std. Error	t	Sig. (P)
LR-1 (Active based)	.927	.859	.281	14.2	.000
LR-2 (Recovered based)	.937	.877	.355	15.4	.000

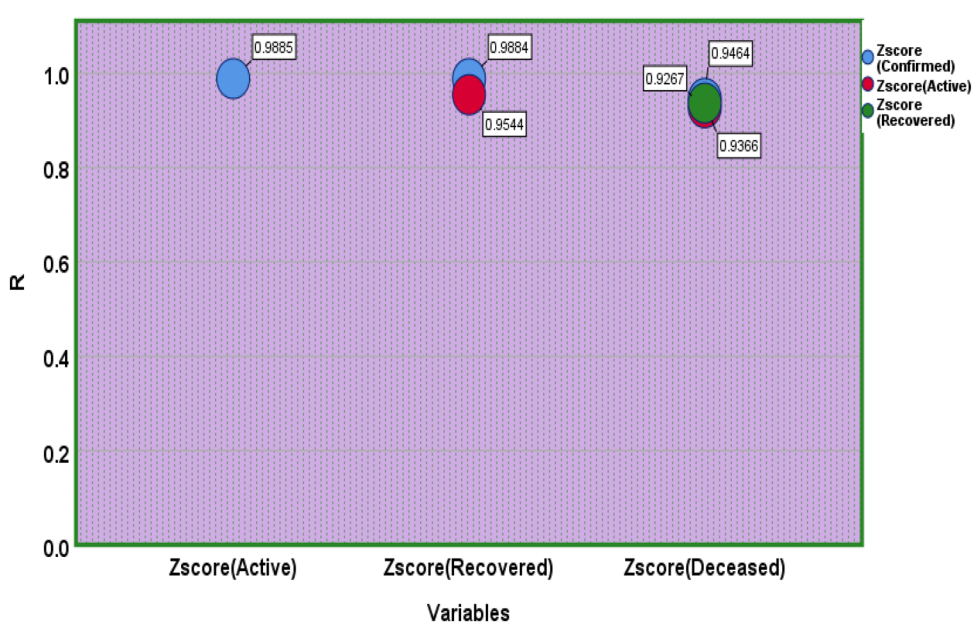

Fig. 8 Covid-19 case correlation at the 0.01 level (2-tailed)

Table 6 ANOVA

	Residual	df	F	Sig. (P)
LR-1(Active based)	4.8	33	200.7	.000
LR-2(Recovered based)	4.1	33	235.5	.000

To measure the collective impact (active and recovered cases) on the deceased, the MLR model is built. On one hand, the LR-1 model's findings signify the effect of active cases on the patient deceased, and on the other hand, the LR-2 model explains the power of recovered cases to predict the deceased.

Table 7 depicts the MLR model metrics, which validated the combined predictive strength of recovered and active cases. The residual error was reduced by 0.4. A bit increment in correlation (R) and coefficient of determination (R^2) was achieved. The autocorrelation score of 2.3 was determined by the Durbin-Watson test, which is close to 2.5 thresholds and infers acceptable autocorrelation between independent variables and adjusted $R^2 = 0.882$ in the MLR model also significant. Further, the model's F value is also found to be significant ($p < 0.005$). Therefore, considering both variable active and recovered cases, predictive strength is improved with the new value of $R^2 = 0.889$.

Three presented models played a vital role in the hypotheses testing, and t values were found significant. In the LR-1, significant t-value is rejected the null hypothesis ERH0: "No effect of recovered cases on patient deceased" and was unable to reject the alternative hypothesis ERHa: "An effect of recovered cases on patient deceased". As a result, recovered cases had a big effect on patients who passed away. Further, the LR-2 model's t-value forced to make a decision failed to accept the null hypothesis EAH0: "No effect of active cases on patient deceased" and alternative hypothesis EAHa: "An effect of active cases on patient deceased" is failed to reject. It proved that active cases have a significant impact on the deceased.

Figure 9 plots the standardized predicted values of the deceased provided with the MLR model. Equation $Y = 1.87\text{E-}16 + 0.93*X$ is significantly explained with the coefficient of determination $R^2 = 0.889$. Only a few records were observed as far away from the benchmark line. Therefore, active and recovered cases are supported to identify the deceased cases effectively. In other words, the accuracy (89%) or explanation strength (0.889) of predictors towards the target variable is most significant.

Figure 10 shows the residual error versus a prediction of the MLR model. The loss curve shows the error points, which does not prove the normal distribution of the residuals because the range does not come in between −2 and +2.

Table 7 MLR model with the impact of active and recovered cases on deceased

	Residual	R	R2	Adj. R2	Std. Error	F	Durbin Watson	Sig. (P)
MLR	3.7	.943	.889	.882	.343	128.4	2.3	.000

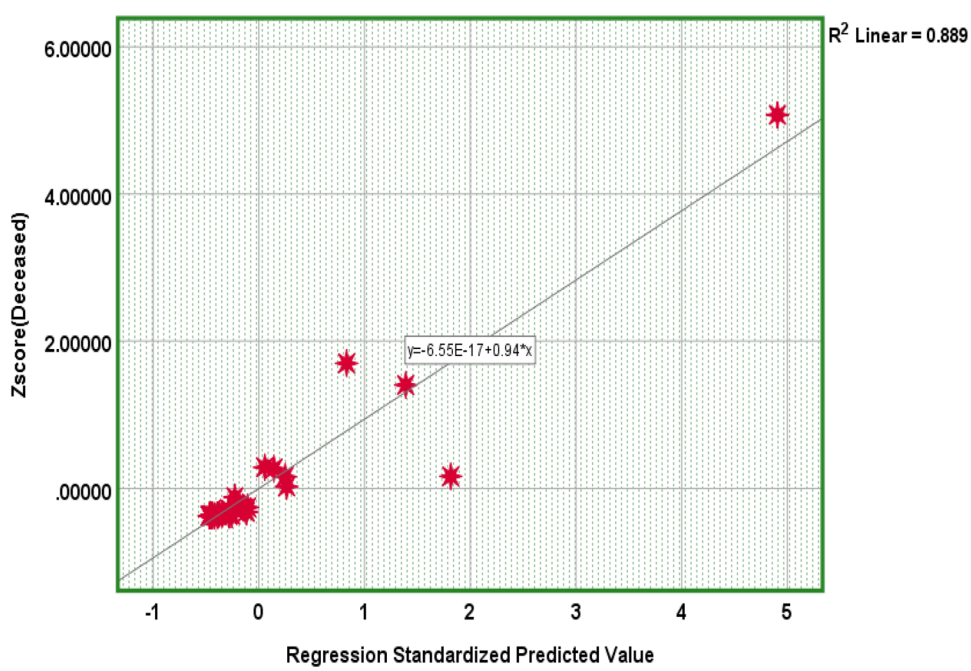

Fig. 9 Standardized deceased predicted value

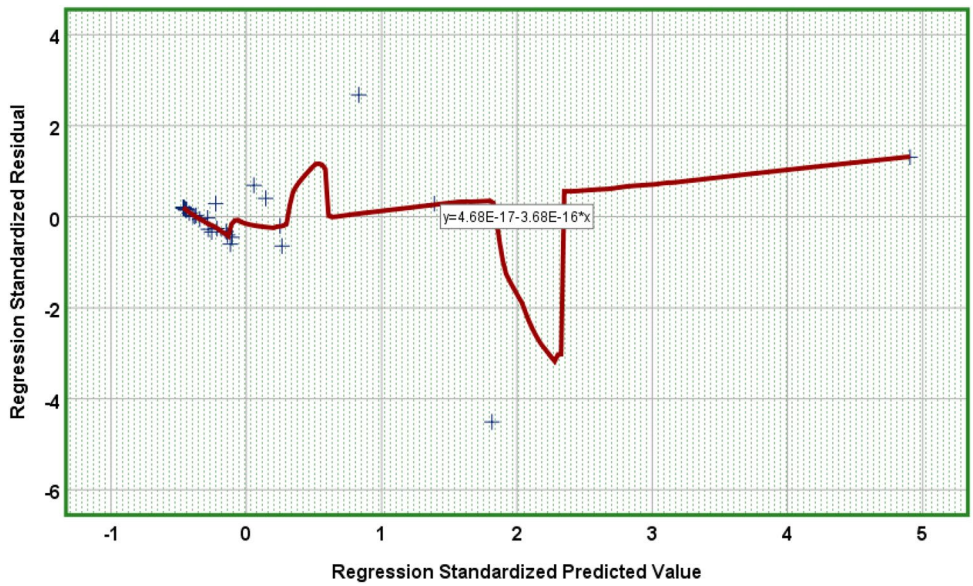

Fig. 10 Standardized deceased predicted value

Hence, the residual is not randomly scattered around zero and linearity achieved. Still, the MLR model is better as compared to the LR-1 and LR-2 of the provinces.

Further, the LR-2 model (recovered based) depicted that the MH, Tamilnadu, and DH are the highest deceased-prone provinces. According to this model, less than 100 deceased may have chances in Assam, Kerala, Uttarakhand, and Jammu Kashmir. The MLR model, estimates the deceased based on recovered and active cases. For the MH, it predicted 3411; for Tamilnadu, it estimated 1422; for Delhi, it observed 1148; for Gujrat, it identified 788 deceased cases.

Figure 11 shows the combined predicted values given by the respective regression models. Starting from the actual deceased count in 19 provinces in India reported up to 13th June. The LR-1 model predicted the deceased cases based on active cases. The highest observed deceased affected states are MH, GUJ, and Delhi. According to the LR-1 model, the possibility of more than 3300 deceased in highly red zone state MH and 1504 life of humans lost in Delhi. More than 300 deceased were reported in Gujrat and WB. Tamilnadu state needs to take care too caused of 1235 deceased. Less than 100 deceased cases predicted in the rest of cases.

Figure 12 visualizes the ARIMA Model's output towards the next 40-days forecasting in India. On the graph, the blue lines signify the forecasting line, red lines show the observed cases, and dotted pink lines denote the UCL and LCL, two control limits: an upper and a lower limit of forecasting. The model predicted the total 753,216 confirmed cases with a lower bound of 704,460 and an upper bound of 801,973 cases. They were forecasting the total number of active cases in the country 331,580 with UCL of 348,581 and LCL of 314,580.

Further, the model proved the forthcoming human losses might be encountered 22,411 with UCL of 24,124 and LCL of 20,699. Recovery of infected cases is predicted at 399,225 with UCL of 431,235 and LCL of 367,214. Therefore, based on the observed values, the forecasting graph proved to enhance all four aspects of Covid-19 affected Indians in the next 40- days.

Figure 13 shows the predicted count measured with the ARIMA Model for the next 40 days. The highest number of confirmed cases to be reported was indicated by vertical blue bars, which is 753,216. Among them, the highest active cases are supposed to be 331,580, and the possibility of 399,225 recovered humans until 23rd July. In the next 40 days, the deceased count may arrive up to 22,411. Therefore, we observed that all cases are rising rapidly.

Figure 14 compares the deceased predictive strength provided by regression and time series forecast methods. Accordingly, the ARIMA model, India maybe lost the highest number of human lives around 22,411, caused by Covid-19. This forecast value is calculated for 23rd July 2020. The lower number of deceased, 8854 predicted with LR-1 and MLR model identified 9429 deceased not confined to any date.

Table 8 depicts the vital performance measures to prove the strength of forecasting models. All models' goodness of fit (the coefficient of determination) was found significant. The Mean Absolute Percentage Error (MAPE) Eq. (17) of active cases is calculated as very low compared to others. The minimum normalized Bayesian Information Criterion (BIC) is 9.8, calculated for the deceased. According to Eq. (18). All four forecast models are found significant ($P < 0.005$) due to the computed t-value with Eq. (19).

$$\mathrm{MAPE} = \left\{ \frac{\left[\sum_{t-1}^{n} \left| yt - \hat{y}t \div yt \right|\right]}{n} \right\} \times 100\% \tag{17}$$

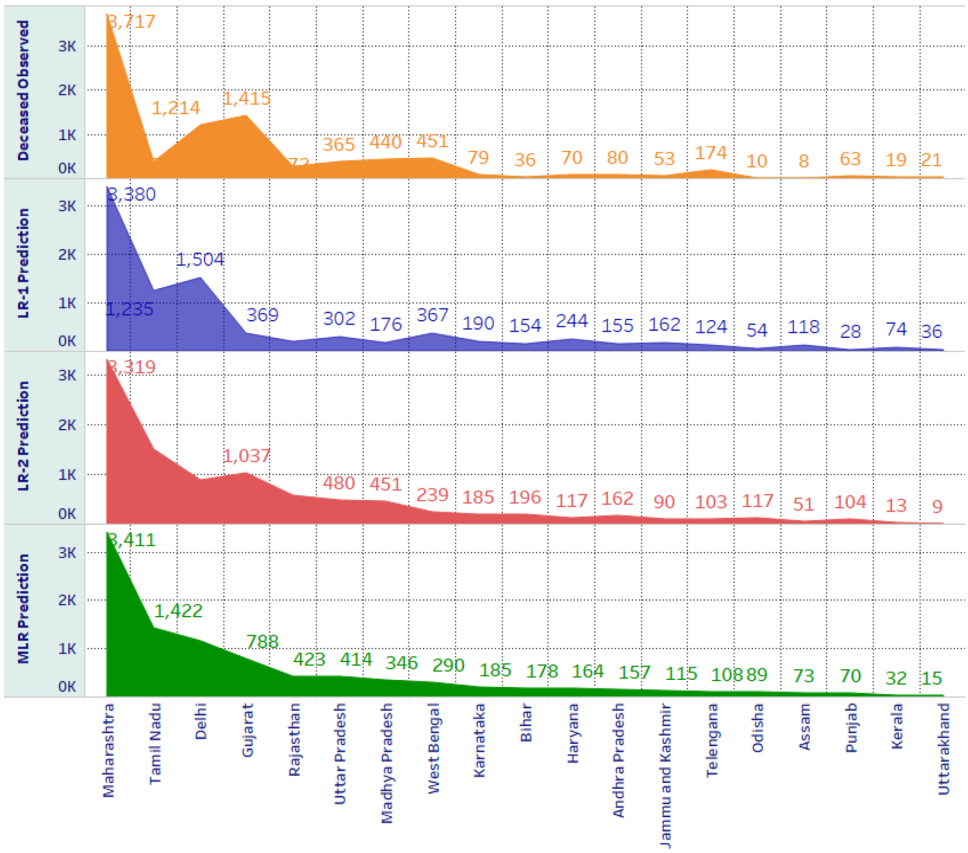

Fig. 11 Regression model's comparison with real deceased predicted values

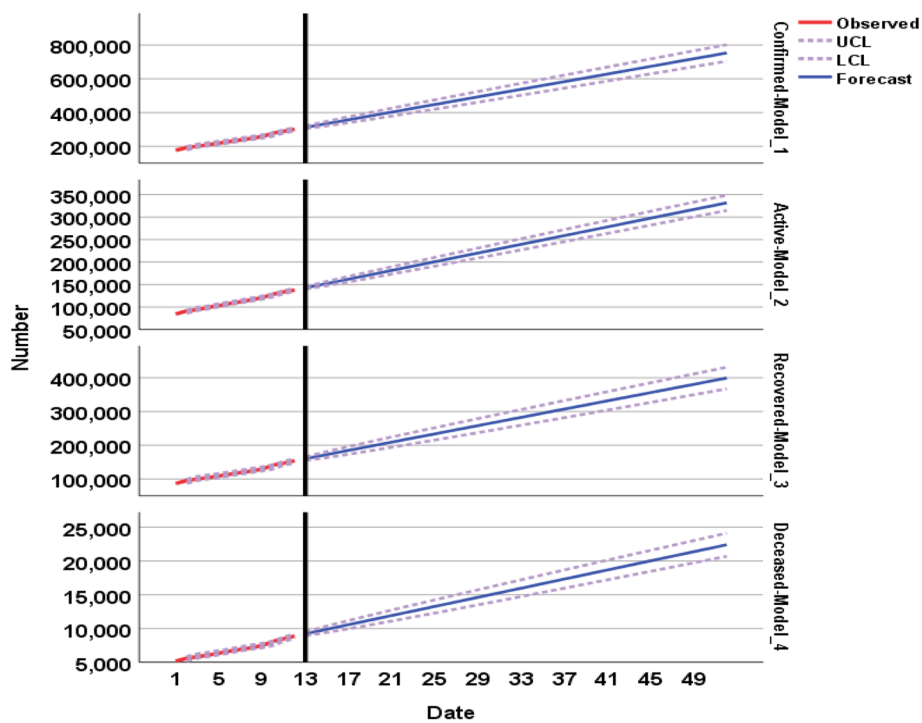

Fig. 12 Forecasting with ARIMA

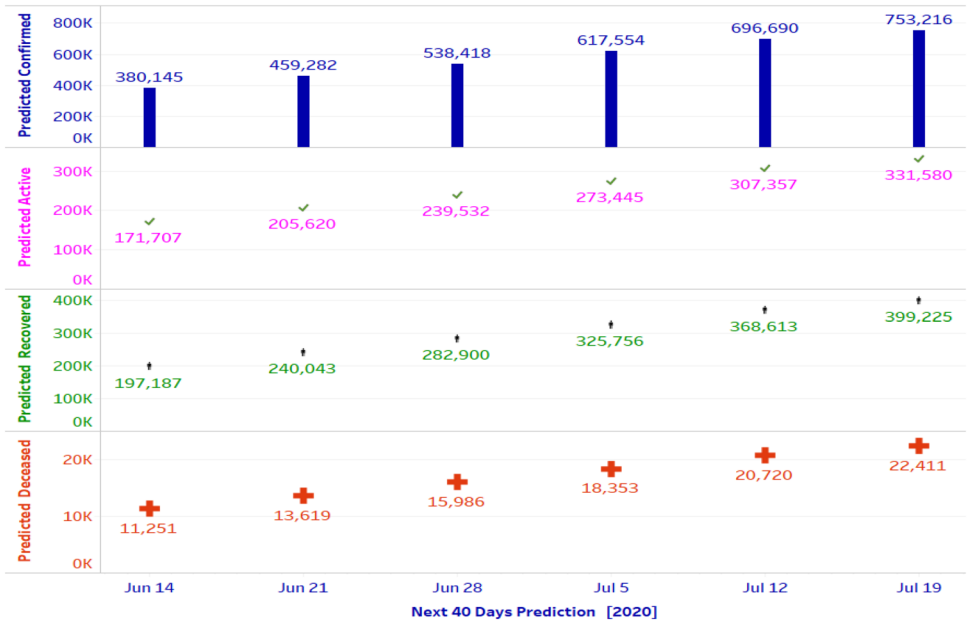

Fig. 13 ARIMA Model's prediction of all cases for next 40 days up to 23rd July 2020

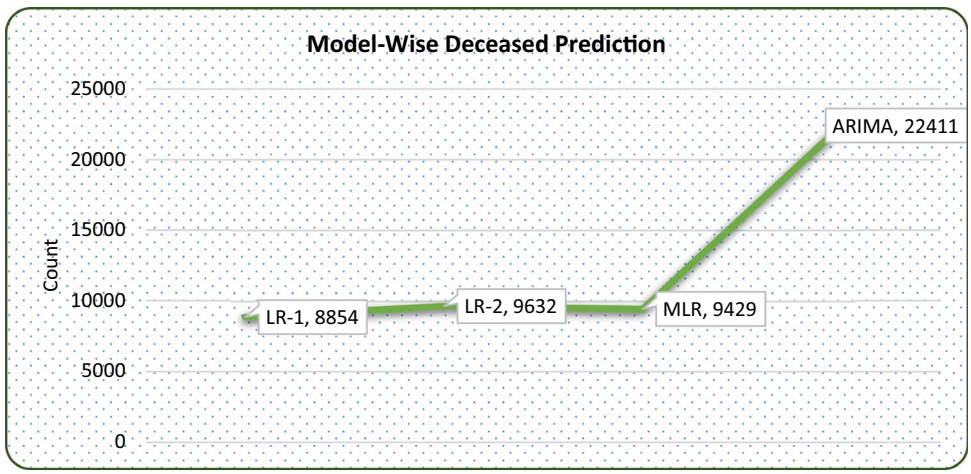

Fig. 14 Total deceased prediction in India suggested by regression and forecasting

Table 8 ARIMA performance (95% confidence interval)

	R2	MAPE	Normalized BIC	t	Sig. (P)
Confirmed	.991	1.06	16.5	10.8	.000
Active	.994	.78	14.4	13.3	.000
Recovered	.986	1.39	15.6	8.9	.000
Deceased	.988	1.29	9.8	9.2	.000

$$BIC = k\ \log(n) - 2\log\left(L\left(\hat{\theta}\right)\right). \tag{18}$$

$$t = x - \widehat{\frac{\mu}{s}} / \sqrt{N} \tag{19}$$

Figure 15 displays residuals of predicted values for four cases, including ACF and PACF. It shows UCL and LCL, which create the residual for various lags. For confirm case prediction, 6, 7, and 8 number lag shows significant autocorrelation. Also, this order was approved with the corresponding PACF. The most considerable lag is 7 and 8 for the forecasted active case. For the recovered case, lag 6, 7, and 8 are significant. In deceased forecasting, lag 1, 6, and 7 show the highest autocorrelation in the series.

6 Conclusion

This study conducted two significant experiments demonstrating regression and time-series forecasting with respect to Covid-19. To estimate the number of Covid-19-infected

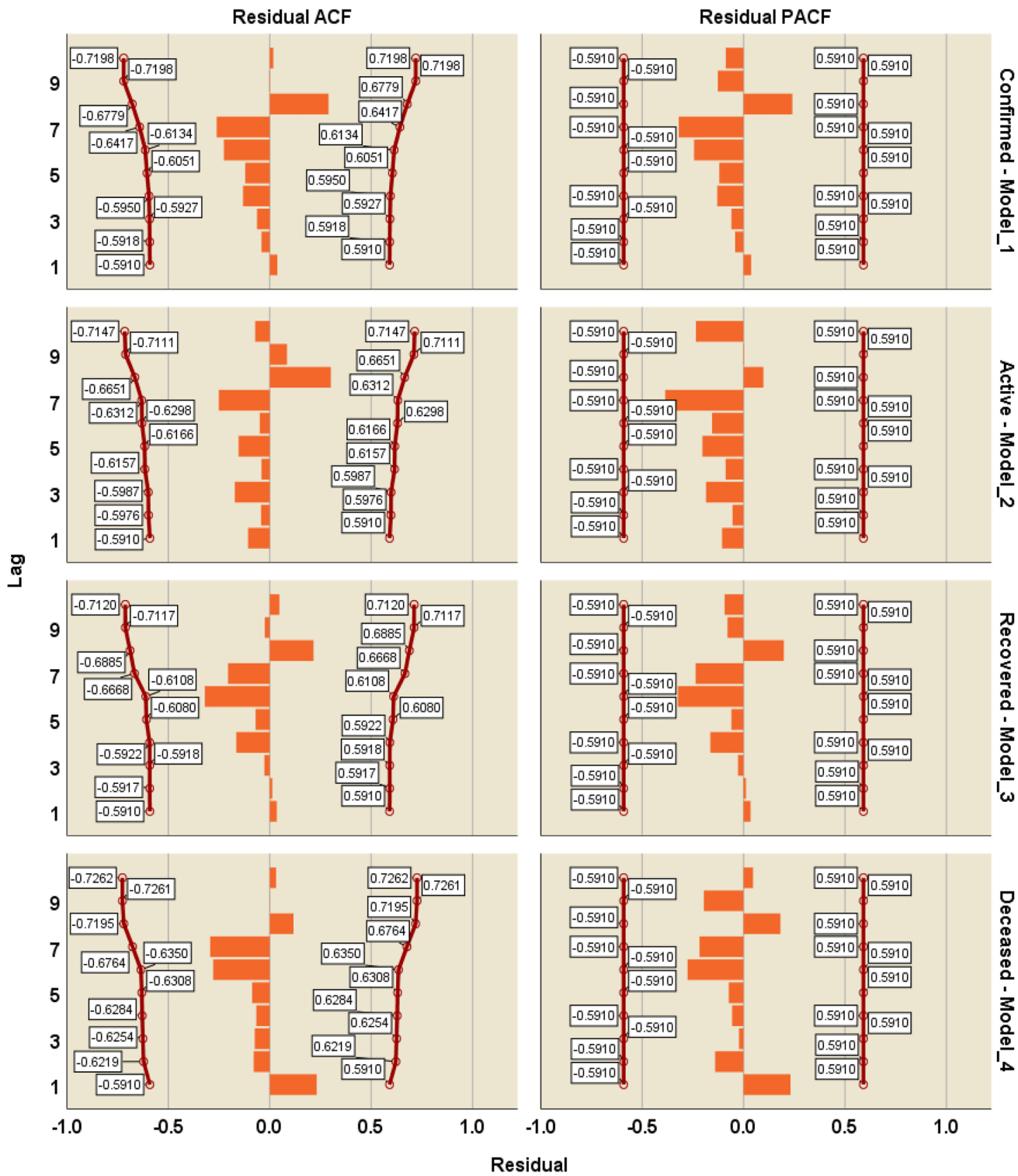

Fig. 15 Residual of ARIMA forecasting

future human deaths in India, we presented four predictive models. The study's findings looked at the significant correlation between the rate of deceased patients and their recovery and active status. Active cases had an impact on the deceased rate on the one hand, and the recovered patient had an impact on the deceased patient on the other. The ARIMA forecasts the highest deceased in the country. Based on results found with MLR, most deceased may be reported in

four provinces (MH, DH, GUJ, Tamilnadu). Overall recovery must be achieved around 400,000, and around 300,000 humans remain active that period. Therefore, profoundly deceased prospects are seen in both cases, even engaged or recovered. Surplus suggestions for recovered patience need to look after more until the pandemic goes down. Additionally, red zone states were warned to take precautions against the epidemic by current models. Observing the government's Covid-19 guidelines and the prohibitions on anticipatory treatment, need to be applied to reduce active and deceased cases, which the model predicted.

Regression models investigated essential elements of the deceased. Based on active and recovered patients, multiple linear regression produced a substantial R^2 of 0.89, which predicted that the patient would pass away. Additionally, the timing of Covid-19 was examined using an ARIMA analysis that predicted confirmed, active, recovered, and deceased cases for the following 40 days.

The future study includes applying a Deep Neural network with appropriate optimization methods. The base samples should be consisted of at least one month. The severe future reporting about the huge count of decease and active cases, and forecasting needs to be estimated for the next six months. Additionally, other public datasets can be evaluated and compared with our results.

7 Limitations

The present study is limited to a specific fixed number of hypotheses. The training samples were used only for 13 days. The days for the forecasting are limited up to 40 days. The particular ARIMA model was applied for forecasting purposes with a random walk. Further, we explored the impact of active and recovered cases on deceased cases. Only a state-wise decease was forecasted instead of districts. We have compared regression and ARIMA approaches on the secondary dataset and found that the ARIMA model is more accurate and worth deploying using Flask technology.

Funding The authors did not receive support from any organization for the submitted work.

Declarations

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

Informed consent The studies are conducted on already available data for which consent not required.

Human or animal participants This is an observational study. This research includes No involvement of Human and Animals, so no ethical approval is required.

References

Almendros Jimenez JM, Becerra Teron A, Torres M (2021) The retrieval of social network data for points of interest in openstreetmap. Human Centr Comput Inform Sci. https://doi.org/10.22967/HCIS.2021.11.010

Boudry L, Essahib W, Mateizel I, Van de Velde H, De Geyter D, Piérard D, De Brucker M (2022) Undetectable viral RNA in follicular fluid, cumulus cells, and endometrial tissue samples in SARS-CoV-2–positive women. Fertil Steril 117(4):771–780. https://doi.org/10.1016/j.fertnstert.2021.12.032

George EP, Box GM, Jenkins GC, Reinsel Ljung GM (2015) Time series analysis: Forecasting and Control - George E. P. Box, Gwilym M. Jenkins." Book.

Bungaro M, Passiglia F, Scagliotti GV (2022) COVID-19 and lung cancer: a comprehensive overview from outbreak to recovery. Biomedicines. https://doi.org/10.3390/biomedicines10040776

Chintalapudi N, Battineni G, Amenta F (2020) COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: a data driven model approach. J Microbiol Immunol Infect Wei Mian Yu Gan Ran Za Zhi 53(3):396–403. https://doi.org/10.1016/J.JMII.2020.04.004

Chitra N, Shanmathi R, Rajesh R (2015) Application of arima model using spss software-a case study in supply chain management. Case Study

Gao Y, Zhang Z, Yao W, Ying Qi, Long C, Xinmiao Fu (2020) Forecasting the cumulative number of COVID-19 deaths in China: a boltzmann function-based modeling study. Infect Control Hosp Epidemiol 41(7):1. https://doi.org/10.1017/ICE.2020.101

Garcia-Flores V, Romero R, Xu Y, Theis KR, Arenas-Hernandez M, Miller D, Gomez-Lopez N (2022) Maternal-fetal immune responses in pregnant women infected with SARS-CoV-2. Nat Commun. https://doi.org/10.1038/s41467-021-27745-z

Hao F, Park DS (2021) CoNavigator: a framework of FCA-based novel coronavirus COVID-19 domain knowledge navigation https://doi.org/10.22967/HCIS.2021.11.006

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Cao B (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan China. The Lancet 395(10223):497–506

Izquierdo-Pujol J, Moron-Lopez S, Dalmau J, Gonzalez-Aumatell A, Carreras-Abad C, Mendez M, Martinez-Picado J (2022) Post COVID-19 condition in children and adolescents: an emerging problem. Front Pediatr. https://doi.org/10.3389/fped.2022.894204

Kotlyar AM, Grechukhina O, Chen A et al (2021) Vertical transmission of coronavirus disease 2019: a systematic review and meta-analysis. Am J Obstet Gynecol 224:35–53

Kumar S, Viral R, Deep V, Sharma P, Kumar M, Mahmud M, Stephan T (2023) Forecasting major impacts of COVID-19 pandemic on country-driven sectors: challenges, lessons, and future roadmap. Pers Ubiquit Comput 27(3):807–830. https://doi.org/10.1007/s00779-021-01530-7

Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, Wang D, Chen G, Zhang J, Peng H, Shao Y (2020) Propagation analysis and prediction of the COVID-19. Infect Dis Model 5:282–292. https://doi.org/10.1016/J.IDM.2020.03.002

Male V (2022) SARS-CoV-2 infection and COVID-19 vaccination in pregnancy. Nat Rev Immunol 22(5):277–282. https://doi.org/10.1038/s41577-022-00703-6

Mohler RR (1990) Nonlinear time series and control applications. In: Proceedings of the IEEE conference on decision and control 2917–19

Rishabh S, Johri A, Deep V, Sharma P (2019) Heart diseases prediction system using CHC-TSS evolutionary, KNN, and decision tree classification algorithm. Adv Intell Syst Comput 813:809–819. https://doi.org/10.1007/978-981-13-1498-8_71

Rizzo G, Mappa I, Pietrolucci ME, Lu JLA, Makatsarya A, D'Antonio F (2022) Effect of SARS-CoV-2 infection on fetal umbilical vein flow and cardiac function: a prospective study. J Perinat Med 50(4):398–403. https://doi.org/10.1515/jpm-2021-0657

Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, Yan P, Chowell G (2020) Real-time forecasts of the COVID-19 epidemic in China from february 5th to february 24th, 2020. Infect Dis Model 5:256–263. https://doi.org/10.1016/J.IDM.2020.02.002

Rujen J, Sharma P, Keshri R, Sharma P (2023) COVID detection using cough soundhttps://doi.org/10.1007/978-981-19-7346-8_69

Sharma P, Saxena K, Sharma R (2016) Heart disease prediction system evaluation using C4.5 rules and partial tree. Adv Intell Syst Comput. https://doi.org/10.1007/978-81-322-2731-1_26

Sharma P, Alshehri M, Sharma R, Alfarraj O (2021) Self-management of low back pain using neural network. Comput Mater Continua 66(1):885–901. https://doi.org/10.32604/CMC.2020.012251

Sharma P, Alshehri M, Sharma R (2022a) Activities tracking by smartphone and smartwatch biometric sensors using fuzzy set theory. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-022-13290-4

Sharma P, Sharma P, Shukla VK (2022b) Covid-19 detection using cough sound with neural networks in 10th international conference on reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2022b, https://doi.org/10.1109/ICRITO56286.2022.9965099.

Shivam B, Alowaidi M, Bhardwaj R, Sharma SK (2021) Machine learned hybrid Gaussian analysis of COVID-19 pandemic in India. Results Phys 30:104630. https://doi.org/10.1016/J.RINP.2021.104630

Sophia S, Vanessa K, Tobias P et al (2022) Effects of SARS-CoV-2 on prenatal lung growth assessed by fetal MRI. Lancet Respir Med. https://doi.org/10.1016/S2213-2600(22)00060-1

Sun H, Koch M (2001) Case study: analysis and forecasting of salinity in Apalachicola bay, Florida, using box-Jenkins ARIMA models. J Hydraulic Eng 127(9):718–727

Tatura SNN (2022) Case report: Severe COVID-19 with late-onset sepsis-like illness in a neonate. Am J Trop Med Hyg 106(4):1098–1103. https://doi.org/10.4269/ajtmh.21-0743

Tomar A, Gupta N (2020) Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ 728:138762. https://doi.org/10.1016/J.SCITOTENV.2020.138762

Tran TT, Pham LT, Ngo QX (2020) Forecasting epidemic spread of SARS-CoV-2 using ARIMA model (case study: Iran). Glob J Environ Sci Manag 6(Special Issue (Covid-19)) https://doi.org/10.22034/GJESM.2019.06.SI.01

Wang L, Li J, Guo S, Xie N, Yao L, Cao Y, Day SW, Howard SC, Carolyn Graff J, Tianshu G, Ji J, Weikuan G, Sun D (2020) Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. Sci Total Environ 727:138394. https://doi.org/10.1016/J.SCITOTENV.2020.138394

Yang Y, Hao F, Park D-S, Peng S, Lee H, Mao M (2021) Modelling prevention and control strategies for COVID-19 propagation with patient contact networks https://doi.org/10.22967/HCIS.2021.11.045

Zhang Z, Jing J, Wang X, Choo KKR, Gupta BB (2020) A crowdsourcing method for online social networks security assessment based on human-centric computing. Human-Centr Comput Inform Sci. https://doi.org/10.1186/s13673-020-00230-0