IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 5, Ser. III (Sep. – Oct. 2020), PP 41-47
www.iosrjournals.org

Spread Analysis and Prediction of COVID-19 in Bangladesh using Linear Regression

Asma Yasmin[1,3], Abdullah Al Rahat[1] , Kamrunnahar Kali[2], Iqbal Aziz Khan[3]
[1](*Department of Computer Science and Engineering, Bangladesh Army University of Engineering & Technology, Natore-6431, Bangladesh)*
[2](*.Department of Physics, Comilla University, Kotbari, Comilla.)*
[3](*Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh)*

Abstract: *COVID-19 is a disease caused by the virus namely SAR-COV-2 which has created death panic all over the world. The acute spread of the virus makes the World Health Organization (WHO) to declare the outbreak as a pandemic. In Bangladesh, the spread of the virus was slow at the early stages but a few month later, the number of infected people and causalities grows rapidly. In this study, we conduct numerical and exploratory analysis on global and local data of COVID-19 to explore reliable trends of COVID-19 transmission and do accurate predictions of the outbreak in Bangladesh. The experimental analysis showed that the outbreak has reached its maximum peak occurring from late May to June. At this time, the recovering rate was very less than that of infection. The experimental analysis also has confirmed that community transmission has already happened in Bangladesh. However, August, daily infection cases going down. We here applied a linear regression technique to predict the infected cases along with recovery and fatal cases.*

Key Words: *COVID-19 disease, Linear Regression, outbreak, pandemic.*

Date of Submission: 31-10-2020 Date of Acceptance: 12-11-2020

I. Introduction

Pandemics are frequently observed over the centuries and the outcomes of these pandemics have consistently had an immense stun on the world. COVID-19 is a pandemic that has influenced more than 170 nations around the globe [1]. The quantity of affirmed and lethal cases have been expanding at a terrible rate in practically all the influenced locales. None of the previous outbreaks can make such a level of panic and economic shutdown across the globe. According to WHO, the current outbreak of coronavirus disease (COVID-19) was first reported from Wuhan, China, on 31 December 2019. Globally, by March approximately 170,000 confirmed cases of COVID-19 caused by the SARS-CoV-2 have been reported along with an estimated 7,000 deaths in approximately 150 countries [2]. On March 11, 2020, the World Health Organization declared the COVID-19 outbreak a pandemic [3]. Primary Data from China have indicated that older adults with serious underlying health complications are at higher risk for suffering from severe COVID-19–associated illness and death than are younger persons [4]. A report published from China showed that COVID-19 cases in China were mild (81%), approximately 80% of deaths come from among adults aged ≥60 years; very low rate (0.1%) death occurred in people aged ≤19 years [4].

To take an effective policy to combat the outbreak, do accurate predictions requires the understanding of the natural succession of the disease. COVID-19 generally progresses through the exposure from the infected person to the non-infected person through bodily contact or through droplets containing virus, getting inside the non-infected person. When an infected host comes in contact with others non-infected people, then the disease begins to spread.

Earlier in May, WHO (2020b) formulated a general guideline for the governments that want to relax lockdown or restriction on economic reopening, where six criteria are stated as follows: (1) infection transmission is under control; (2) health system can "detect, test, isolate, treat every infection case and track every case"; (3) risks are minimized for the vulnerable hot spot areas, such as nursing homes; (4) protective measures are established for educational institutes, workplaces, and other essential places; (5) the probable risk of imported new cases can be managed; and (6) the communities are thoroughly educated, engaged, empowered, and willing to function according to the new standard or normal. But, alarming was for a country like Bangladesh is lifting the lockdown too early or too quickly can raise the rate of the infection because Bangladesh has no trustworthy and technically sounds health management policy and legislative structures to fight against the COVID-19 outbreak. At the early stage, there are only 399 Intensive Care Units (ICUs) in the government hospitals in Bangladesh—of which 218 are in the Dhaka city alone and 20% of COVID-19 patients require ICU, due to the shortage of ICU the hospitals were unable to provide it [5].

Bangladesh confirmed the first coronavirus case on 8 March 2020 and In response to the COVID-19 pandemic, the Government of Bangladesh declared shutdown the educational institutional from 17 march and a general leave had been imposed from 26 March in the name of "lockdown" and extended it up to 30 May 2020 in seven different time slots [5]. In addition, the closure of the educational institution has been prolonged on 31 October, and there is still uncertainty regarding the reopening of the educational institution. The government banned all the movements and restricted and urged people to stay at home. The citizens were only allowed to step out only in emergencies. All these steps were taken in the hope of flattening the curve of infected cases and to limit the exponential growth of the patients in Bangladesh [6].

Previously, Kaplan's model was used to predict the spread of the HIV/AIDS pandemic [7]. This prediction focused on the spread pattern related to the specific group of people who takes drug-using injector/syringe. Another virus namely MERS has been studied to analyze the transmission route using Decision Tree and Apriori Algorithms [8]. In [9] support vector machine (SVM), a popular machine learning technique was used to access the spread pattern. Moreover, A neural forecasting model was used in [10] for obtaining a forecast for swine flu. In [11] a maximum likelihood method was used to assess the spread of the SARS epidemic using the construction of a phylogenetic tree.

We use a linear regression model with polynomial features to predict the confirmed cases, recovery cases, and fatal cases in Bangladesh. The remaining part of this study is organized as follows:

Section 2 describes the overview of COVID-19, section 3 depicts the linear regression model as our prediction model. Section 4 illustrates the experimental analysis and discussion, finally, Section 5 concludes the paper.

II. Overview of COVID-19

COVID-19 is found highly contagious disease caused by coronavirus-2 which affects the respiratory system of the human body. The incubation period of this disease is 1–14 days or even longer [12]. Most common symptoms include fever, tiredness, and dry cough and less common symptoms: aches and pains, diarrhea, conjunctivitis, headache, loss of taste or smell, shortness of breath, aches and pains, and sore throat, Serious symptoms: difficulty breathing or shortness of breath chest pain or pressure loss of speech or movement [13]

To combat the spread at an unprecedented rate, some preventive measures like a complete lockdown of the heavily infected areas, ban on international travels, suspending schools and other non-essential daily activities are taken in most of the infected countries to limit interpersonal contact, considering the contagious nature of the disease. Experts assumed that the older individuals were profoundly defenseless to the virus and statistical data also showed that because of the weak immune system the elderly succumb to the disease, the young children of developing immune systems are at higher risk [1]. Besides, people who have diseases like diabetes, high BP, asthma, cancer, cardiovascular disease, etc. become affected critically and also experienced fatal cases[14].

III. Learning and Prediction Methods

A prediction has been done based on various prediction techniques and different data sources. Population statistics are an easy way to perform prediction tasks because it does not require the sampling as the entire population is present in the dataset. Population statistics also help to make reliable predictions and estimates with less computational overhead and there is a lack of bias. In the literature, many studies are also carried out on clinical data. These studies may be useful for a physician, doctors, and researchers in the medical domain for investigating better diagnostic methods and for pharmaceutical industries in formulating vaccines, drugs in a short time. Here we have studied machine learning techniques that are used for predicting confirmed cases, recovery cases, and death cases that help policymakers to draw proper guidelines to combat the spread. Major significant parameters are listed below, Daily death count, Number of carriers, Incubation period, Environmental parameters, i.e., temperature, humidity, wind speed., Social distancing, quarantine, isolation, Transmission rate, Mobility, Geographical location, Age and Underlying disease, and many more. Apart from these above-mentioned parameters, there can be many influential factors that need to be further investigated.

Machine learning techniques are used worldwide for predictions due to their accuracy. However, to use machine learning (ML) techniques, there are a few challenges as very little data is available. For instance, the challenges involved in training a model are the appropriate selection of parameters and the selection of the best ML model for prediction. Researchers have done predictions based on datasets that are available and used the best ML model as per the dataset [14, 15, 16, 17]. Kumar and Hembram [18] presented a model based on the Logistic equation, of the virus. DeCapprio et al. [19] proposed a model using logistic regression, gradient boosted trees, and a hybrid model using Medicare data. The outcome of these models will help to initiate control strategies and to initiate corrective measures in time to control the outbreaks.

Linear regression is a linear approach to model the relationship between a scalar dependent response and one or more explanatory independent variables. For one explanatory variable, the regression is said simple linear regression while for more than one explanatory variable, the process is called **multiple linear regression** [20]. Once a regression relationship is obtained, it can be used to predict the values of the response variable, and sometimes it is also used to recognize variables that mostly influence the response. Let establish a linear relationship between the response variable, y and the predictor variable, $x_i, i = 1,2..., n$ of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

where

$\beta_0, \beta_1 ..... \beta_n$ are regression coefficients (unknown model parameters), and

ε is the error due to variability in the observed responses.

IV. Experimental Analysis and Discussion

4.1 Dataset and Experimental Setup

For data analysis, COVID-19 dataset of John Hopkins University is collected from the Github public repository [21]. Data is also taken from the official GitHub repository of the COVID 19 dataset for competition (week 4) in Kaggle. Coding is done in python language in the Kaggle notebook environment.

4.2 Data Analysis

Fig.1 depicted the global cumulative confirmed cases over time (more than 200 days from January 2020), this global plotting showed that global case still raising slowly i.e near to linearly and recovering cases hopefully increasing exponentially. Besides, active cases also flattening and deaths case highly flattened. In fig 2., global spread over time has been depicted only for the countries that were high test rates. Here, cumulative confirmed cases linearly increasing which was exponential trends at the early stage of the outbreak. However, recovery and active cases similarly increasing. Fig 3. Illustrated the global spread overtime only for low test rate countries. Here it is observed that the cumulative confirmed cases still somehow exponentially increasing, but death cases are not as alarming as before seen as it is also flattened like high test rate countries. Active cases also going to flattened.

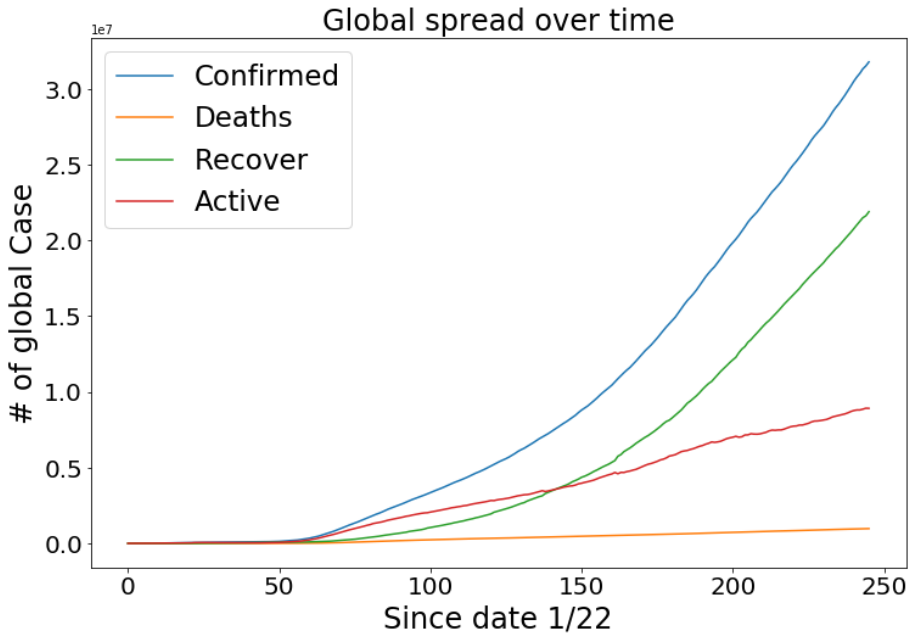

Fig.1 Spread globally over time.

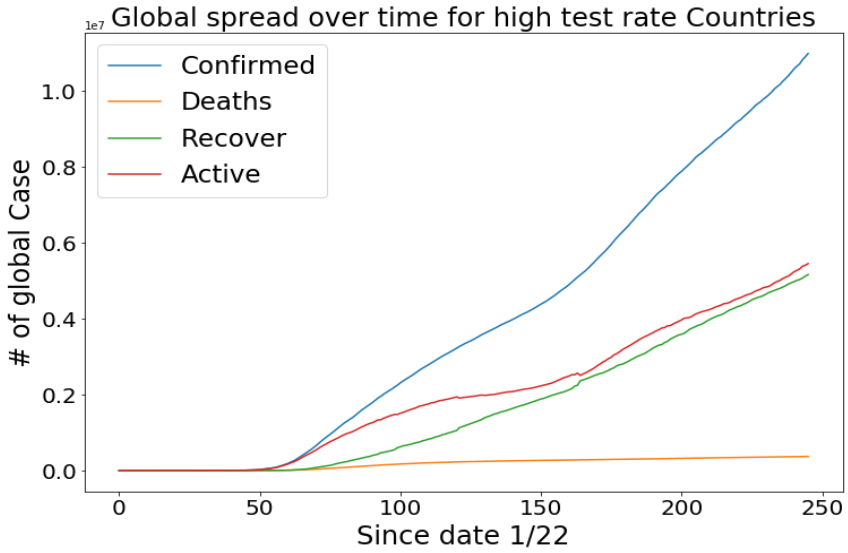

Fig.2 Spread globally over time for high test rate countries.

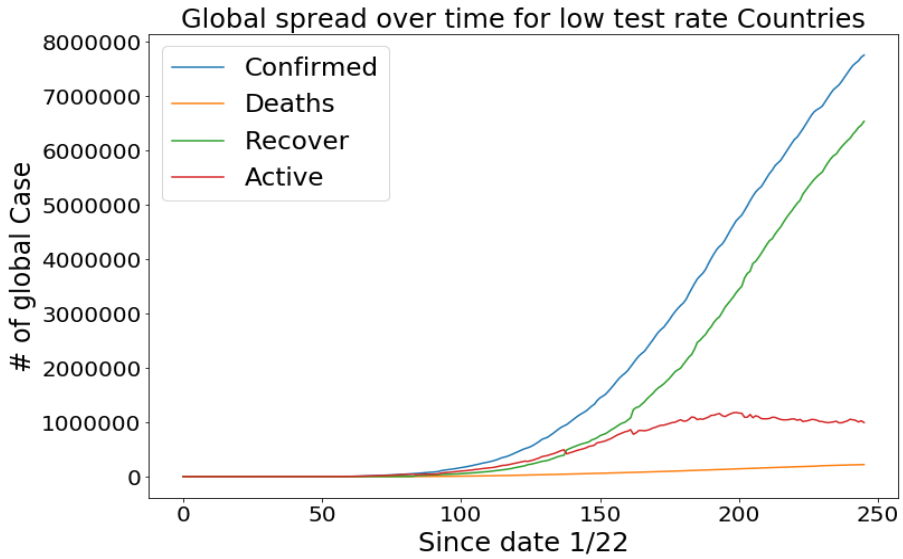

Fig. 3 Spread globally over time for low test rate countries.

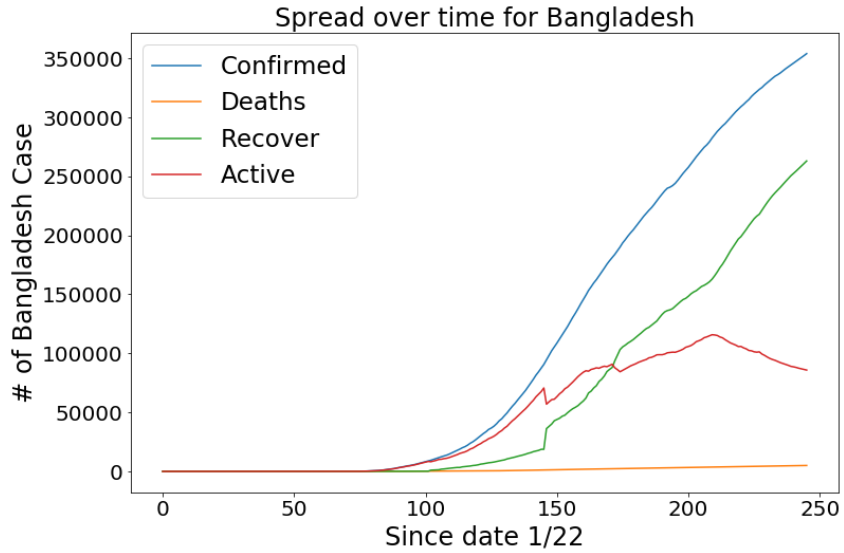

Fig 4: Spread over time for Bangladesh.

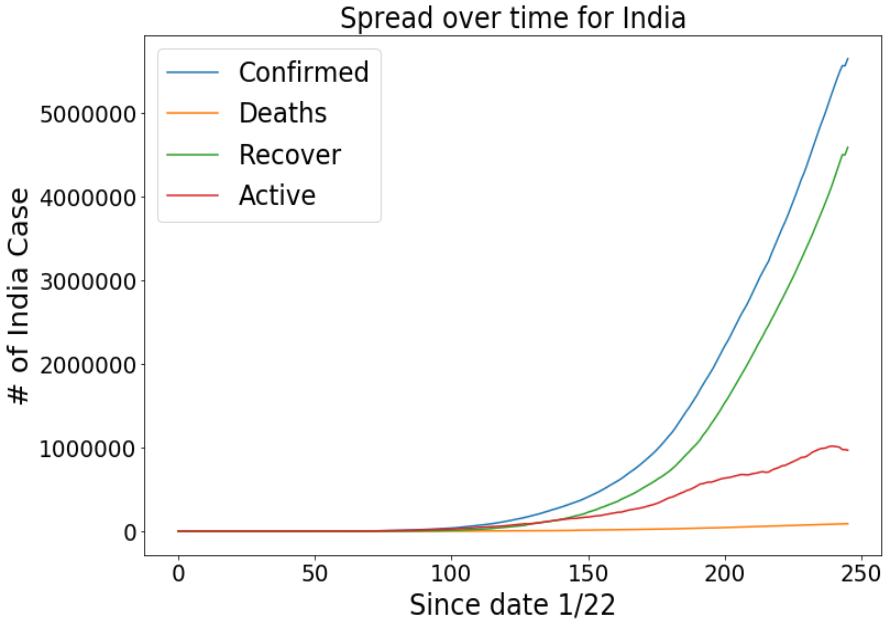

Fig 5: Spread over time for India.

Fig. 4 demonstrates the spread over time for Bangladesh. This plotting is following as same trends as low test rate countries. However, the pandemic followed the exponential graphing after 80 days from the beginning of the outbreak. And in May and June were the peak for infection. Though death cases are not improving and cumulative confirmed cases still following an exponential path in the plotting. Fig. 5 depicted that Bangladesh's neighboring country India cases where this plotting also following a similar pattern of Bangladesh. From the above plotting, as a whole, we can conclude that countries of low test rate and densely populated still suffering from an exponential increase in cumulative confirmed cases than that of high test rated countries.

4.3 **Prediction of COVID-19 cases in Bangladesh :**

Here we used Linear Regression with polynomial features to predict confirmed case,death case and recovery cases in Bangladesh.

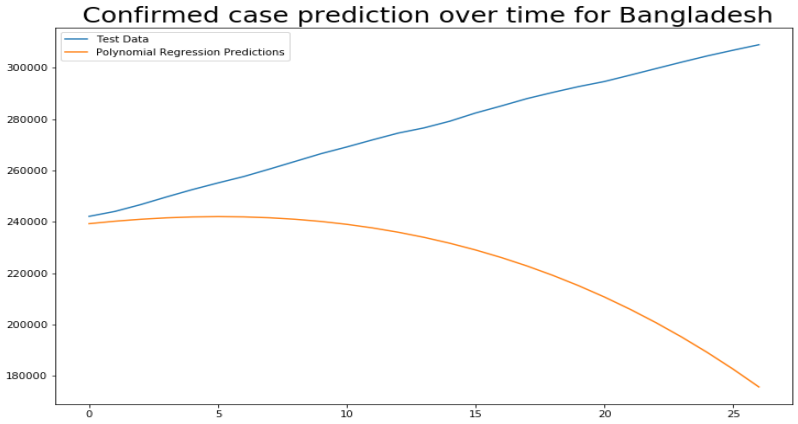

Fig. 6A: Confirmed case prediction in Bangladesh

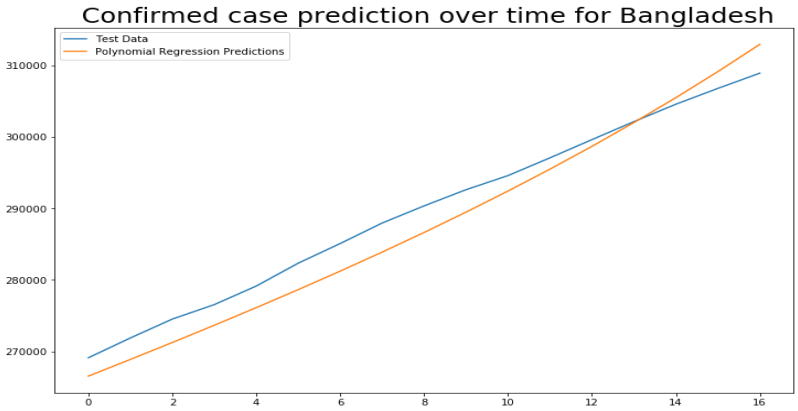

Fig. 6B: Confirmed case prediction in Bangladesh

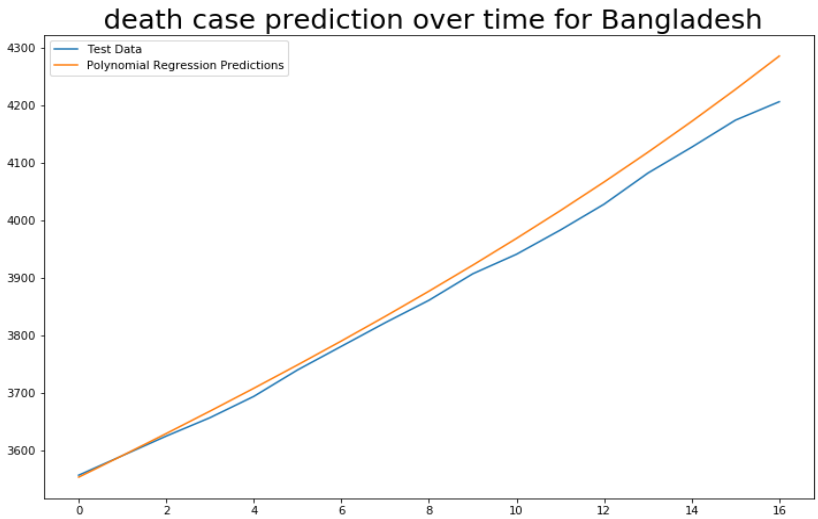

Fig. 7: Death case prediction in Bangladesh

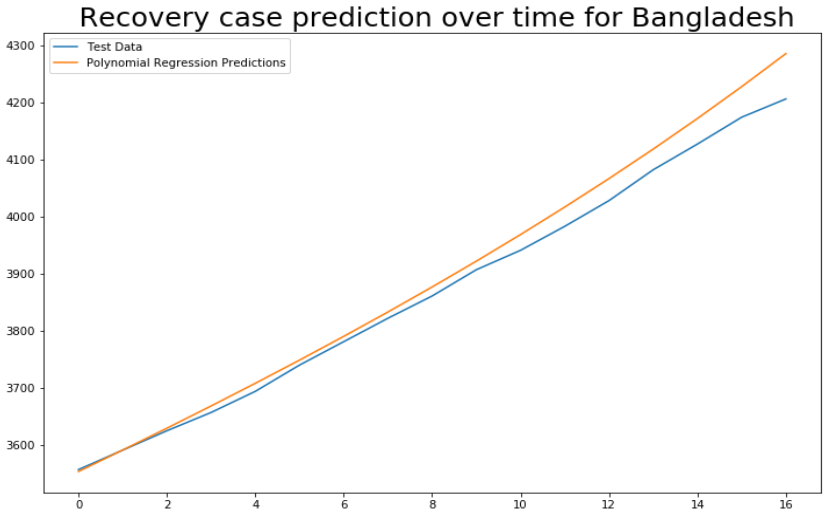

Fig.8 : Recovery case prediction in Bangladesh

Fig. 6A depicted the Confirmed case prediction from the beginning of the outbreaks. From the figure, it is clear that the prediction does not fit the actual data. After analysis, we found that earlier data showed less pattern and does not follow a linear relationship, and Fig. 6B showed the linear mapping of the predicted result with the actual result. This pattern was found after 80 days from the beginning. Fig. 7 and Fig 8 represent the death prediction and recovery prediction in Bangladesh respectively. All three predictions fit with the actual test data.

Table No. 1: Shows the Results of Regression Analysis.

PREDICTION type		MEAN SQUARED ERROR (MSE)	MEAN ABSOLUTE ERROR (MAE)	CO-EFFICIENT OF REGRESSOR
Confirmed case	Fig. 6A (Early stage)	4245703373.7582927	51898.24514820785	3401.86043 -794.35808 40.779562 -0.770038 0.005812 -0.000014
	Fig. 6B (After 80 days of begining)	8433561.65092303	2668.1668083290842	-2779683.59329 110083.223498 -1682.309721 12.303603 -0.042695 0.000057
Death case		998.4780883276758	23.9363624064375503	-31624.550968 1251.03274 -19.087174 0.139472 -0.000484 0.000001
Recovery case		998.4780883276758	23.966362406437503	-31624.550968 1251.03274 -19.087174 0.139472 -0.000484 0.00001

V. Conclusion

In this study, we performed the numerical and exploratory analysis to explore the relations of COVID-19 transmission in Bangladesh comparing with global trends and neighboring country India. We also applied linear regression to predict the infection, death, and recovery cases in Bangladesh. The experimental analysis explores that the infection and death cases had reached its maximum peak from late May to June. The infection pattern indicated that community transmission has been started in Bangladesh by May-June. From August, daily infection cases were going to down. Linear regression was used to predict the infected cases along with recovery and fatal cases.

References:

[1]. Shinde, G.R., Kalamkar, A.B., Mahalle, P.N. et al. Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art. SN COMPUT. SCI. **1,** 197 (2020). https://doi.org/10.1007/s42979-020-00209-9

[2]. World Health Organization. Coronavirus disease 2019 (COVID-19) situation report–57. Geneva, Switzerland: World Health Organization; 2020. https://www.who.int/docs/default-source/coronaviruse/situationreports/20200317-sitrep-57-covid-19.pdf?sfvrsn=a26922f2_2 2.

[3]. World Health Organization. Coronavirus disease 2019 (COVID-19) situation report–51. Geneva, Switzerland: World Health Organization; 2020. https://www.who.int/docs/default-source/coronaviruse/situationreports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10

[4]. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China [Chinese]. Chinese Center for Disease Control and Prevention Weekly 2020;41:145–51.

[5]. Shammi, M., Bodrud-Doza, M., Islam, A.R.M.T. et al. Strategic assessment of COVID-19 pandemic in Bangladesh: comparative lockdown scenario analysis, public perception, and management for sustainability. Environ Dev Sustain (2020). https://doi.org/10.1007/s10668-020-00867-y

[6]. Md. Hasinur Rahman Khan, Ahmad Hossain. COVID-19 Outbreak Situations in Bangladesh : An Empirical Analysis. **doi:** https://doi.org/10.1101/2020.04.16.20068312

[7]. Greenhalgh D, HAY G. Mathematical modeling of the spread of HIV/AIDS amongst injecting drug users. Math Med Biol J IMA. 1997;14(1):11–38

[8]. Kim D, Hong S, Choi S, Yoon T. Analysis of transmission route of MERS coronavirus using decision tree and Apriori algorithm. In: 2016 18th International conference on advanced communication technology (ICACT). 2016. (pp 559–565). IEEE.

[9]. Hu B, Gong J. Support vector machine-based classification analysis of SARS spatial distribution. In: 2010 Sixth international conference on natural computation. 2010 (vol. 2, pp. 924–927). IEEE.

[10]. Sultana N, Sharma N. Statistical models for predicting swine flu incidences in India. In: 2018 First international conference on secure cyber computing and communication (ICSCCC). 2018 (pp. 134–138). IEEE.

[11]. Amiroch S, Pradana MS, Irawan MI, Mukhlash I. Maximum likelihood method on the construction of phylogenetic tree for identification the spreading of SARS epidemic. In: 2018 International symposium on advanced intelligent informatics (SAIN) 2018. (pp 137–141). IEEE.

[12]. World Health Organization online available on https://www.who.int/emergencies/diseases/novel-coronavirus-2019/

[13]. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[14]. Fong SJ, Li G, Dey N, Crespo RG Herrera-Viedma E. Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak. arXiv preprint arXiv:2003.10776. 2020

[15]. Batista M. Estimation of the final size of the second phase of the coronavirus COVID-19 epidemic by the logistic model.

[16]. Hu Z, Ge Q, Li S, Jin L, Xiong M. Evaluating the effect of public health intervention on the global-wide spread trajectory of Covid-19. medRxiv. 2020.

[17]. Jia L, Li K, Jiang Y, Guo X. Prediction and analysis of coronavirus disease 2019. arXiv preprint https://arXiv:2003.05447. 2020.

[18]. Kumar J, Hembram KPSS. Epidemiological study of novel coronavirus (COVID-19). 2020 arXiv preprint https://arXiv:2003.11376

[19]. DeCaprio D, Gartner J, Burgess T, Kothari S, Sayed S. Building a COVID-19 vulnerability index. arXiv preprint https://arXiv:2003.07347. 2020.

[20]. David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 26. A simple regression equation has on the right-hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression e right-hand side, each with its own slope coefficient

[21]. "CSSEGISandData/COVID-19", GitHub, 2020. [Online]. Available: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series. [Accessed: 21- Oct- 2020].