2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), 18-19 December, Dhaka

Forecasting the COVID-19 pandemic in Bangladesh using ARIMA model

Julshan Alam Ratu[a*] , Md. Abdul Masud[b] , Md. Munim Hossain[a] , Md. Samsuzzaman[c]

[a]Faculty of Computer Science and Engineering, Patuakhali Science and Technology University, Bangladesh

[b]Department of Computer Science and Information Technology, Patuakhali Science and Technology University, Bangladesh

[c]Department of Computer and Communication Engineering, Patuakhali Science and Technology University, Bangladesh

Email: ratualam14@cse.pstu.ac.bd*, masud@pstu.ac.bd, munimhossain14@cse.pstu.ac.bd, sobuz@cse.pstu.ac.bd

2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI) | 978-1-6654-0009-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/STI53101.2021.9732576

***Abstract*— The effects of the coronavirus disease in 2019 are visible in every corner of the globe. The public health system is mostly affected, and the economic and social crises are also increasing day by day. Due to the widespread nature and the unavailability of drugs or vaccines for this pandemic, it is urgent to predict the COVID-19 infected cases to handle the situation more efficiently. Time series prediction is a crucial technique of the machine learning domain to deal with the issue. This research aims to predict the number of daily confirmed COVID-19 cases for a successful time. To forecast COVID-19 instances in Bangladesh, we use the Autoregressive Integrated Moving Average (ARIMA) model. The experimental results show that the estimated best models are: ARIMA(3,1,0) with drift, ARIMA(3,1,2) with drift, ARIMA(5,1,0) perform significant predictions on three different kinds of COVID-19 datasets.**

Keywords*— *Akaike Information Criterion, ARIMA, COVID-19, Forecasting, Time Series

I. Introduction

The global epidemic caused by the coronavirus, a contagious virus, is a matter of concern to mankind today. The COVID-19 pandemic is the most serious global health threat we've seen since World War II. On December 8th, 2019, a novel Corona Virus Disease (COVID-19), a member of the SARS Corona-virus-2 (SARS-CoV-2) family, began infecting patients in Wuhan, China [1]. Three highly virulent and lethal human coronaviruses, SARS-CoV, MERS-CoV, and SARS-CoV-2, have developed in the last two decades [2]. Both SARS-CoV, which arose in China in 2003, and MERS-CoV, which emerged from the Middle East in 2012, caused severe symptoms [3, 4]. SARS primarily occurred in China and quickly spread all around the world, with over 8000 infected individuals and 776 diseases. In 2012, some Saudi Arabian people were found to be infected with a new coronavirus, ten years after the first. The Middle East Respiratory Syndrome Coronavirus was named after the virus was recognized as a member of the coronavirus family (MERS-CoV) [5].

In this research, we apply the ARIMA model to forecast the new confirmed cases of COVID-19 in Bangladesh during the coming week. This model is very popular for forecasting, and it provides good accuracy for time series data. We apply the ARIMA model and it performs almost an identical result to its real value. Bangladesh, being a country with a large population, is prone to the spread of COVID-19. Bangladesh's healthcare condition is deteriorating because of the number of COVID-19 affected people. So the rationale for this study's focus on Bangladesh is worth highlighting.

However, the COVID-19 virus was declared a pandemic on March 11, 2020, because of its spread from China to neighboring countries [6]. Following the stay-at-home, forced face masks, and social-distancing directives, most nations are currently near a breaking point in terms of health care. Around 190 nations have been impacted, with big outbreaks in the United States, Italy, Spain, France, and China [7].

The study's purpose is to predict the newly confirmed COVID-19 cases in Bangladesh over the next seven days. Check the datasets stationarity using the ADF test. Following the removal of the trends and seasonality, we chose the best model according to AIC values from the three datasets, and it demonstrates a good result for every dataset. This will aid in determining the future state of epidemics in Bangladesh. The ARIMA models: ARIMA(3,1,0) with drift, ARIMA(3,1,2) with drift and ARIMA(5,1,0) perform significantly and show a good result in the upcoming days in Bangladesh.

II. Related Work

Several research studies have been proposed to estimate the spread of the disease. Isra Al-Turaiki et al. proposed a method using seven forecasting models for future forecasting [8]. The results indicated that among the seven forecasting models, in most analyses, ARIMA exhibited the lowest forecasting prediction error. But there are some drawbacks to this forecasting approach. Firstly, it might not be precise or dependable enough for long-term forecasting. Secondly, pinpointing the precise reasons why some models outperform others is difficult. Gülhan Toğa et al. developed an approach for examining the dynamics of COVID-19 prevalence in Turkey using ARIMA and Artificial Neural Networks. The findings showed that the methodologies used were quite effective in determining prevalence in Turkey [9]. However, for a more accurate assessment, data needs to be up to date and new parameters in real-time with a view to affecting the prevalence of the outbreak have to be taken into consideration. Qiuying Yang et al. applied the ARIMA model to Italy, based on new cases and deaths in Hubei, to track the epidemic's progress [10]. It will serve as a theoretical foundation for the future evolution of pandemics in some nations. However, their model is more suited to short-term forecasting. In our study, we predicted new cases with AIC criteria. We choose the best ARIMA models from the various ARIMA models with identical AIC values and find the best model from all the datasets.

978-1-6654-0009-1/21/$31.00 ©2021 IEEE

III. METHODOLOGY

A. Model Description

The ARIMA model is the most commonly used model for forecasting future data on time series data. Demand is forecasted using time series forecasting models, which employ mathematical methodologies based on past data [11]. In the ARIMA model, the differenced autoregressive and moving average models are combined into one. The ARIMA model has three fundamental forms if the data evidence is stationary in the distinct parts of the regression: the autoregressive (AR) model, the moving average (MA) model, and the autoregressive integral moving average (ARIMA) model [12].

An AR model is a strategy for predicting future or current behavior in a time series based on data from previous behaviors in the same time series. This model is based on the data's lag values. A pure AR model is one in which y_t is solely reliant on its own latency. As a result, the y_t specific lagged values are employed as predictor value. Lags affect the outcome of one time period on to the next. The equation that defines the AR model is below:

$$y_t = \delta + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \varepsilon_t \quad (1)$$

Where:

- $y_{t-1}, y_{t-2} \ldots y_{t-p}$ are values from previous series (lags).
- ϵ_t is white noise (i.e. randomness).
- and δ is given by the equation below

$$\delta = \left(1 - \sum_{i=1}^{p} \Phi_i\right) \mu \quad (2)$$

The process mean is denoted by μ.

The dependence between a moving average model's residual error and observation is used in an MA model. The only thing that determines y_t in a pure moving average model is the lagged prediction errors, rather than historical values of the forecast variable in a regression. The equation that defines the MA model is:

$$y_t = \alpha + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q} \quad (3)$$

Where:

- The trend parameter q is a moving average.
- $\epsilon_{t-1} + \epsilon_{t-2} + \cdots + \epsilon_{t-q}$ are the errors at pervious time period.
- α is a constant.
- Φ is the numeric coefficient.
- ϵ_t is white noise (i.e. randomness)

An ARIMA model requires the AR and MA terms to be combined after that to make a time series stationary, it has been differenced at least once. The ARIMA model may be stated as follows if the series is stationary:

$$Y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_p y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_q \varepsilon_{t-q} \quad (4)$$

ARIMA version is diagnosed with the aid of putting the order in for 3 phases: p and q stand for AR and MA. The order of difference stands for d.

B. Parameter Estimation

We must first determine if the data is stationary or not before finding the parameters for the ARIMA model. Each ARIMA component functions as a parameter with a defined syntax. The typical notation is p, d, and q, with integer values indicating the ARIMA model employed. The parameters are as follows:

- p: the number of lag observations in the model is known as lag order.
- d: number of times that the data are differenced.
- q: the moving average's order.

The Augmented Dickey-Fuller test (ADF test) is a standard statistical technique for determining if data is stationary. The null hypothesis H_0 is that the Non-stationary time series are required. As we use three different datasets, we need to check for all three datasets. The ADF test result suggests that ($p > 0.05$) and we assume the data is non-stationary. After the first difference between all three datasets, the datasets became stationary. The p-value ($p < 0.05$) obtained was, but the importance level and thus the ADF value were less than the critical values. The d value for each dataset is 1, and the null hypothesis was also rejected.

To determine the order of the ARIMA model, the autocorrelation function (ACF) and partial autocorrelation function (PACF) employ data as the key tool. The delayed correlation, which is the relationship between two time series data, is displayed by ACF. The correlation coefficients among the series and lags of itself MA(q) are plotted in the ACF plot, while the partial correlation coefficients between the series and lags of itself are plotted in the PACF plot. The PACF aids in the identification of potential orders for the term AR(p).There is a more systematic way to do this. The ARIMA model's parameters were chosen by the Akaike information criterion (AIC). The fundamental formula is as follows:

$$AIC = -2(\text{log-likelihood}) + 2K \quad (5)$$

Where:

- K is the number of model parameters.
- Model fit is measured using log-likelihood. The better the match, the higher the number. This is typically gleaned through statistical data.

C. Select Best Arima Model

To compare and choose the best model, we can successfully visualize the ARIMA models. The AIC is used to assess the quality of the model. We calculate the AIC and find that the model with the lowest AIC is superior to the others. In terms of several models with almost identical AIC values, we can plot them and reduce the confusion and by doing this procedure we can select the proper model.

D. Prediction with Best ARIMA Model

From the time series dataset, we predict the future behaviour using the best ARIMA model. This best model is

likely to give us a decent forecasted result. The ARIMA model's forecasting procedure is depicted in fig. 1.

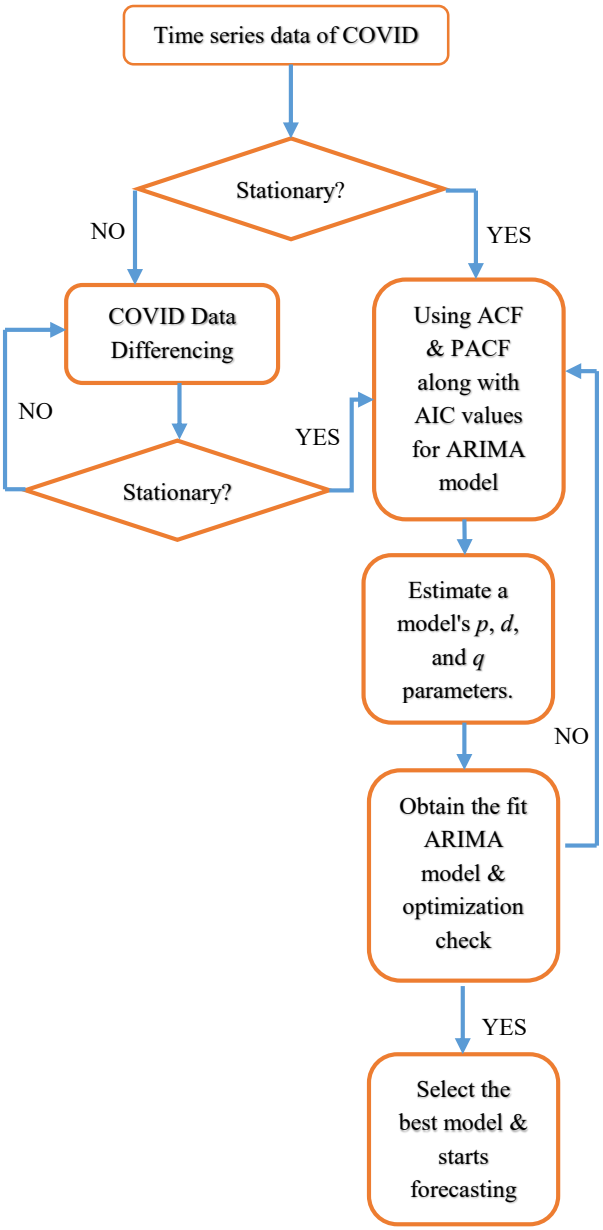

Fig. 1. Model selection procedure for forecasting.

IV. Experiments

A. Data Description

The datasets used in this work are from the World Health Organization's (WHO) authoritative site and refer to daily confirmed COVID-19 cases in Bangladesh. In this study, three datasets were used. This data set pertains to the number of confirmed cases in Bangladesh on a daily basis. The first dataset shows the confirmed COVID-19 cases in Bangladesh on a daily basis from 3rd April to 2nd June, 2020. We used a total of 61 data points in the first dataset. On April 3rd, the confirmed cases of COVID-19 were 5. Following a week, the count of affirmed cases came to 112. Then, at that point, the number of affirmed cases quickly expanded day by day. In the second dataset, we use more data than in the first dataset. The second dataset shows the confirmed COVID-19 cases in Bangladesh on a daily basis from 3rd April to 24th July, 2020. In the second dataset, the most elevated daily affirmed cases were 4019, which were accounted for on July 2nd. In July 2020, the growth rate was much higher than in April 2020. However, in the third dataset, we used a total of 164 datasets. In that dataset, we take the confirmed new cases in Bangladesh from 3rd April to 13th September, 2020. We forecast the next one-week's daily anticipated COVID-19 confirmed cases in each dataset.

B. Experimental Setting

Several types of experiments were done for the model selection. The ARIMA model's parameters are determined by the AIC value that is the lowest. Tables I, II, and III exhibit the values of the AIC for various p and q parameters. We tried taking various upsides of the two parameters p and q, going from 0 to 5 with the float in each of the three datasets of every day affirmed instances of COVID-19, depending on ADF test d was picked 1. In the event that d=1, there is a pattern with incline μ. In this case, the value of μ is also an estimate of the mean of the differenced data. So the presence of a trend is of interest, we should check the standard error of μ to ensure the slope is significantly different from zero. The output will include the drift coefficient and standard error.

TABLE I. Selection Of ARIMA Order Based On AIC For First Dataset.

Models	AIC
ARIMA(0,1,2)	826.4783
ARIMA(1,1,0) with drift	820.6894
ARIMA(3,1,0) with drift	**818.5061**
ARIMA(3,1,2)	819.6658
ARIMA(4,1,0) with drift	818.8721
ARIMA(5,1,0) with drift	820.7813
ARIMA(1,1,0)	824.7408

TABLE II. Selection Of ARIMA Order Based On AIC For Second Dataset.

Models	AIC
ARIMA(0,1,1) with drift	1581.641
ARIMA(0,1,1)	1582.803
ARIMA(0,1,4) with drift	1582.602
ARIMA(1,1,1) with drift	1583.313
ARIMA(3,1,2) with drift	**1578.224**
ARIMA(0,1,4)	1584.08
ARIMA(1,1,1)	1584.746

TABLE III. Selection Of ARIMA Order Based On AIC For Third Dataset.

Models	AIC
ARIMA(0,1,3)	2343.84
ARIMA(3,1,1)	2342.526
ARIMA(5,1,0)	**2342.038**
ARIMA(0,1,3)	2343.84
ARIMA(4,1,0)	2344.99
ARIMA(4,1,1)	2343.285
ARIMA(1,1,1)	2345.608

With drift, it means fitting a model that fluctuates around a trend (up or down). Models with the lowest AIC value are chosen as the best. Accordingly, ARIMA(3,1,0) with drift, ARIMA(3,1,2) with drift, ARIMA(5,1,0) are the best models for each of the datasets, which are shown in Tables I, II, and III.

We perform forecasting of the next one-week confirmed cases using three time-series datasets. We divided the data into two sets: training and testing. After training the datasets, the testing datasets are used for forecasting the confirmed number of cases for each day. The plot of total daily confirmed new cases in Bangladesh from 3rd April to 2nd June 2020 is shown in fig. 2, from 3rd April to 24th July 2020 is shown in fig. 3, and the data from 3rd April to 13th September 2020 is shown in fig. 4. The figs. 2, 3, and 4 show that the trend of confirmed cases is increasing with time and the variances of time series are not stable, which leads to the variables being non-stationary. Therefore, data needs to be stationary. We use a common statistical test, namely, the Augmented Dickey-Fuller Test (ADF Test) is used to check the stationarity of datasets. After completing the first difference of the ADF test, the three datasets become stationary. This stabilized the time series at the required level of 5% significance and indicates that it is suitable for further analysis.

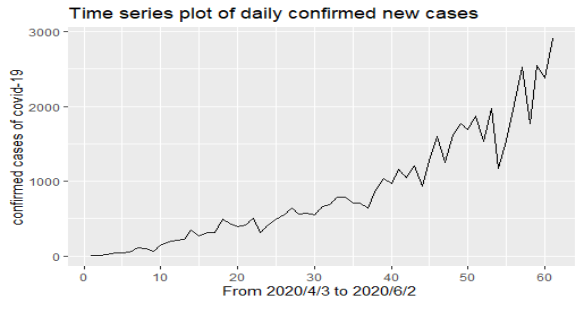

Fig. 2. Trend of daily confirmed cases on 1st dataset.

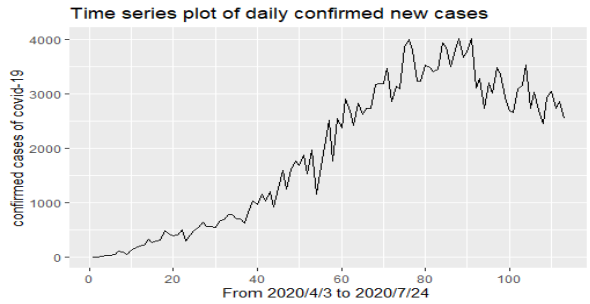

Fig. 3. Trend of daily confirmed cases on 2nd dataset.

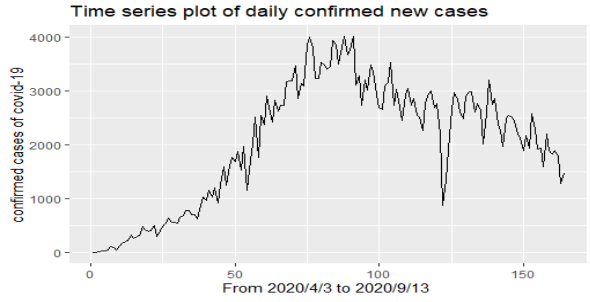

Fig. 4. Trend of daily confirmed cases on 3rd dataset.

C. Result and Discussion

The data for these experiments was gathered from the WHO's official website, which included daily verified cases in Bangladesh. We assigned the first dataset with the daily number of confirmed cases of COVID-19 in Bangladesh from 3rd April to 2nd June, 2020. Similarly, we collect the confirmed cases from 3rd April to 24th July, 2020 as the second dataset and from 3rd April to 13th September, 2020 as the third dataset. We present the forecasting performance of three datasets in figs. 5, 6, and 7.

If the confidence level of residuals is at 95%, the Ljung-Box Q test is used to examine pure white noise when the value(p>0.05) for models of the ARIMA(3,1,0) with drift, ARIMA(3,1,2) with drift, ARIMA(5,1,0) models. Because these models already have a strong fit, they need to be able to forecast the COVID-19 confirmed cases in Bangladesh for the next week. Figs. 5, 6, and 7 show the predicted values with the fitted models from 3rd June to 9th June 2020, 25th July to 31st July 2020, and from 14th September to 20th September 2020, respectively. Fig. 5 shows an upward trend in forecasted value. Figs. 6 and 7 show an upward and downward trend in forecasted values. From figs. 5, 6, and 7, we can notice that the forecasted value is indicated in the blue area.

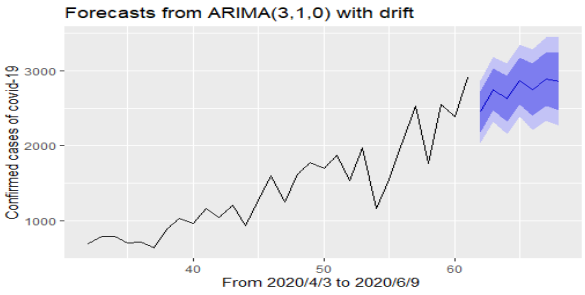

Fig. 5. Forecasting of COVID-19 daily confirmed cases on 1st dataset.

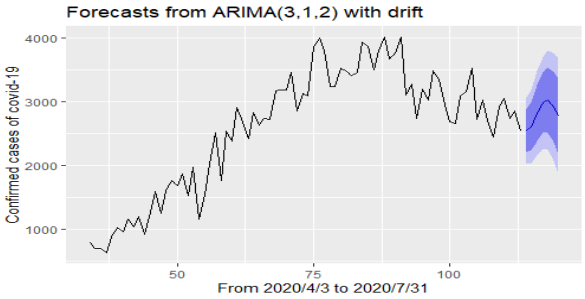

Fig. 6. Forecasting of COVID-19 daily confirmed cases on 2nd dataset.

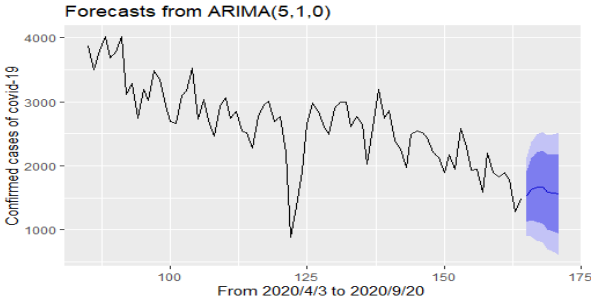

Fig. 7. Forecasting of COVID-19 daily confirmed cases on 3rd dataset.

TABLE IV. THE ARIMA MODEL'S PREDICTED VERSUS ACTUAL VALUES.

Date	Predicted	Actual	Date	Predicted	Actual	Date	Predicted	Actual
3/6/20	2450	2695	25/7/20	2551	2520	14/9/20	1523	1812
4/6/20	2748	2423	26/7/20	2616	2275	15/9/20	1631	1724
5/6/20	2628	2828	27/7/20	2826	2772	16/9/20	1664	1615
6/6/20	2861	2635	28/7/20	2991	2960	17/9/20	1667	1593
7/6/20	2747	2743	29/7/20	3030	3009	18/9/20	1586	1541
8/6/20	2890	2735	30/7/20	2926	2695	19/9/20	1570	1567
9/6/20	2857	3171	31/7/20	2780	2772	20/9/20	1562	1544

Indisputably, the world has never seen a pandemic like COVID-19. The prime objective of this study is to track and anticipate the projected number of newly confirmed coronavirus patients in Bangladesh using the ARIMA model, which is derived from the total confirmed cases on a daily basis. In this research, we found ARIMA models that closely matched the distribution of COVID-19 in Bangladesh. This modeling gives us an understandable pandemic situation and helps authorities become effective with epidemic response strategies and make productive decisions. As a result, its influence on society, the environment, medical management, the economy, and education is limited.

As previously stated, our ARIMA model prediction technique outperformed many current models based on many methodologies and results. We predict the daily COVID-19 confirmed cases from 3rd June to 9th June, 2020 in Bangladesh using ARIMA(3,1,0) with drift and fig.8 shows us the comparison of actual confirmed instances against predicted cases. The blue lines are actual values and the red ones are the forecasted values. From fig.8, it's a reasonable estimate that daily confirmed cases will grow from 2450 to 2857.

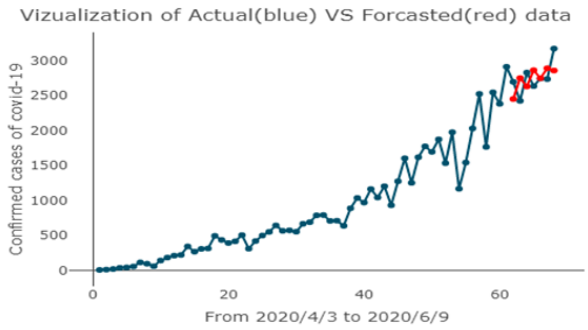

Fig. 8. Representation of Actual confirmed cases and predicted cases on 1st dataset.

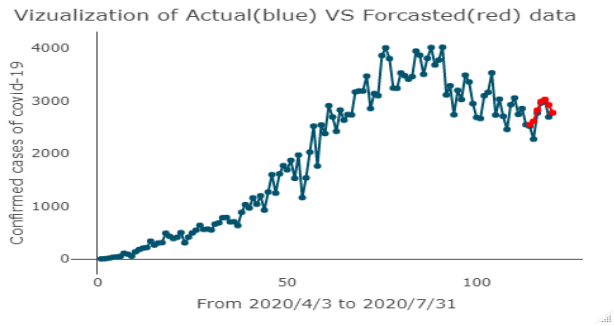

Fig. 9. Representation of Actual confirmed cases and predicted cases on 2nd dataset.

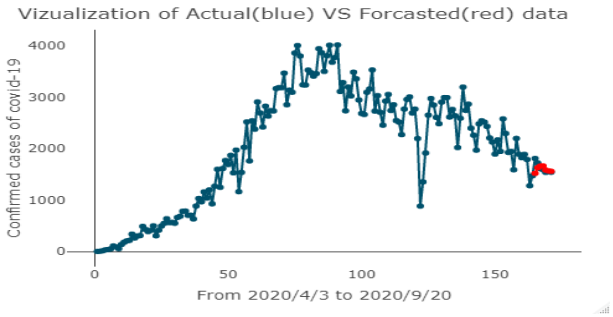

Fig. 10. Representation of Actual confirmed cases and predicted cases on 3rd dataset.

Fig. 9 shows that the predicted values from 25th July to 31st July, 2020 are also pretty close to actual values. The daily confirmed cases could rise from 2551 to 2780 daily cases. Fig. 10 also shows that the actual values and predicted values in the graph are nearly identical in the visualization from September 14th to September 20th, 2020, where the confirmed cases remain between 1523 and 1562. The confirmed cases show a decreasing curve in the latter parts of the pandemic. Alternatively, Table IV is reported to better understand the relationship between actual and predicted confirmed cases.

ACKNOWLEDGMENT

This research was supported by the Research and Training Center (RTC) under project Code No. 3631108 in Patuakhali Science and Technology University, Bangladesh.

V. CONCLUSION AND FUTURE WORK

It has been quite a long time since the world has seen such a tough time. A considerable number of people have died from the COVID-19 pandemic all over the world, and it is a threat on an unparalleled scale to every possible sector. The economic and social structures have been devastated by this disease. Furthermore, the pandemic has had a negative impact on every industry in Bangladesh. We discovered accurate forecasting models for projecting the trend of coronavirus cases using different ARIMA settings. In this tough time, the government of Bangladesh needs to take some serious steps to minimize the loss of people's lives. Besides all of this, people must wear a mask and maintain physical and social distance. This study forecasts the total number of new cases in Bangladesh during the next seven days. The model's forecast of the current condition will be useful in predicting the future. This will help the authorities make valuable decisions. By incorporating death cases and recoveries into the current work, it will be possible to predict the spread of COVID-19 in the

future. Also, it can be developed for complex time series data sets with precise forecasting.

REFERENCES

[1] C. C. Lai, T. P. Shih, W. C. Ko, H. J. Tang, P. R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges," International journal of antimicrobial agents, 2020, vol. 55, no. 3, pp. 105924.

[2] Z. Zhu, X. Lian, X. Su, W. Wu, G. A. Marraro & Y. Zeng , "From SARS and MERS to COVID-19: a brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses," Respiratory research, 2020, vol. 21, no. 1, pp. 1-14.

[3] D. S. C. Hui, A. Zumla, "Severe Acute Respiratory Syndrome: Historical, Epidemiologic, and Clinical Features," Infectious Disease Clinics, 2019, vol. 33, no. 4, pp. 869-889.

[4] E. I. Azhar, D. S. C. Hui, Z. A. Memish, C. Drosten, A. Zumla, "The middle east respiratory syndrome (MERS)," Infectious Disease Clinics, 2019, vol. 33, no. 4, pp. 891-905.

[5] N. Kaur, A. Sethi, HC Patil, S. Singh, H. Kaur, U. K. Mishra, "Origin and Evaluation of Pathogenic Coronavirus," International Journal of Health Sciences and Research, 2020, vol. 10, no. 7, pp. 207-217

[6] O. A. Adegboye, A. I. Adekunle, E. Gayawan, "Early transmission dynamics of novel coronavirus (COVID-19) in Nigeria," International Journal of Environmental Research and Public Health, 2020, vol. 17, no. 9, pp. 3054.

[7] N. Chintalapudi, G. Battineni, F. Amenta, "COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach," Journal of Microbiology, Immunology and Infection, 2020, vol. 53, no. 3, pp. 396-403.

[8] I. A. Turaiki, F. Almutlaq, H. Alrasheed, N. Alballa, "Empirical Evaluation of Alternative Time-Series Models for COVID-19 Forecasting in Saudi Arabia," International Journal of Environmental Research and Public Health, 2021, vol. 18, no. 16, pp. 8660.

[9] G. Toğa, B. Atalay, M. D. Toksari, "COVID-19 prevalence forecasting using Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN): Case of Turkey," Journal of Infection and Public Health, July 2021, vol. 14, no. 7, pp. 811-816.

[10] Q. Yang, J. Wang, H. Ma, X. Wangb, "Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy," Journal of Infection and Public Health, 2020, vol. 13, no. 10, pp. 1415-1418.

[11] J. Fattah, L. Ezzine, Z. Aman, H. E. Moussami, A. Lachhab, "Forecasting of demand using ARIMA model," International Journal of Engineering Business Management, 2018, vol. 10, pp. 1-9.

[12] Q. Yang, J. Wang, H. Ma, X. Wang, "Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy," Journal of Infection and Public Health, 2020, vol. 13, no. 10, pp. 1415-1418.

[13] F. Mahia, A. R. Dey, M. A. Masud and M. S. Mahmud, "Forecasting Electricity Consumption using ARIMA Model," 2019 International Conference on Sustainable Technology for Industry 4.0(STI), 2019, pp. 1-6.