2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS) | 979-8-3503-4616-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACCESS57397.2023.10199278

Analysis and Forecasting of COVID-19 Pandemic Using ARIMA Model

Soni Singh
Department of Computer Science and Engineering,
School of Computer Science and Engineering, Lovely Professional University, Jalandhar-Delhi G.T. Road, Phagwara 144411,
Punjab, India
soni.30409@lpu.co.in

Sonam Mittal
Chitkara University Institute of Engineering & Technology
Chitkara University,
Punjab, India
sonam.mittal@chitkara.edu.in

Sunaina Singh
Department of Electrical & Electronics School of Engineering, University of Petroleum and Energy Studies (UPES))
Bidholi, Dehradun, India.
nainasingh0306@gmail.com

***Abstract*-The global community is now seriously threatened by the COVID-19 pandemic. The government of every nation must pay close attention to the analysis of this disease to take the required actions to lessen the impact of this worldwide epidemic. This research focused on the disease outbreak in the Indian region through July 21st, 2021, and evaluated the incidence and mortality. Machine learning techniques, such as the ARIMA model, are applied to perform the prediction analysis on collected data from the World Health Organization (WHO) official portal for India between January 20, 2020, and July 21, 2021. Mean Square Error (MSE), a measure of model performance, was used to assess performance, and it came in between 2170.636098 and 46.839689. In the four weeks of test data, the Expected instances are estimated to be between 192K and 230K, which is fairly similar to the actual figures. The government and physicians will be able to make future strategies with the aid of this study**.

Keywords: *Machine Learning, Analysis, ARIMA Model, COVID-19, Prediction.*

I. INTRODUCTION

The novel coronavirus (CoV) is a novel variant of the coronavirus. In China, Wuhan city, the first infected person with coronavirus, was found, later infection was termed COVID-19. Each of the three diseases—corona, virus, and disease—is portrayed. This virus was once referred to as the "2019 novel coronavirus" or "2019-nCoV." The COVID-19 new virus belongs to the same virus family as SARS and numerous cold virus subtypes [1]. Direct contact with an infected person having a cough and cold, and the respiratory droplets due to coughing and sneezing of the infected person may infect the healthy person, as well as the surfaces that have been exposed to the virus may also cause infection. Simple disinfectants can eliminate the COVID-19 virus, even though it may survive on surfaces for several hours. Since, December 2019, the coronavirus has infected millions of individuals all over the world [1,2]. The coronavirus has a competency for mutation and is exceedingly contagious. People who suffered from coronavirus experience severe problems related to respiration and lungs, and they are more likely to become ill if they have chronic illnesses [2] like diabetes or cardiovascular disease, a weakened immune system, or they are older [14]. On March 11, 2020, the WHO declared the COVID-19 disease as a pandemic. Since an infected person may show symptoms years after catching the illness or may not, the condition is challenging to control.

For COVID-19, the vaccine has been developed. In this instance, social exclusion and using testing to identify the positive cases. Studies that demonstrate the use of statistical analysis, modelling, and artificial intelligence to slow the spread of the virus and highlight its effects in the days to come have been published [4]. These preliminary analyses were conducted utilising the scant information that was available at the beginning of the outbreak.

Now that the infection has spread widely, there is a wealth of information that can be analysed. To help governments and health services plan for and control the development of infectious diseases, predictive analysis of COVID-19 has emerged as a hot research topic [3]. The health systems can be prepared to handle the approaching quantity of patients by modelling and anticipating the virus's daily spread behaviour. It is important to accurately predict the disease since it could have an impact on government policy, containment measures, the health system, and social life. We investigate the forecasting model ARIMA's [6] propensity to predict in this setting. Due to their greater predicting accuracy, the models are generally recognised and used. For our analytic study, we employ the day-level cumulative COVID-19 India instances. The confirmed coronavirus cases, deaths, and cures in India from Jan 2021 to Aug 2021 are depicted in Fig. 1 below. As we can see from the graph, the number of COVID-19 instances significantly increased from May 2021 to July 2021.

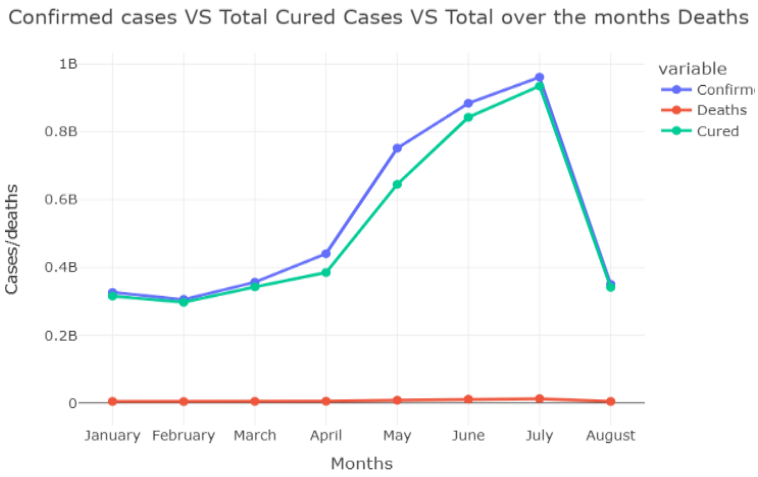

Fig. 1 COVID-19 Total Cases in India

The goal of this work is to give an evaluation of a prediction model utilising COVID-19 India cases and to protect against the

979-8-3503-4616-9/23/$31.00 ©2023 IEEE

virus's effects in India. We analyse trends in COVID-19 instances and demonstrate how well the model performs using by measuring the performance of the model using mean square error (MSE) [6]. For confirmed, ongoing, recovered, and fatal COVID-19 cases, we produce forecasting findings. The findings demonstrate that for the totality of COVID-19 cases, the ARIMA model performs better. The remaining sections of the paper are organised as follows. A review of the literature is included in Section II. Using COVID-19 data, Section III analyses trends. The time series forecasting models are described in Section IV. Section V of the result describes the analysis of actual and predicted cases for the COVID-19 dataset. The Section VI presentation includes statistical analysis and model evolution. The paper is wrapped up in Section VII.

II. Literature Review

In the paper [4], Analysis, the author measures the effect of Covid-19 in various states of India and makes predictions using Machine Learning (ML) models. The WHO datasets have been used for the analysis of Covid-19. The paper shows that linear regression models have been utilized for prediction.

In paper [5], (KNN) the author used the Covid-19 datasets for the prediction of infection rate. KNN- classifier was used to measure the degree of closeness and strength of each neighbour for classification. The study demonstrates that the KNN is more accurate in predicting the infectious rate in COVID-19 patients.

The authors of the paper [6] suggested the SCIR model using ML models and regression models, for making predictions using Covid-19 data sets. The data was collected from the authenticated website of the Indian government. The performance of the proposed model was evaluated using the RMSE score and Error rate of SCIR and the Regression Model. It has been observed that the SCIR model outperforms better in comparison to the regression model.

The author [7], used a different regression analysis model for data analysis of Covid-19. The study utilized a regression model based on 3rd, 4th, 5th, and 6th degree and exponential polynomials. The results show that the sixth-degree polynomial regression model is performing better in comparison to other models.

In the paper [8], the authors consider trend analysis models that are logistic, exponential and susceptible-infectious susceptible (SIS) models for predictions. The author used to measure two weeks of Covid-19 trend. It has been classified that the R-square value of logistic and exponential models of ML is above 0.90. They have also created a web-based application to measure the daily base updates of Covid-19 cases.

In the paper [9], the author has done Covid-19 analysis globally for the last 2 years. The phenomenon, like the COVID-19 pandemic, has spread across the world. More than 418.6 million confirmed cases and 5.8 million death cases have occurred in the past 2 years. The study also shows that 55.04 % of people are fully vaccinated for approaching herd immunity globally. The study concluded that the covid-19 infection can only be controlled through the proper imposing of lockdowns and vaccination of people.

In paper [10], the author has done a comprehensive study in India, for predicting the infection rate of COVID-19 cases, forecasting models. The autocorrelation and autoregressive models have been utilized to enhance the performance of forecasting models using accuracy matric, along with multiple linear regression and correlation coefficients. The actual number of cases and predicted values are agreed closely i.e., 0.9992- R square score. The author concluded that the complete isolation of patients and proper imposition of lockdown are two major elements that can lead to slowing the spreading of COVID-19, disease.

III. Materials and Methods

The patients of confirmed, and death cases of COVID-19, including the cured or recovered people from COVID-19, are predicted using the ARIMA model. According to the rate of Covid-19 infection, the number of cases has increased. In India, cases can be controlled by enforcing a lockdown. For the above-mentioned prediction, data analysis and visualisation are carried out in the research. The following points in Fig. 2 are used to discuss the methods used in the prediction of COVID-19 cases:

1. Data Collection
2. Data Analysis
3. ARIMA Model

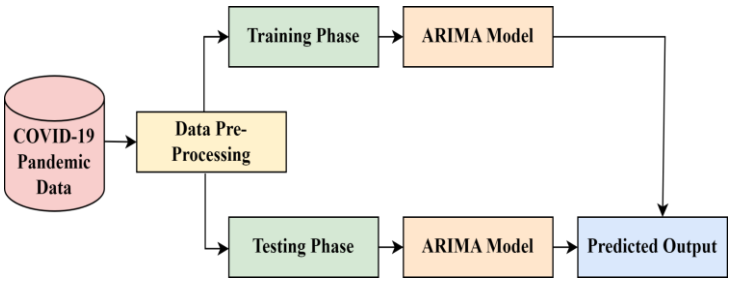

Fig. 2 Work-flow of ARIMA Model

For making predictions of COVID-19 instances in India, a prediction model is applied. Prediction is made using the ARIMA model. The data used for prediction includes a variety of attributable information, including confirmed, death cases of coronavirus and recovered patients from COVID-19 cases in India.

A. Data Collection

The COVID-19 dataset was taken from the WHO website [2]. The dataset used for prediction analysis is collected for the time of 18 months from Jan 2020 to Jul 2021. The data includes the total number of confirmed occurrences. There are live cases, dead cases, and cured cases in India. Datasets of the.csv type can be found in the file. By deleting the empty values, the data is updated. In this instance, the dataset is divided into two different sets of training and testing samples. 75 % dataset is fed to the model to train it, while 25 % dataset is for testing the model.

B. Data Analysis

In India, the first case of COVID-19 was reported on date Jan 30, 2020. In Feb, there were three instances reported, and this number remained constant the entire month. The disease's rate of

spread greatly accelerated in March 2020. The variety in reported instances and fatalities from January 22, 2020, to July 20, 2021, is shown in the statistics [2,4,6]. In March, the disease spread widely in India at a faster rate. Taken, the dataset is examined to check the fitness of the model for COVID-19 prediction analysis, before deploying it. Figure 3 displays state wise analysis of COVID-19 confirmed cases in India. As we can see through the figure that it shows that few of the state has very higher confirmed individual of COVID-19 The government of India impose proper lockdown and vaccination of people. It also indicates that many states of India reached their peak stage to manage the pandemic [7].

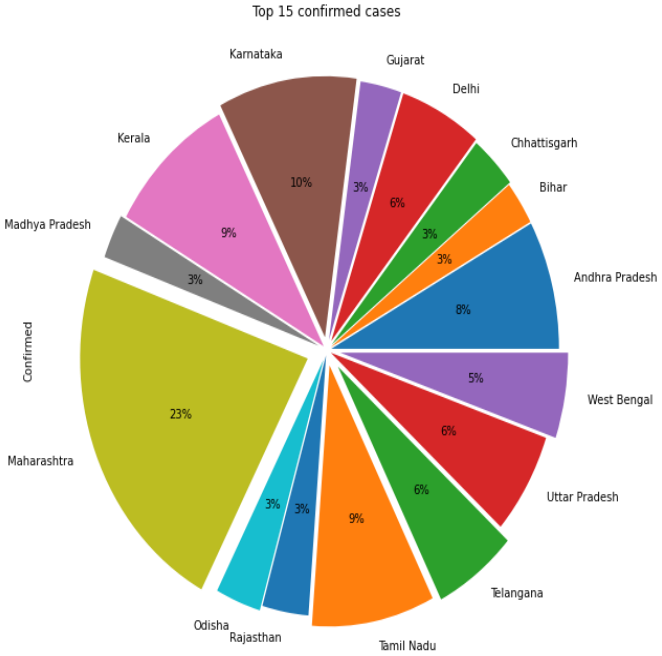

Fig. 3 State-wise Confirmed Cases in India

Figure 4 below displays the top 5 Indian states by the number of active cases. The graph indicates that Kerala, Tamil Nadu, Karnataka, Maharashtra, and Andhra Pradesh had the largest number of COVID-19 active cases in India from Jan 2020 to Jul 2021. In comparison to other states, these states are highly affected states [8,10].

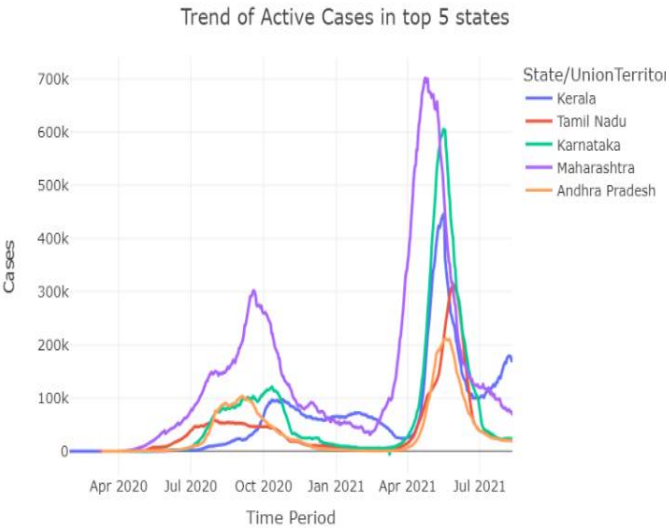

Fig. 4 Trend of COVID-19 Active Cases in Top 5 States

C. *ARIMA Model*

ARIMA model stands for Autoregressive Integrated Moving Average Model (ARIMA) model. It is a combination of Auto Regression (AR), Integrated (I) and Moving Average. The model uses the notation of integration with time series data [9].

The aspects of the model are defined as-

- AR: The AR model is used to work for the dependent relationship between observations and other backward observations.
- I: I stand for integration. It mainly uses for the difference of row observation to subtract observation from previous timestamp observation to do time series data.
- MA: MA is used to model the dependence between observation and residual error in applied lagged observations.

All the aspects used in the ARIMA model are taken as a parameter. ARIMA used a standard notation ARIMA (p, q, d) and these parameters were used as integer values to identify which model was being used for parameter substitution [10].
Each parameter of ARIMA is defined below-

- p: It is used as a lag observation which is also known as lag order.
- q: This is referred to as the degree of difference for the number of raw observation differences.
- d: This is referred to as a number of moving averages for the size of the moving average window.

In terms of k, the general forecasting equation for the ARIMA model is given in Eq(1):

$$k_t = \mu+\phi_1 k_{t-1} +\ldots+ \phi_p k_{t-p} - \theta_1 c_{t-1} -\ldots- \theta_q c_{t-q} \quad (1)$$

k denotes the difference, θ is the moving average parameters and ϕ_1 is the slope coefficient.

IV EXPERIMENTAL RESULTS

The forecasting model, ARIMA is used to deal with time series data. The forecasting result shows the prediction of COVID-19 cases, cured and mortality rate. The model employs the integration notation [11,12]. In the below figure prediction of confirmed, Deaths, and cured patients over the following 30 days is shown.

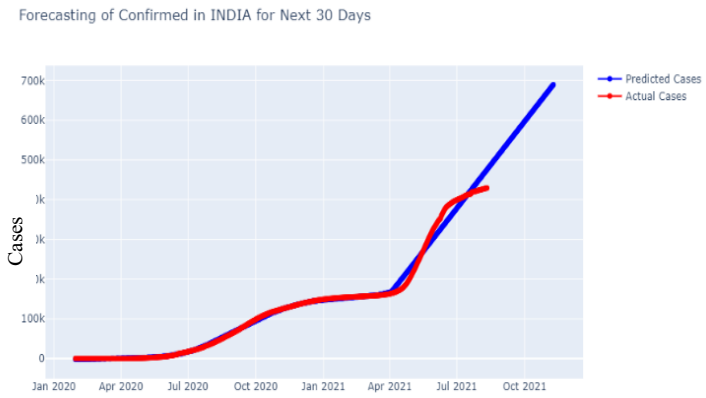

Fig. 5 Prediction using ARIMA model for Confirmed Cases

The predicted outcome of confirmed cases of COVID-19 in India is displayed in Fig. 5. The result defines an increase in cases between March and July. Based on the trend line [13], it is observed that in the upcoming month, the number of confirmed cases will significantly increase.

The predicted of cured cases in India for COVID-19 are displayed in Fig. 6. The prediction indicates a rise in the number of cured cases from March to July [11,18]. The trend line indicates that there will be a noticeably increased number of cases that are cured in the coming month. The increase in cured cases totally depends upon the proper imposition of lockdown and vaccination.

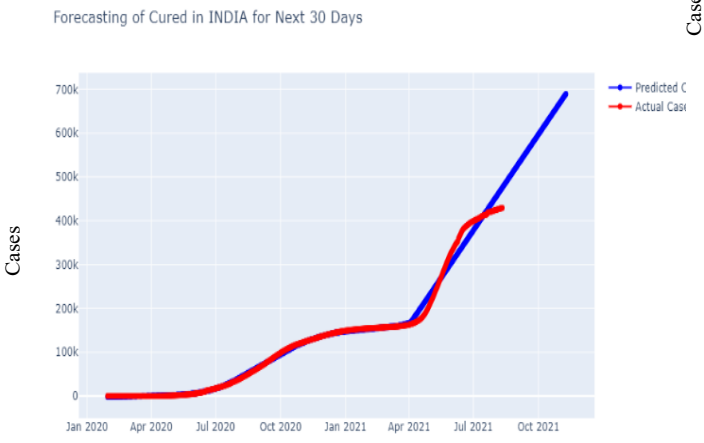

Fig. 6 of Cured Cases Prediction using ARIMA Model

The forecasting result of death cases in India are displayed in Fig. 7. The graph shows a rise in death cases between March and July. The tendency indicates that there will be a consequently higher mortality rate in the coming next month.

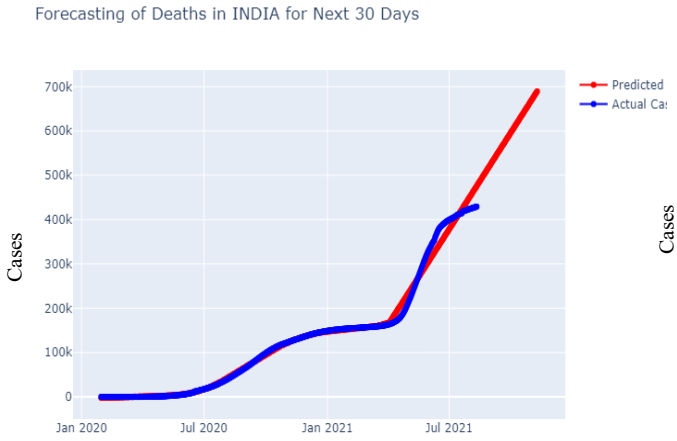

Fig. 7 Deaths Cases Prediction using ARIMA Model

V RESULTS ANALYSIS

The analysis using the ARIMA model is displayed in Fig. 8 below. The graph depicts the trend in the ARIMA model using data from India [13, 17]. The orange line represents the trend growth in the exposed population with respect to y for confirmed cases, the blue line represents the trend forecasting result of growth in the confirmed population, and the grey line represents the trend growth in the confidence interval of the confirmed cases of the given population.

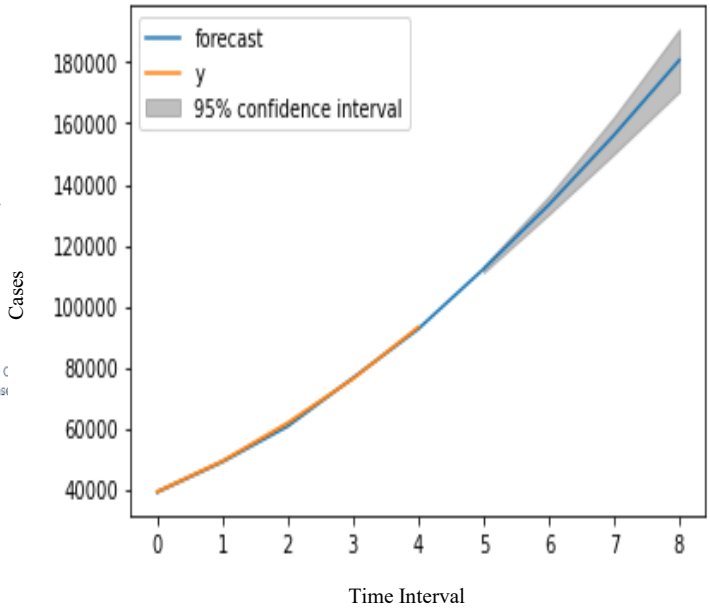

Fig. 8 Analysis of COVID-19 Confirmed Cases

The analysis result using the ARIMA model is depicted in Fig. 9, where the confirmed cases in India were represented by a red line for the actual values and the green line represents confirmed cases from predicted values for COVID-19 data [14,15]. The outcome shows a mean squared error with 2170.636098 confirmed cases in India.

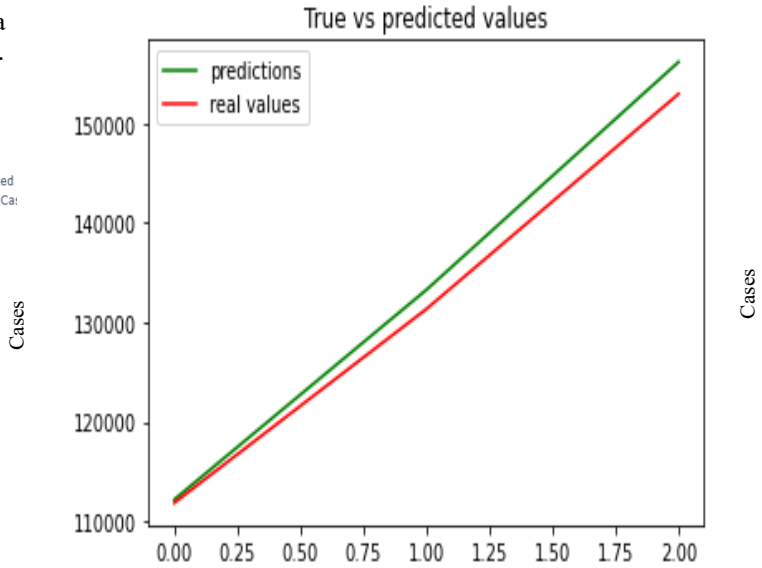

Fig. 9 Confirmed Cases for Actual and Predicted Values

Fig. 10, illustrates the analysis using the ARIMA model. The blue line in the figure represents deaths from the classification model, the orange line represents cumulative deaths cases with respect to y, and the grey line represents deaths at a specific specified time interval.

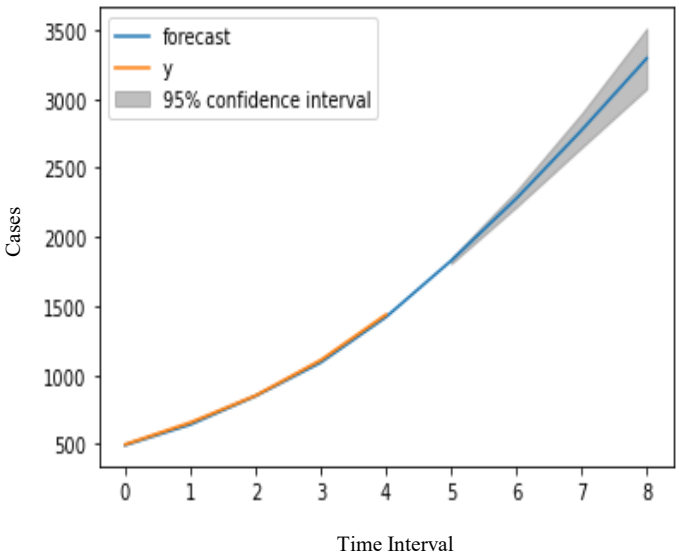

Fig. 10 Death Cases Analysis of COVID-19

The analysis of death cases using the ARIMA model is shown in Fig. 11, where the actual values for the death rate in India are displayed in the red line and the green line represents death cases as predicted from COVID-19 data. The outcome shows a mean squared error for the number of mortality cases in India of 46.839689.

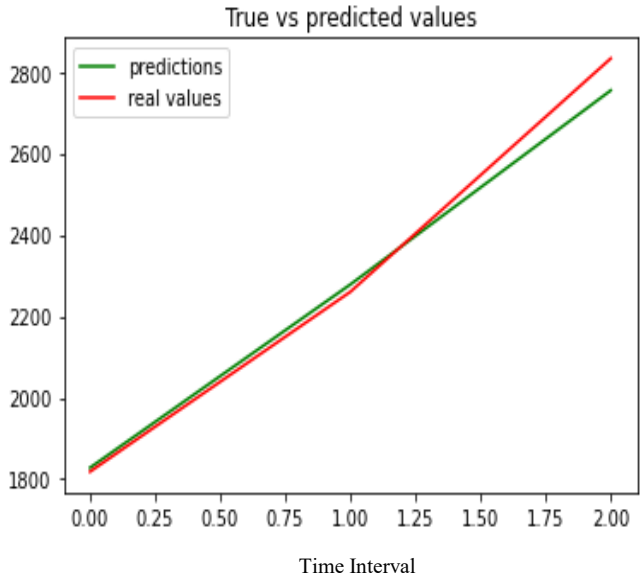

Fig. 11 COVID-19 Death Cases for Actual and Predicted Values

In this research, an ARIMA model is used to analyse and forecast changes in the COVID-19 disease's spread. The result illustrates that the ARIMA prediction result is very near to the actual results Additionally, we forecast the total confirmed, fatal, and recovered COVID-19 cases for the 30 days beginning on July 21 and ending on August 1, 2021, which was quite similar to the actual number of cases that occurred in India [16,19]. Our model calculated the mean square error for the ARIMA model to be between 2170.636098 and 46.839689 during performance evaluation.

VI CONCLUSION

The results of this research were derived from training data up to and including Jan 2022, to Jul 2021. Additionally, based on the current trend, there will undoubtedly be an increase in the number of instances. According to established medical standards, health professionals, and others included in contributing critical services must be guarded. The number of cases may rise exponentially as a result of future community spreading brought on by negligence on the part of both individuals and groups. Since the peak has not yet arrived, the Indian government must exercise increased caution and strictly enforce its regulations. Additionally, there must be a vigorous increase in the availability of medical facilities throughout the nation. For data that is collected on a weekly or biweekly basis, an instinctive system can be created in the future to retrieve data often and forecast the cases. Government agencies and medical facilities may keep an eye on demand and the level of care and isolation needed for new patients in this way. Data scientists from other regions can use this study to compare the performance of different ML models on the Indian dataset. Administrators and healthcare professionals can use this study to evaluate the condition in the coming future.

REFERENCES

[1] Li, Q., & Feng, W.," Trend and forecasting of the COVID-19 outbreak in China.", arXiv preprint arXiv:2002.05866, 2020

[2] World Health Organization., "Coronavirus disease 2019 (COVID-19), Situation report, 51., 2020

[3] Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., &Chowell, G., "Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020", Infectious Disease Modelling, vol. 5, pp. 256-263, 2020.

[4] Raji, P. and Lakshmi, G.D., "Covid-19 pandemic Analysis using Regression.", medRxiv, 2020

[5] Shaban, W.M., Rabie, A.H., Saleh, A.I. and Abo-Elsoud, M.A., "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier.", Knowledge-Based Systems, vol. 205, p.106270, 2020.

[6] Gupta, R., Pandey, G., Chaudhary, P. and Pal, S.K., "Machine learning models for government to predict COVID-19 outbreak", Digital Government: Research and Practice", vol. 1, no. 4, pp.1-6, 2020.

[7] Yadav, R.S., "Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India", International Journal of Information Technology, vol.12, no. 4, pp.1321-1330, 2020.

[8] Ghosh, P., Ghosh, R. and Chakraborty, B., "COVID-19 in India: statewise analysis and prediction.", JMIR public health and surveillance, vol. 6, No. 3, p.e20341, 2020.

[9] Zhou, C.M., Qin, X.R., Yan, L.N., Jiang, Y., Ke, H.N. and Yu, X.J., "Global trends in COVID-19.", Infectious Medicine., vol. I, pp. 31-39, 2022

[10] Kumari, R., Kumar, S., Poonia, R.C., Singh, V., Raja, L., Bhatnagar, V. and Agarwal, P., 2021. Analysis and predictions of spread, recovery, and death caused by COVID-19 in India. Big Data Mining and Analytics, vol. 4, no. 2, pp.65-75, 2021.

[11] Dharmshaktu, N. S., "The Lessons Learned from Current ongoing Pandemic Public Health Crisis of COVID 19 and its Management in India from Various Different Angles, Perspectives and way forward", Epidemiology International (E-ISSN: 2455-7048), vol. 5, no. 1, pp. 1-4, 2020.

[12] Sharma, S. and Guleria, K., 2022, April. Deep learning models for image classification: comparison and applications. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1733-1738). IEEE.

[13] Sharma, S., Guleria, K., Tiwari, S. and Kumar, S., 2022. A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans. Measurement: Sensors, 24, p.100506.

[14] Singh, S., Ramkumar, K. R., & Kukkar, A., "Machine Learning Techniques and Implementation of Different ML Algorithms". In 2021 2nd Global Conference for Advancement in Technology (GCAT) (pp. 1-6). IEEE, 2021.

[15] Singh, S., Ramkumar, K.R. and Kukkar, A., 2023. Analysis and Implementation of Microsoft Azure Machine Learning Studio Services with Respect to Machine Learning Algorithms. In Modern Electronics Devices and Communication Systems: Select Proceedings of MEDCOM 2021 (pp. 91-106). Singapore: Springer Nature Singapore.

[16] Singh, S. and Ramkumar, K.R., "Significance of Machine Learning Algorithms to Predict the Growth and Trend of COVID-19 Pandemic.", ECS Transactions, 107(1), p.5449. 2022.

[17] Liu, Z., Magal, P., Seydi, O., & Webb, G., "Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data". arXiv preprint arXiv:2002.12298, 2020

[18] F. Jiang, L. Deng, L. Zhang, Y. Cai, C. W. Cheung, and Z. Xia. 2020. "Review of the clinical characteristics of coronavirus disease 2019, (COVID-19)", J. Gen. Intern. Med., vol. 35 pp. 1545–1549, 2020

[19] Rajan Gupta and Saibal K. Pal. 2020. "Trend analysis and forecasting of COVID-19 outbreak in India", Retrieved from https://www.medrxiv.org/content/10.1101/2020.03.26.20044511v1.