# COVID-19 Pandemic Data Analysis and Prediction Using Machine-Learning Algorithms

Md Rasel Uddin [1]

***ABSTRACT: The COVID-19 pandemic is a disaster for the world. Researchers across the world are working hard to control the pandemic fully. Though the vaccine has been invented, the pandemic seems not to be in full control. So, researchers are analyzing the COVID-19 data and trying to predict the pandemic trend. In this paper, using the current COVID-19 pandemic data, we will analyze and predict the behavior of the COVID-19 pandemic using machine learning algorithms to help make decisions to control the pandemic fully. Many machine learning models have been used for prediction around the world. We have collected data on the COVID-19 pandemic from the period December 2019 to 27 August 2021. This paper has analyzed COVID-19 patient data to help to resist the potential COVID-19 infection. This study will help governments, the authority, doctors, and leaders prepare a plan to control pandemics in the future.***

***Keywords: COVID-19, Machine learning, Coronavirus, Data Analysis, Regression.***

## 1. INTRODUCTION

COVID-19 is known as coronavirus in the world. It is the biggest threat of modern times to humans. The virus was first detected in Wuhan, China in December 2019 [1]. COVID-19 is fatal, estimated at 4.5%, and it is 80% for the age group 70-79. But it is 14.8% for the age group 80 and up [1]. Elder persons with underlying diseases such as diabetes, Parkinson’s disease, and Cardiovascular disease aged greater than 50 have the highest risk to be infected by the virus. The disease is caused by SARS-CoV-2. Though the virus was first in China, now it has spread all over the world. As of August 2021, the number of confirmed cases of the virus has exceeded 214,468,601 in more than 200 countries around the world and more than 4,470,969 people have died due to this virus till August 27, 2021 [2]. The consequences of this virus are dangerous. This is disastrous for the health system also, and it is most important to predict the virus situation before taking a crucial decision. The role of scientists, and researchers are to analyze the data to understand the virus and its characteristics. This kind of research will help make the right decisions and appropriate plans.  Many previous tests have been used in research such as blood tests, CT images, socio-demographics, and more. However, the majority of the studies rely on an algorithm trained on a single prognostic outcome [3].

Some more research papers also have been published related to predicting COVID-19 using machine learning algorithms such as [4]. Besides, there are also some more papers. Another paper [5] is also very much effective where they discussed algorithms but didn’t apply them to data. The objective of this study is to analyze if it is possible to predict the overall poor prognosis for COVID-19 patients and the spread of disease using machine learning algorithms or models such as the SEIR model and Regression models. The data used in this paper has been taken from Word Health Organization’s (WHO) official website and ‘Our World in Data’ from December 2019 to 27, August 2021 [2], [6]. This data includes confirmed cases, death cases, and recovered cases from all countries. The motivation behind this research is to control and decrease the infection by controlling the virus using machine learning algorithms by analyzing and predicting the virus’s new cases, new deaths, and total cases in the period. As the COVID-19 pandemic is ongoing, the data is being updated regularly. So, data analysis and prediction of COVID-19 in this paper may vary from time to time. The limitation of

[1] Research fellow, Department of CSE, UITS
E-mail: raseluddin102@gmail.com

this paper is data is updated regularly. So we experimented with the data of the current time. This is done till 27 August 2021. The rest of the paper is designed in the following sections. In section 2, the literature review is discussed. The proposed methodology is presented in section 3. There are multiple subcategories in section 3. Section 4 presents the results and discussion and the final section 5 is the conclusion.

## 2. LITERATURE REVIEW

Machine learning is being used for many sectors of life. It is being used for predicting disease for so long. Many researchers are researching these things. Many research papers have also been published related to predicting COVID-19 since the pandemic spreads across the world. F.T. Fernandes et al [3] have discussed machine learning algorithms and predicted the COVID-19 trend based in Brazil. An India-based research paper has also been published to predict this pandemic which is based in India by R. Gupta [1] and around the world [3], [7], [8]. As well based in Bangladesh, there has also been some research paper using Machine learning algorithms. Using machine learning algorithms many research papers published since the beginning of the pandemic [9]. C. Shorten et al [5] have discussed many things on this topic. This paper is very useful for an ML algorithm. The authors have discussed many algorithms. These research papers have been done using machine learning algorithms. As Machine Learning algorithms work with data, there have some limitations related to data. This paper used data before vaccine invention. So after vaccination, the data varies. But in this paper, we are going to discuss all machine learning and deep learning algorithms which can be used to control pandemics and also predict the trends of the pandemic.

## 3. PROPOSED METHODOLOGY

In this proposed study, we have analyzed and predicted the upcoming trend of COVID-19 infections. The dataset which has been used in this study is collected from ourworldindata.org [2]. Here we have used a multiple linear regression algorithm. After collecting the dataset, first, we split our dataset into parts the training set, testing set, and validation set. Then we implemented the machine learning regression model.  We have calculated the performance of the proposed model. We have calculated the accuracy and used Root Mean Square Error (RMSE) to test our model. The proposed methodology is presented in Figure 1.

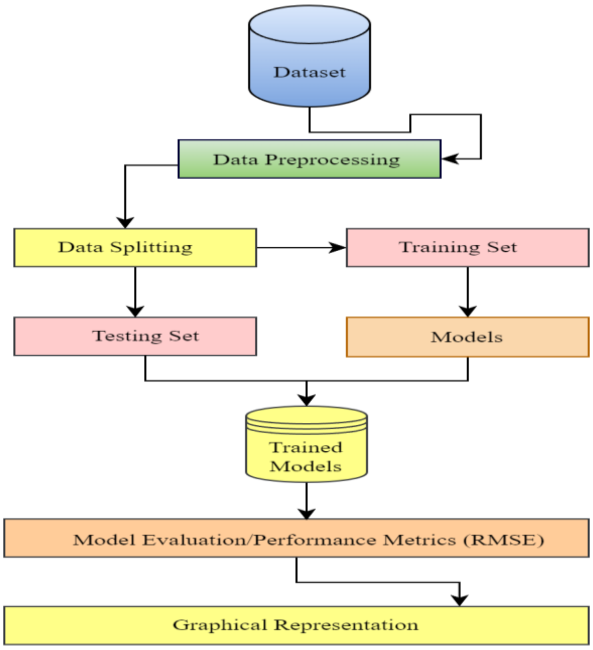


**Figure 1:** Proposed Methodology

Data collection is a mandatory step for implementing the proposed model. The analysis data is done based on the collected dataset. Then we used machine learning algorithms such as linear regression and polynomial regression to predict the pandemic. The methodology which we have used here is a suitable approach. Now we will discuss the algorithms of both machine learning and deep learning which are used to predict and control pandemics such as the SEIR model, SIR model, and NLP, and also discuss literature mining.

## *A. Machine Learning Models*

### *A.1 SEIR model*

The two most popular machine learning models for disease infection forecasting are the SEIR model and the Regression model. SEIR model is a combination of four components. They are susceptible(S), Exposed(E), Infection (I), and Recovered (R) as shown in Figure 2.

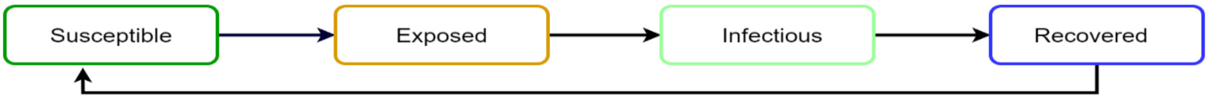


**Figure 2:** SEIR model Architecture

'S' is the fraction of susceptible individuals. It means those who can confront the disease. 'E' is the fraction of individuals who are exposed, those who have been infected but are not yet infectious. 'I' is the fraction of infective individuals (those capable of transmitting the disease). In the final word, 'R' is the fraction of recovered individuals. Those who have become immune. Let's assume that the incubation period is a random variable with exponential distribution with parameter a. And assuming the presence of vital dynamics with birth rate A equal to death rate N (so that the total number N is constant).

$$\frac{dS}{dt} = \mu N - \mu S - \frac{\beta IS}{N} \quad (1)$$

$$\frac{dE}{dt} = \frac{\beta IS}{N} - (\mu + \alpha)E \quad (2)$$

$$\frac{dI}{dx} = \alpha E - (\Upsilon + \mu)l \quad (3)$$

$$\frac{dR}{dx} = \Upsilon I - \mu R \quad (4)$$

We have S+E+I+R=N, but this is only constant because of the simplifying assumption that birth and death rates are equal in general N is a variable. For this model, the basic representation number is:

$$R_o = \frac{a}{\mu + a} \frac{\beta}{\mu + \Upsilon} \quad (5)$$

The most important part of the SEIR model is to calculate the $R_o$ Value. $R_o$ indicates the contagiousness of the disease. Simply Ro determines the number of people who can be affected by a single infected person over the course of time. If the value of $R_o < 1$, this means the spread is expected to stop. If the $R_o$ value is equal to 1, it signifies the spread is stable or endemic. If $R_o > 1$, this signifies the spread is increasing in the absence of intervention. A regression model is a form of prediction modeling technique that investigates the relationship between a dependent variable (target) and an independent variable (predictor). There are many regression models such as linear regression, polynomial regression, and more. This study has used linear and polynomial regression to predict the data.

### *A.2 SIR model*

The SIR model is one of the core models of epidemiology. It can be The SIR model can tell about the spread of infectious diseases and the population stands for Susceptible (S), Infected (I), and Recovered (R). Sometimes this model extended to the exposed population which has been discussed in detail before. The SIR model describes how a population transmission from healthy or ”Susceptible” to ”infected” and ”Recovered” through a set of three differential equations. These equations solve for the infection and recovery rates from data of initial and recovered populations. The challenge with these SIR models is that they have to limit assumptions. Many research papers published about the SIR model. The paper [5] discussed SIR models.

### *A.3 NLP and Literature Mining:*

Natural Language Processing is known as NLP. NLP applications are being used by researchers for COVID-19. NLP has been a boom of interest due to the invention of the transformer Neural Network. The advancement of NLP is the success of self-supervised pre-training and transfer learning. It would be extremely challenging to find a big dataset of question-answer pairs related to COVID-19 The research paper by C. Shorten et al [5] also discussed in depth how NLP is being used for COVID-19.
Literature datasets such as CORD-19 contain over 128000 papers that contain information about SARS-CoV-2 as well as related coronaviruses such as SARS and MERNS, information about COVID-19, and relevant papers concerning drugs [5]. Many NLP systems were developed to organize a massive scale of data. This task is not possible for a single person or a group of people. Here are some literature mining systems built from datasets such as CORD-19 and TREC-COVID. CO-SEARCH, COVIDEX, SLEDGE, and CAiRE-COVID are some popular data mining systems. These systems use a combination of information, retrieval, knowledge graph consultation, question answering, and summarization to facilitate exploration into the COVID-19 scientific literature. Many papers have also discussed literature mining such as [5].

### *A.4 Linear Regression*

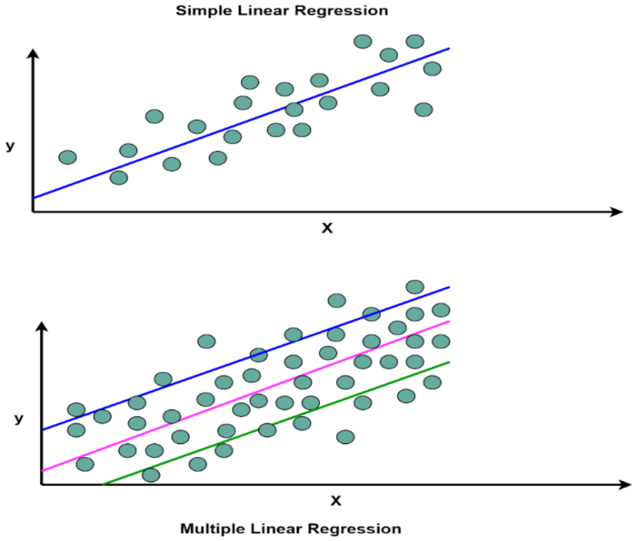


**Figure 2:** Simple and Multiple Linear Regression Structure

Linear regression can be used when we try to find the relationship between one or more predictors and targets [10]. Linear regression can be divided into two types. One is simple linear regression and another is multiple linear regression. Predicting a response using a single feature is known as simple linear regression. Assume two variables x and y which are linearly related. So we can find a linear function which will be between x and y that can predict y which is the response value as accurately as possible as a function of the x which is the independent variable or features. The equations of simple and multiple linear regression are presented in equations (6) and equation(7).

$$y = mx + c \tag{6}$$

$$y = \ m_1 x_1 + m_2 x_2 + \cdots + m_n b_n + c \tag{7}$$

Multiple Linear regression is called regression which builds a relationship between two or more features and a response by befitting to the observed data. It is the extension of simple linear Polynomial regression, which is slightly different from linear regression. In polynomial regression, we fit the polynomial equation on the data [11]. We also fit data with a curvilinear relationship between the independent and target variable. The value of the target variable changes in a non-uniform manner concerning the predictor in a curvilinear relationship.

### *B. preparing dataset*

Preparing datasets is the first and fundamental part of implementing models. To implement machine learning models, preparing datasets is most important. It provides the best performance. In this part, we have to find the null values by checking data qualities, data cleaning of the dataset as well as feature engineering next. The dataset we are using here is a preprocessed dataset collected from and WHO website and our word in data, from their official repository. They have collected data from across the world. This is a secondary dataset.

### *C. Feature Extraction*

In the dataset, every feature is not important for predicting or analyzing the dataset. We can do a reduction of dimensionality according to our model. Every feature is not useful for an independent variable. If all feature is used in the model, it will produce a bad score for the model and also impact accuracy. So feature extraction is much more important. In our dataset, there have lots of features such as date, new cases, new deaths, origin, and location. All of them are not useful for the model. So we have done feature engineering for our model for the best performance.

### *D. Root Mean Square Error*

Root Mean Square Error (RMSE) is one of the best performance measurement metrics. To find out or measure the error of any machine learning model, RMSE is standard for predicting quantitative data. RMSE is defined as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_{i-} y_i)^2}{n}} \tag{8}$$

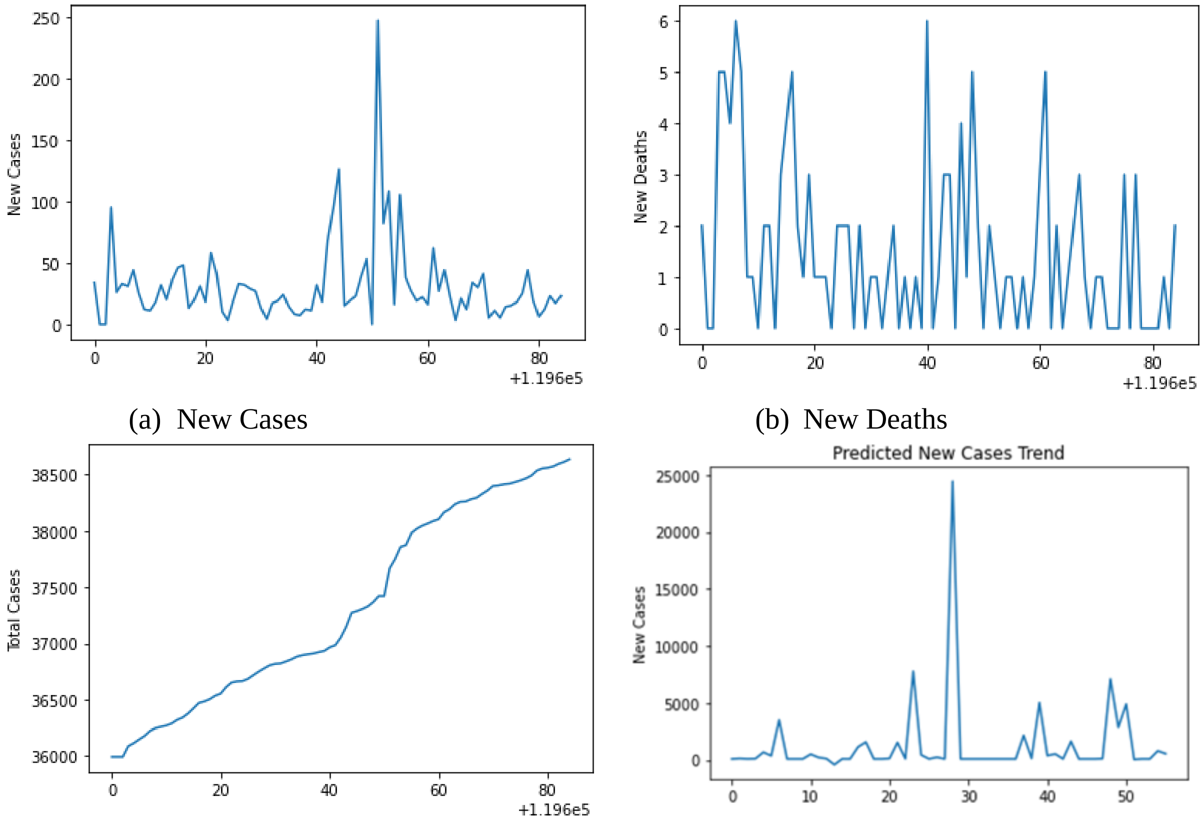
Here $\hat{y}_i$ is the predicted values, $y_i$ is the observed values and n is the number of observations. In our model, RMSE has been used for measuring the error of the model. We have also measured more tests, and we have done the accuracy score of our model for both linear regression and polynomial regression.

### *E. Accuracy*

Machine learning model accuracy is different from the usual method for finding accuracy. In machine learning, accuracy is found by identifying relationships and patterns between variables in a dataset. This happens based on training data or input. There are some metrics by which we can easily measure the accuracy of a model and can find out the best model. Using some matrices, we can easily find the accuracy of our model also. Some matrices are Confusion Matrix, Classification Accuracy, Logarithmic Loss, Area in the der curve, F1 Score, Mean Absolute Error, and Mean Squared Error. In this paper, we will use Root Mean Square Error (RMSE) to find out the accuracy of our model.

## 4. RESULT AND DISCUSSION

The first COVID-19 was detected in December 2019. Since then, the pandemic is still ongoing. Though vaccination is provided, the pandemic seems still dangerous. This analysis is done based on data from December 2019 to August 27, 2021. Data were collected from international sources to predict the outbreak. The graphs in Figure 3 show the fluctuations of new cases, new deaths, total confirmed cases, and predicted new cases from the period December 2019 to May 26, 2021. After the pandemic outbreak, many numbers of people have been infected with COVID-19 daily. New cases of COVID-19 were very high at the beginning. The new trend of COVID-19 can be seen in this graph. These graphs are generated by implementing the Python code. Libraries like Matplotlib have been used in codes to simulate the data of the COVID-19 pandemic. Matplotlib library is a library that is used for creating interactive and animated visualizations. The library is written in Python.



(a) New Cases

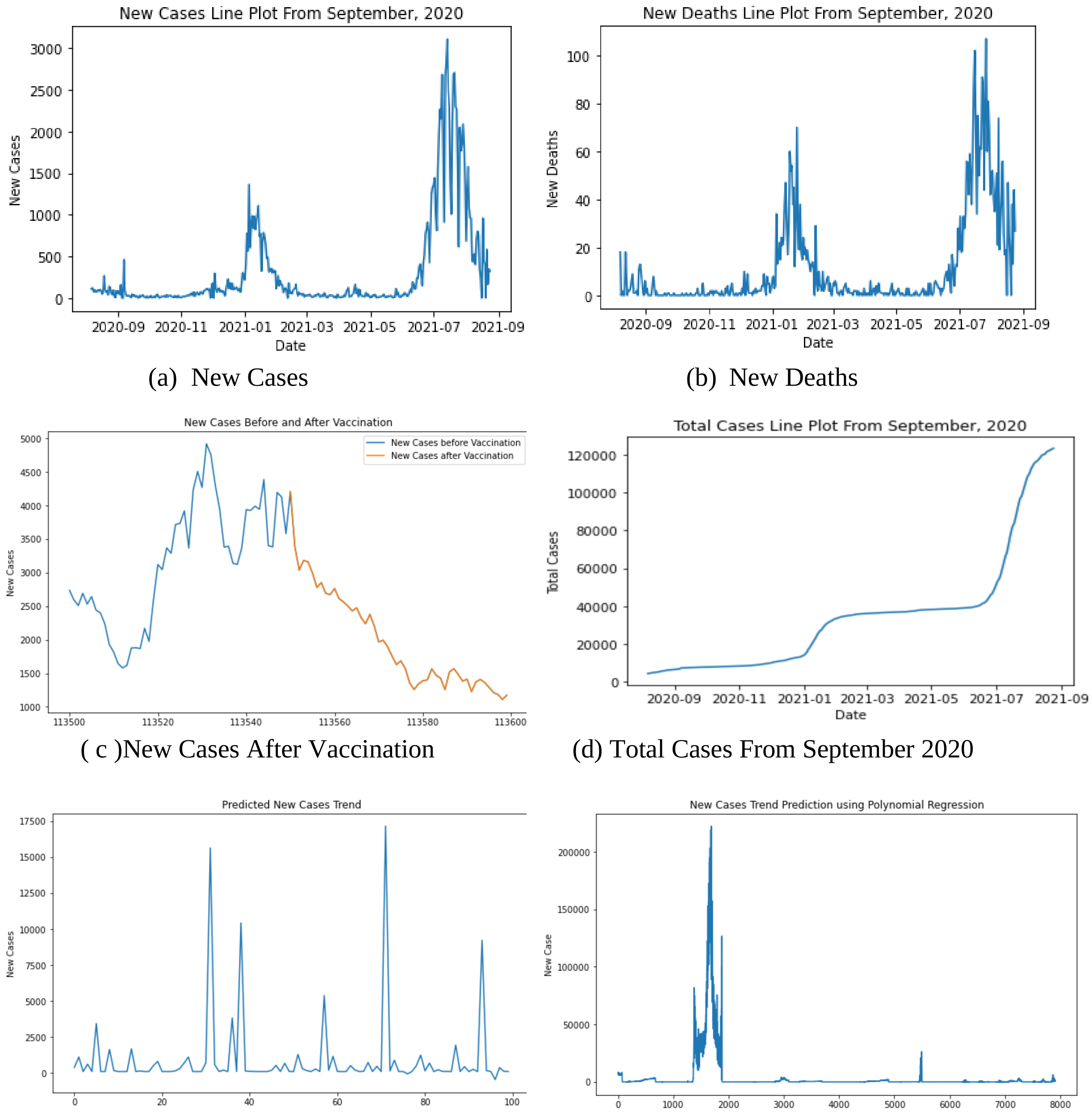
(b) New Deaths

(c) Total Cases

(d) Predicted New Cases

**Figure 3:** COVID-19 Cases from December 2019 to May 26, 2021

It is shown that the number of new cases is downwards through the middle of time it was upward in Figure 3(a). The number of deaths was always high in Figure 3(b). Many infected people were so vulnerable to death due to COVID-19. To get an idea about the trend of COVID-19 deaths nowadays, this plotted graph in Figure 3(b) shows that the number of COVID-19 death cases was pretty increasing at first, though it is decreasing day by day. Nowadays, it is downwards. The total number of COVID-19 cases around the world is around 169 million till May 26, 2021, in Figure 3(c). The number of new cases was too high in the second wave. Though it is seen now through this graph that the total number of cases is not increasing too fast now. Based on these data from the period December 2019 to May 26, 2021, it is predicted the trend of new cases of COVID-19 in Figure 3(d). Using linear regression, we have predicted the trend of the new cases of COVID-19 day by day based on data on total cases, new deaths, and deaths. In Figure 3(d), it is shown that the new cases were downwards. This prediction shows that the number of COVID-19 new cases is decreasing gradually. This analysis was done before the acceleration of vaccination.



(a) New Cases

(b) New Deaths

( c )New Cases After Vaccination

(d) Total Cases From September 2020

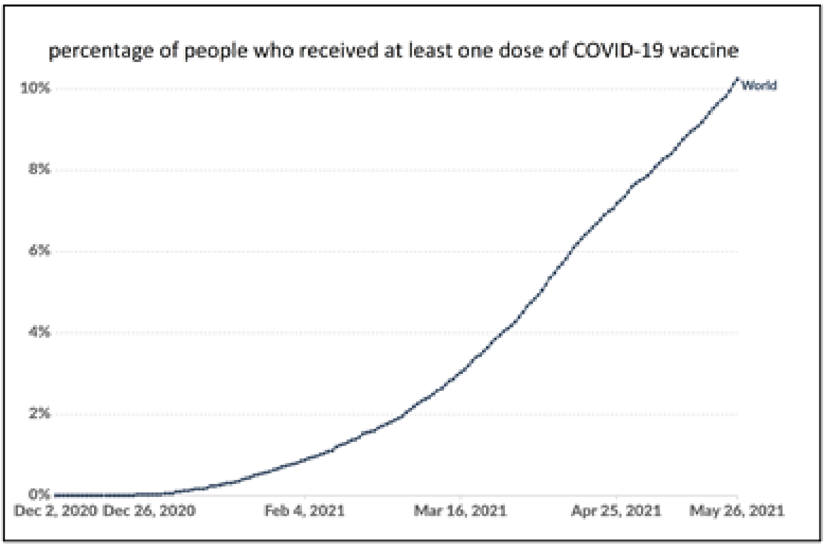
(e) Predicted New Cases Using Linear Regression

(f) Predicted New Cases Using Polynomial Regression

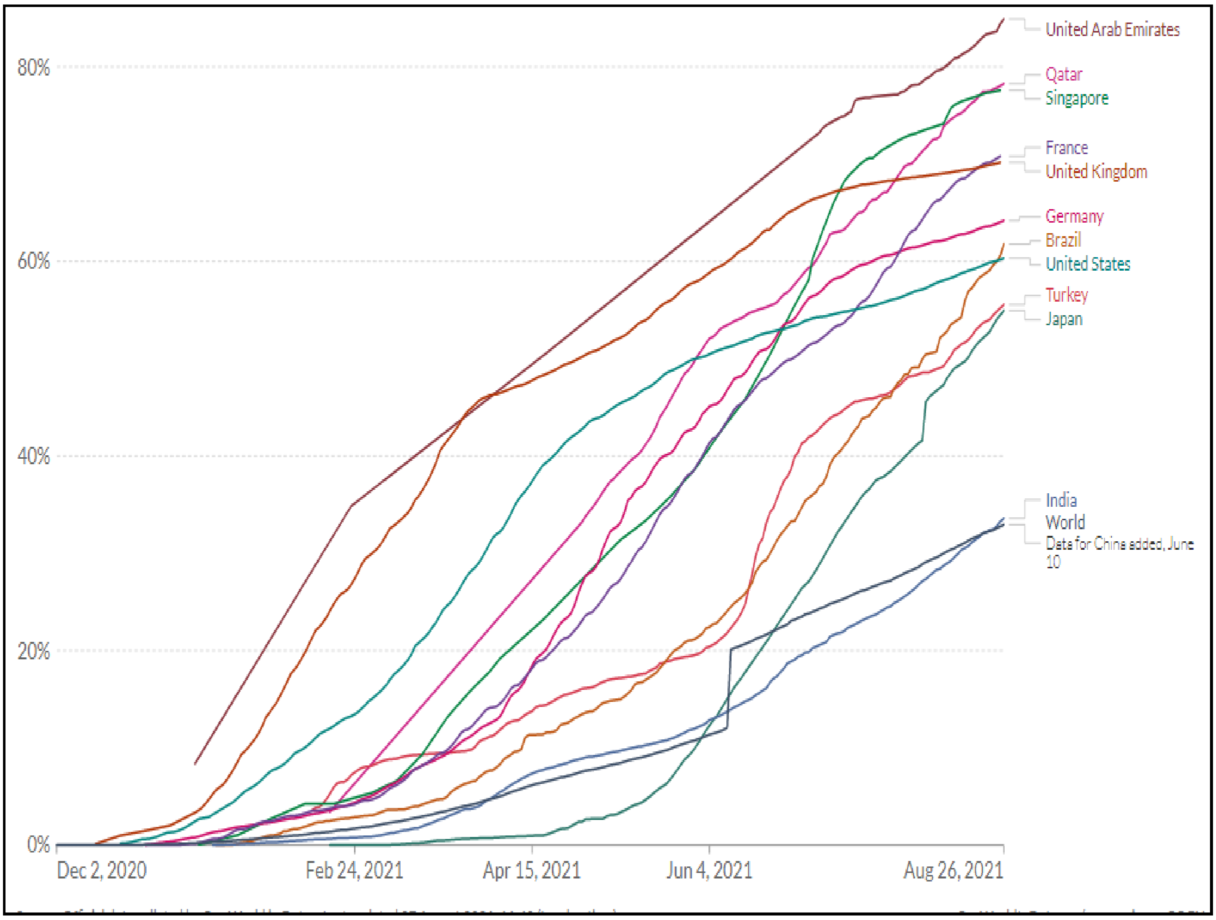
**Figure 4:** COVID-19 Cases from May 26, 2021, to August 27, 2021

In Figure 4, The COVID-19 cases were analyzed by the collected data from the period May 26, 2021, to August 27, 2021. As the world is under vaccination, the number of cases is decreasing. Though the number of vaccinated people around the world was very low during that period, it increased later. Only a few countries vaccinated more than 50% such as Canada, the United Kingdom om, and Mongolia then. According to the graph, the total number of vaccinated people around the world is

10% till May 26, 2021 [2]. From May 26, 2021, to August 27, 2021, the world has given importance to vaccination. Every country is providing vaccines to its citizens to control the pandemic. If we analyze the data of the new cases, new deaths, and total cases of COVID-19 from May 2021 till now August 2021, we can see the nowadays pandemic situation. We a make comparison of new cases, new deaths, and total cases of whether the vaccine is effective to control pandemics or not. Figure 4(a) is the graph of new cases from September 2020 to August 2021. At the beginning of the pandemic, the cases were fluctuating and the number of new cases was pretty high in the mid of 2021. Then it goes downwards. The number of new deaths along with new cases was also high in Figure 4(b). The number of new deaths was compared daring with new cases. Besides the total number of cases is also increasing slowly. Figure 4(c), represents a comparison how was the number of new cases before and after vaccination. It has been shown that the number of new cases is decreasing after the vaccination. The models which have been used here are the Linear Regression model and the Polynomial regression model. To predict the COVID-19 trend, these models have been used. After using these models, it is shown that the pandemic is downward mostly after the efforts of vaccination. In figure 4( e ), the plot is the prediction of new cases using linear regression and polynomial regression has been used in figure 4(f) to predict the new cases of the pandemic. According to the graph, the number of new cases is decreasing and will be in control. The variance score of the linear regression model is above 72% and polynomial regression is more than 69%.



**(a)** Percentage of people who received one dose of the COVID-19 Vaccine



**(b)** Total number of people who received at least one dose of COVID-19 vaccine per country

**Figure 5:** Vaccination till August 27, 2021

As the pandemic is seemed more controllable after vaccination, the vaccination effort should speed up. As of May 2021, more than 10% of the world population has given the vaccine at least one dose which has been shown in Figure 5. As of August 27, 2021, the number of people who received at least one dose of the COVID-19 vaccine around is 4 953 887 422 in Figure 5(a). Figure 5(c) shows that UAE has given more vaccines to its population till August 27, 2021. Qatar, Singapore, Prance, UK has also given the vaccine to more than 60% of their population [6]. The United States has given vaccines to 60% of the equal population whereas India has given them to more than 20% of its population. Bangladesh has given vaccines to more than 20 million of its population in both doses on August 27, 2021.

## 5. CONCLUSION

Based on the above discussion, it can be concluded that the pandemic is going under more control as the number of both new cases and the death rate is decreasing. So vaccination should be broadened around the world as well as the importance of hygienic life should be promoted. As is shown that the number of new COVID-19 cases is decreasing with the increasing ratio of vaccination. Many developed countries have taken too much effort into the vaccination step. Now the number of new cases is very low in those countries. Besides seeing the graph, it can be said that the pandemic can be controlled through the broad step of vaccination. Though day by day we also see that the COVID-19 virus changes its pattern. It shows some variants. Then the effectiveness of the vaccine becomes questionable though scientists assure that some of its vaccines are effective in its new variants. World most country is facing COVID-19 new variants. But these variant is less dangerous to those who have taken the vaccine. The analysis above which was done using data from World Health Organization shows the pandemic seems in control. The machine learning models are showing that the number of new cases is decreasing. But as this virus change its pattern, its new variant can increase the cases. So to control the virus, everybody should be cautious and should take the vaccine.

## REFERENCES

[1] Gupta R, Pandey G, Chaudhary P, and Pal S K,(2020), Machine Learning Models for Government to Predict COVID-19 Outbreak, Digit. Gov. Res. Pract., vol. 1, no. 4, pp. 1–6, 2020.

[2] W. H. O., (2021), WHO Coronavirus (COVID-19) Dashboard, 2021. [Online]

[3] Fernandes F T, Oliveira T A de, Teixeira C E, Batista A F d M, Dalla Costa G, and Chiavegatto Filho A D P, (2021), A multipurpose machine learning approach to predict COVID-19 negative prognosis in Sao Paulo, Brazil, ˜ Sci. Rep., vol. 11, no. 1, pp. 1–7, 2021. [Online].

[4] Leon M I, Iqbal M I, Azim S M, Mamun K A,(2020), " Analysing and Predicting Coronavirus Infections and Deaths in Bangladesh Using Machine Learning Algorithms" pp. 1–10, 2020.

[5] Shorten C, Khoshgoftaar T M, and Furht B, (2021), Deep Learning applications for COVID-19, J. Big Data, vol. 8, no. 1, 2021. [Online].

[6] O. W. in Data., (2021), Coronavirus Pandemic (COVID19) – the data., 2021, 2021. [Online].

[7] Zoabi Y, Deri-Rozov S, and Shomron N,(2021), Machine learning-based prediction of COVID-19 diagnosis based on symptoms, npj Digit. Med., vol. 4, no. 1, pp. 1–5, 2021. [Online].

[8] Carcione J M, Santos J E, Bagaini C, and Ba J,(2020), A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model, Front. Public Heal., vol. 8, no. May, 2020.

[9] Islam M N, Inan T T, Rafi S, Akter S S, Sarker I H, and Islam A K M N, (2020), A systematic review on the use of ai and ml for fighting the covid-19 pandemic, IEEE Transactions on Artificial Intelligence, vol. 1, no. 3, pp. 258–270, 2020.

[10] Su X, Yan X, and Tsai C L, (2012), Linear regression, Wiley Interdisciplinary Reviews: Computational Statistics, vol. 4, no. 3, pp. 275–294, 2012.

[11] Cheng C L and Schneeweiss H, (1998), Polynomial regression with errors in the variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 60, no. 1, pp. 189–199, 1998.

**Md Rasel Uddin** is an undergraduate student in the department of computer science and engineering of the University of Information Technology and Sciences shortly known as UITS. He was born on January 1st, 1997 in a village named Choto Jaynagar, under the Sadar South Upazila of Cumilla district in Bangladesh. He has completed his primary and higher secondary education at Kamalpur High School and Cumilla Cambrian College. Now he is completing his graduation in Computer Science and Engineering. He is a final year student of 4 years long bachelor of science(B.Sc) degree.

He is a young researcher in the field of Machine Learning, Deep Learning, Computer Vision, and Natural Language Processing. His main job is to implement artificial intelligence techniques to solve real-life problems. Currently, he is a STUDENT. Previously he was a web developer. Besides he is a programmer who loves to solve problems on various online platforms. Gradually he become interested in research. In the beginning, his research interest was only machine learning, nowadays he is working with Deep Learning and Natural Language Processing also.

Mr. Uddin is a research enthusiast as well as a social concern person. He is promoting research among young people and students. He is a member of multiple research society that promotes research. He is a campus ambassador of a research platform also. He is a senior executive of his university computer club whose primary task is to encourage students to research.