Prediction of number of fatalities due to Covid-19 using Machine Learning

Manpinder Singh
CSE,DIT University
Dehradun,India
manpindersingh417@gmail.com

Saiba Dalmia
CSE, DIT University
Dehradun,India
saibadalmia2513@gmail.com

2020 IEEE 17th India Council International Conference (INDICON) | 978-1-7281-6916-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/INDICON49873.2020.9342390

***Abstract*—Throughout our history, pandemics have severely devastated civilization, even resulted in their end. The latest pandemic to cause global outbreak is SARS-CoV-2(Covid-19). It originated in the city of Wuhan, China in December,2019. It is transmitted to humans mostly due to contact with an infected animal (the most common carriers are bats). India had its first case of Covid-19 on 30 January 2020. If we could predict the number of deaths at an early stage then millions of lives could be saved. The government would be able to forecast the infection rate and the spread of the pandemic which would result in better planning of resources, a better prepared response and improved medical facilities. Machine learning plays a really important role in such pandemic situation like helping in making new vaccines or suggest ongoing medicines. Inspired by this, we propose a Machine Learning model which can predict future number of deaths in India.**

Keywords—Linear Regression, Polynomial Regression, Coronavirus, COVID-19, Machine Learning, Web Scraping

I. INTRODUCTION

Throughout our history, pandemics have severely devastated civilization, even resulted in their end. They have changed the course of history many times. When large groups of people started living close at hand to each other and with animals, it became fertile grounds for diseases as humans failed to have proper sanitation and nutrition. Far-flung trade sped up these pandemics. Progression of civilization, development in infrastructure, urbanization are few of many driving forces for the spread of pandemic. The first epidemics to appear in early times were Malaria, tuberculosis, smallpox, leprosy, influenza, etc. Antonine Plague (165-180 AD) also known as the plague of Galen was of the first epidemics to ravaged human population. It resulted in deaths of five million people. There were many pandemics to follow that shook the human civilization, like Black Death (Bubonic Plague) that occurred between 1347-1351 AD. It killed 200 million people, wiping out 30-50% of Europe's population. It originated in rats and extended to human beings via infected fleas. Fig. 1 represents the number of deaths in various pandemics throughout the history. Smallpox(1520 AD)-killed 56 million people, Spanish Flu(1918-1919)-killed 40-50 million people ,Plague of Justinian(541-542 AD)-killed 30-50 million people just to name a few[1], [2].

The latest pandemic to cause global outbreak is SARS-CoV-2(Covid-19). In December 2019, a type of Coronavirus, never seen before in humans began to spread in the city of Wuhan, China. Coronaviruses are a family of viruses that affects the vital respiratory organs of mammals.

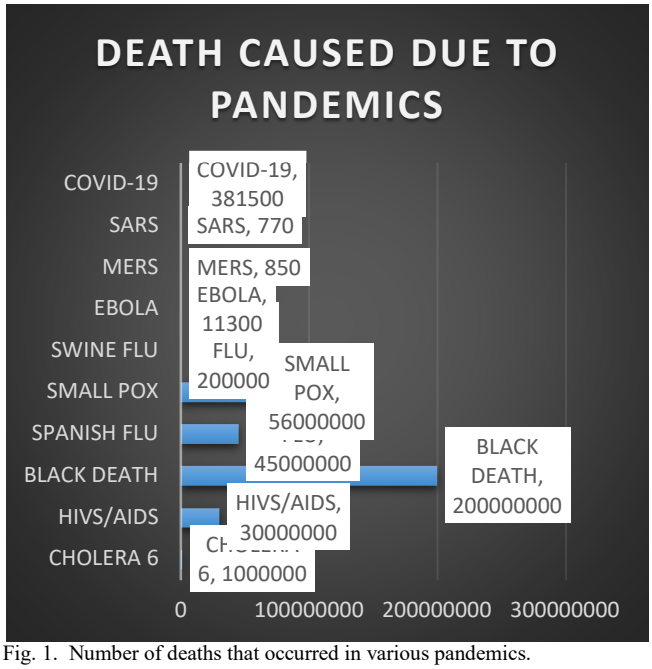

Fig. 1. Number of deaths that occurred in various pandemics.

This is not the first Coronavirus that caused global outbreak. The first was SARS (Severe Acute Respiratory Syndrome) Coronavirus spread in 2002 in China and the second was MERS (Middle East Respiratory Syndrome) Coronavirus spread in 2012 in Saudi Arabia. It is transmitted to humans mostly due to contact with an infected animal. Bats are among the most carrier of coronavirus. It does not transmit directly to humans, instead it might occur via an intermediary animal. SARS-CoV-2 is found identical to 2 bat coronaviruses: bat-SL-CoVZC45 and bat-SL-CoVZXC21. Its genomic sequence is 88% similar to them.

India had its first case of Covid-19 on 30 January 2020.As of 14th June 2020,the total number of confirmed cases are 3,21,000 cases while the number of recovered patients are 1,62,000 and number of deaths in the country are 9,195.As we have seen ,hundreds of millions people died all over the world due to such contagious pandemics[3]. If we could predict the number of deaths at an early stage then the whole world scenario would be different. The government officials and the concerned authorities would be able to forecast the infection rate and the spread of the pandemic which would result in better planning of resources and a better prepared response. It would help them in their decision-making process.

Machine learning plays a really important role in such pandemic situation. Machine Learning has helped in predicting the structure of proteins and their interaction with

other chemical compounds in order to make new medicines or vaccine for Coronavirus or suggest ongoing medicine. It has helped in predicting the rate at which the coronavirus spread thus helping hospitals, health officials as well government to get prepared with resources and planning. It has helped in screening process thus making the screening of patients faster. As we know that these days social media are rapidly growing in popularity all over the world. Here machine learning is helpful for extracting the useful and relevant information about the disease. Thus, helping the general public to know about the true symptoms and details about the disease.

Inspired by this, we propose a Machine Learning model which can predict future number of deaths in India if the number of confirmed cases reaches a certain number. The government and the concerned authorities can decide better when to provide relaxations & when to impose stricter rules like a nation-wide lockdown. As the spread of Covid-19 is becoming harder and harder to contain, by knowing the number of fatalities in the future, the government could better strategize and formulate measures to improve the medical facilities. It will also help the government to better prepare the budget to be spent on medical equipment like ventilators, kits, masks, protection gear for essential workers etc.

II. Related Work

A Machine Learning model has proven to be valuable in the research of current antiviral that will help the patient to get recovered from Covid -19 They can use the model with the RNA sequences of coronavirus together with chemical compounds to predict the best medicine that a patient can have for coronavirus. They have used Densely Connected Convolutional Networks but have used fully connected layer instead of convolution[4]. A model is developed which considers all the commercially available medicine. The molecular transformer drug target network was trained on Smile Dataset in which molecules are presented in a text that can be encoded and decoded. This model predicts the binding affinity values between the target proteins and antiviral medicine. It was observed that "the 2019-nCoV 3C-like proteinase was predicted to bind with atazanavir". Atazanavir is an antiviral drug which is used to treat HIVs/Aids[5]. Another model was used to predict whether the patient having Coronavirus will survive or not on the basis of age and other factors. XG Booster Classifier was used in this model. It will be useful to detect which patients should keep themselves in isolation the most[6]. Machine Learning has greatly benefited the radiologist by decreasing their reading time by 65% and optimizing their workflow. Therefore they can examine more CT Scans in the same amount of time[7].

III. Detailed Description

A. Machine Learning

Human being tends to learn from their past experiences but a computer cannot. Hence, this is what machine learning does! Machine learning comes under the field of Artificial Intelligence that helps the system and gives them the ability to learn from their past experiences and improve without any human intervention. In simple words, we can say that:

Machine learning is like gardening. Seeds are the algorithms, nutrients is the data, the gardener is you and plants is the programs[8].Its primary aim is to develop computer programs. Machine Learning covers numerous fields like medical, military, marketing, security etc.[9]

B. Linear Regression

Linear regression is a machine learning algorithm. It helps in finding a linear relationship between two continuous variables and in forecasting. It is based on supervised learning. It predicts a dependent variable (predictor) value based on a given dependent variable (response). It does not look for a deterministic relationship but for a statistical relationship . A statistical relationship is one in which one variable cannot be accurately expressed by the other. For example, the relationship between height and weight[10].
The main idea behind Linear Regression is to obtain a line that best fits our data.

The best line is the line for which the total prediction error for all the points is as minimal as possible. The Euclidean distance between the points to the regression line is considered as error[11].

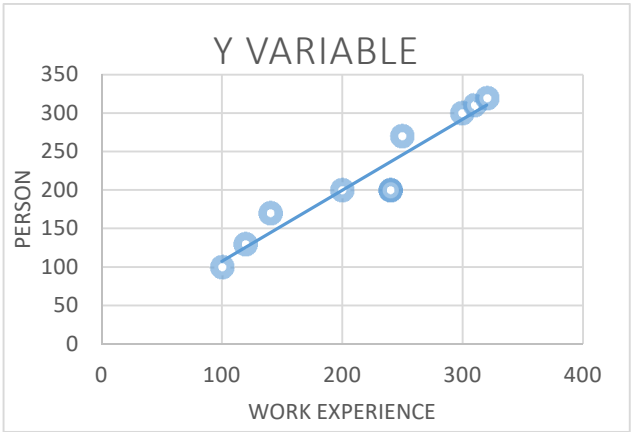

Fig. 2. Linear Regression of Work Experience versus Person.

In the Fig. 2, X (input) represents the work experience and Y (output) of a person. The regression line is the best line that fits our model and having minimal total prediction error.

1) Hypothesis function for Linear Regression:

It is calculated while training the model to obtain the regression line that best fits our data. Here, X is the input training data, Y is the target variable, θ_1 is the intercept and θ_2 is the coefficient of X. We update θ_1 and θ_2 until we find the best values for them. Once we find the best θ_1 and θ_2 values, we get the best fit line[11].

$$Y=\theta_1+\theta_2.X \quad (1)$$

2) Cost Function (J):

Root Mean Squared Error (RMSE) is the cost function(J) of Linear Regression. It is calculated between the predicted value and the true value of Y. Regression model always aims to minimize the cost function in order to get the best fit line that can predict the value of Y for a given value of X [11].

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n}(\text{pred}_i - y_i)^2 \quad (2)$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(\text{pred}_i - y_i)^2 \quad (3)$$

C. *Polynomial Regression*

It is a regression algorithm where we fit the polynomial equation on the data with a curvilinear relationship between the target variable and independent variable it is called special case of Multiple Linear Regression as we add some polynomial terms in a multiple linear regression to make it polynomial[12][13].

Equation of multiple linear regression:

$$Y = b_0 + b_1x + b_2x_2 + b_3x_3 + ... + b_nx_n \quad (4)$$

Equation of Polynomial regression:

$$Y = b_0 + b_1x + b_2x^2 + b_3x^3 + ... + b_nx^n \quad (5)$$

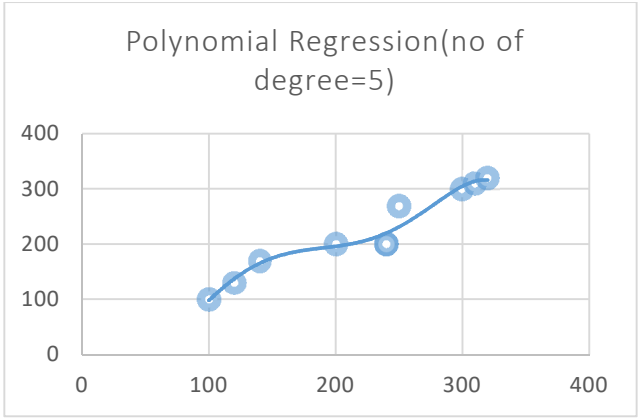

Fig. 3. Polynomial Regression of degree 5.

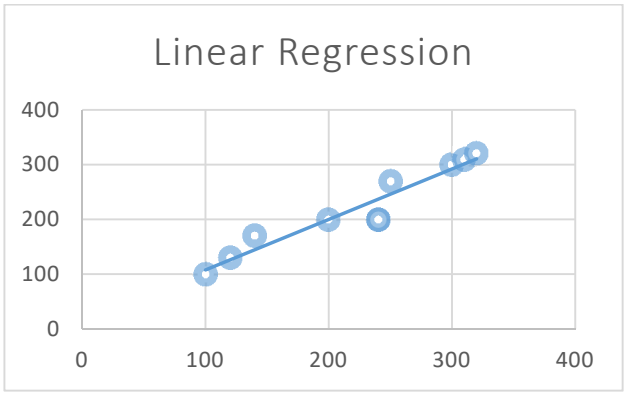

Fig. 4. Linear Regression.

For the non-linear dataset like in Fig. 3 and Fig. 4 ,the accuracy predicted by the simple linear model is less as compared to the accuracy predicted by the Polynomial model as the loss function and error rate in a simple linear model is high. So Polynomial Regression is a linear model with some changes which helps to increase the accuracy for non-linear and complicated datasets.

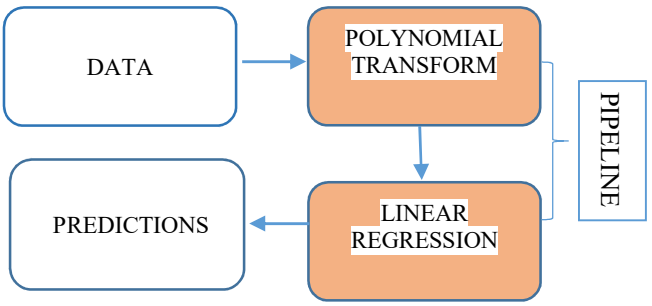

Fig. 5. Pictorial representation of implementation of Polynomial Regression.

In polynomial regression, first, we transform the data into polynomial using PolynomialFeatures function and then use linear regression to fit the parameter. Fig. 5 gives the pictorial representation of this. The polynomial features transform the equation into n^{th} (number of degrees) equation. So we have to choose it precisely because if we choose the degree of polynomial less it won't fit the data and if we choose it high it would overfit the data[12].

D. *Root Mean Square Error*

Root Mean Square Error (RMSE) is the square root of the average of squared differences between predicted and actual values of y. The formula of RMSE is:

$$\text{RMSE} = \sqrt{\overline{(F - O)^2}} \quad (6)$$

where F is the predicted value and O is the observed value. The bar above (F-O) represents the mean (similar to $\bar{x}$).

This formula can be represented slightly differently as (Barnston, 1992):

$$\text{RMSE}_{FO} = [\sum_{i=1}^{N}(Z_{Fi} - Z_{Oi})^2/N]^{1/2} \quad (7)$$

where Σ is summation , $(Z_{Fi} - Z_{Oi})^2$ = square of the difference, N is the sample size[14].

It is the measure of how spread out the residuals are. Residuals are a measure of how far the data points are from the regression lines. It can range from 0 to ∞ and does not depend on the direction of errors. RMSE tells us how concentrated or spreads out the data is around the regression line. The lower the value of RMSE, the better is the model. RMSE is more useful when particularly large errors are undesirable because it gives relatively higher weight to large errors since errors are squared before they are averaged. Use of RMSE can be commonly found in climatology, regression analysis to verify experimental results and forecasting[15].

E. *R-Square*

In simple words, it represents how well the data fit in the regression model. R-squared tells the scatter of the data points around the fitted regression line. It is also called the coefficient of determination, or the coefficient of multiple determination for multiple regression [16].

$$R^2 = \frac{Variance\ explained\ by\ the\ model}{Total\ Variance} \quad (8)$$

R-square values vary from 0 to 1 or from 0% to 100%. 0% represents that the model does not have any variation in the y-variable around its mean. 100% represents that the model has variation in the y-variable around it means.

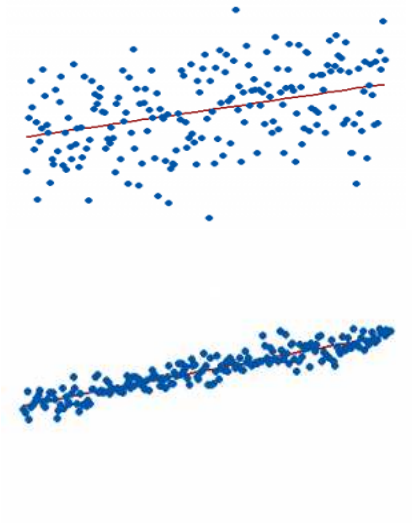

Fig. 6. Comparison of R^2 score between two graphs.

In Fig. 6, we can see that the graph on the left has R^2 score of 15%, whereas on the right-hand side the score is 85%. So, we can conclude that for the same dataset, higher the value of R^2 score denotes smaller the difference between the observed data values and the fitted values. Generally, higher R^2 score fits the graph properly.[17]

F. Experimental Data

In this study, the data were collected from "Ourworldindata.org" [18] through web scraping. Our dataset consists of 33 columns from which we chosen 3 columns location (It includes the name of the countries, islands, etc.), Total cases and Total Deaths. Our dataset contains information about the total cases and total deaths of 211 locations from 31st December,2019 to the present date. Our dataset is updated every day. We have considered the data for our model from the date where India had its first Covid-19 case. We have normalized our data as part of data preprocessing by dividing each value of total cases and total deaths with the maximum values present in the total cases and total deaths columns respectively.

G. Model Description

In this study, both Linear Regression model and Polynomial Regression model were created. Out of which Polynomial Regression model was chosen.

Since the nature of the dataset used is non-linear, the accuracy predicted by the simple linear model was less as compared to the accuracy predicted by the Polynomial model as the loss function and error rate in the Simple linear model is high.

Data Pre-processing

Only the data of India were considered for our model. The total number of cases column was chosen as a feature while the total number of deaths is our target column. We have considered all the data for our model from the date where India had its first Covid-19 case. We have normalized our data by dividing each value of total cases and total deaths with the maximum values present in the total cases and total deaths columns respectively.

1) Model Training

The dataset is randomly divided into two independent datasets, that is, 67% for training and 33% for testing. For the Linear Regression model, the fit_intercept, normalize and n_jobs parameters are experimentally set to True, False and None respectively. For Polynomial Regression model, degree, interaction_only and include_bias parameters are experimentally set to 3, False and True respectively.

2) Model Prediction

The Polynomial Model was chosen for predicting the number of deaths for the given number of confirmed cases in India. The predicted and actual value of deaths were used to calculate standard performance metrics for a regression model.

IV. Implementation

A. Experimental setup

We have used Python Programming Language to train the proposed Regression model. All experiments were performed on Jupyter Notebook 5.7.8 with Windows 10, version 1909 using Intel HD Graphics 5500 and 8GB of RAM. For the Linear Regression model, the fit_intercept, normalize and n_jobs parameters were experimentally set to True, False and None respectively. For Polynomial Regression model, degree, interaction_only and include_bias parameters were experimentally set to 3, False and True respectively. The dataset is randomly divided into two independent datasets, that is, 67% for training and 33% for testing using train_test_split() which was imported from sklearn.model_selection. TABLE I. and TABLE II. gives the Dataset description and the libraries used for our Machine Learning model.

TABLE I. Dataset Description (as of 14th June, 2020)

DESCRIPTION	VALUE
Number of features	1
Number of target variable	20584
Number of null values in Instances	0

TABLE II. LIBRARIES USED

LIBRARIES USED
NUMPY
PANDAS
MATPLOTLIB
SKLEARN
BEAUTIFUL SOUP
REQUESTS

B. *Performance Metrics*

RMSE (Root Mean Square Error) and R^2 score were used as performance metrics for our regression models (Linear Regression and Polynomial Regression models) TABLE III. shows the values of accuracy, RMSE and R^2 score for both the models.

TABLE III. PERFORMANCE MEASURES OF MODELS

MODELS	PERFORMANCE METRICS		
	ACCURACY	*RMSE*	*R^2*
LINEAR REGRESSION	0.9979	0.000124	0.9979
POLYNOMIAL REGRESSION	0.9998	0.000010	0.9998

V. CONCLUSION

India had its first case of Covid-19 on 30 January 2020.As of 14th June 2020, the total number of confirmed cases are 321,000 cases while the number of recovered patients are 162,000 and number of deaths in the country are 9,195. We have developed a Machine Learning model which can predict the future number of deaths in India for a given number of Cases. The data for our Machine Learning model were collected from "Ourworldindata.org" [18] through web scraping. It is updated daily. Both Linear Regression model and Polynomial Regression model were created. Out of which Polynomial Regression model was chosen due to its higher accuracy for non-linear dataset. The dataset is randomly divided into two independent datasets, that is, 67% for training and 33% for testing. For testing the model, we have developed an API which gives the number of deaths in India for an input number of confirmed cases (as shown in Fig. 9). Fig. 7 shows that our API gives live updates of the number of confirmed cases, recovered cases and death and also shows the graph of total death versus total cases of six countries (India, USA, China, Canada, Russia and Italy) as shown in Fig. 8.

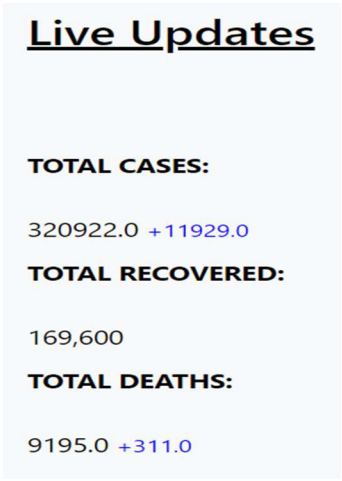

Fig. 7. Live updates of Covid-19 cases in India.

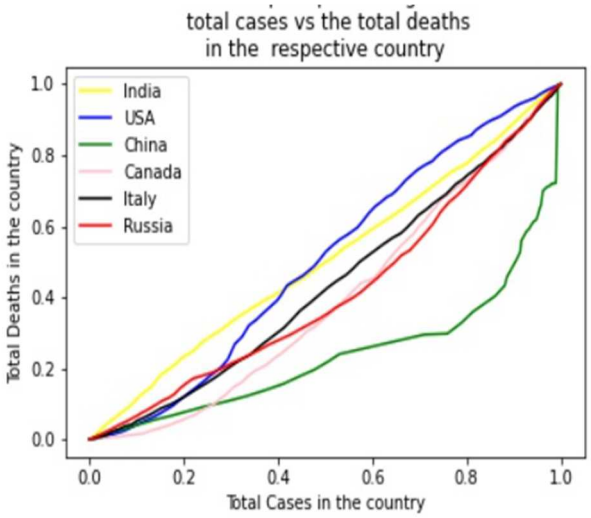

Fig. 8. Total deaths versus Total cases graph of various countries.

Polynomial Regression model was chosen for prediction and it gave an accuracy of 99.98%. It gave exceptional values for performance metrics like RMSE and R^2 score. RMSE for the model 0.000010 was and R2 score 0.9998. With the help of this, the government and the concerned authorities could better strategize and formulate measures to improve the medical facilities. They would be able to forecast the infection rate and the spread of the pandemic, which would result in better planning of resources and would prepare a better response. Our model could be extended and trained on data of different countries which will help them in their decision-making process and help them prepare the budget to be spent on medical equipment like ventilators, kits, masks, protection gear for essential workers etc.

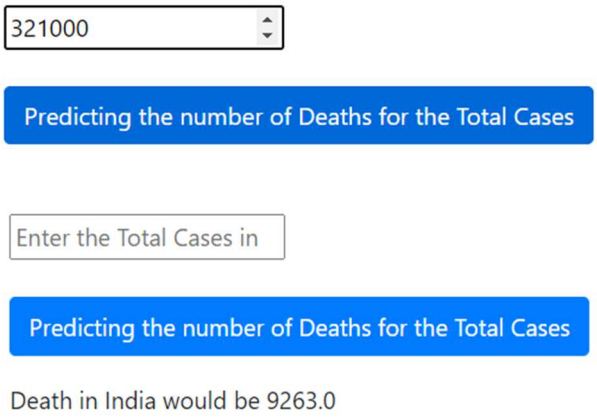

Fig. 9. Predicting the number of deaths with the help of our API.

REFERENCES

[1] Nicholas LePan. (2020, March 14). Visualising the History of Pandemics.Link:.https://www.visualcapitalist.com/history-of-pandemics-deadliest/, Last Accessed: 14-06-2020.

[2] Dave Roos. (2020, March 27). How 5 of History's Worst Pandemic Finally Ended. Link:https://www.history.com/news/pandemics-end-plague-cholera-black-death-smallpox, Last Accessed: 14-06-2020.

[3] Maria Cohut. (2020, June 3). Novel coronavirus: Your questions, answers.Link:.https://www.medicalnewstoday.com/articles/novel-coronavirus-your-questions-answered, Last Accessed: 14-06-2020.

[4] Zhang, Haiping, Konda Mani Saravanan, Yang Yang, and Tofazzal Hossain. 2020. "Deep Learning Based Drug Screening for Novel Coronavirus 2019-NCov." *Preprints (Www.Preprints.Org)* 19(8):1–17

[5] B. R. Beck, B. Shin, Y. Choi, S. Park, and K. Kang, "Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 784–790, 2020, doi: 10.1016/j.csbj.2020.03.025.

[6] L. Yan *et al.*, "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan," *medRxiv*, p. 2020.02.27.20028027, 2020, doi: 10.1101/2020.02.27.20028027.

[7] J. Chen *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *medRxiv*, p. 2020.02.25.20021568, 2020, doi: 10.1101/2020.02.25.20021568.

[8] Jason Brownlee. (2020, December 25). Basic Concepts in Machine Learning.Link:.https://machinelearningmastery.com/basic-concepts-in-machine-learning/, Last Accessed: 14-06-2020.

[9] Yann Mulonda. (2018, December 5). What is Machine Learning?"In Simple English".Link:https://medium.com/@yannmjl/what-is-machine-learning-in-simple-english-b0aaa251cb60, Last Accessed: 14-06-2020.

[10] Saishruthi Swaminathan. (2018, February 26). Linear Regression-Detailed View.Link:https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86 , Last Accessed: 14-06-2020.

[11] Mohit Gupta. ML | Linear Regression. Link:https://www.geeksforgeeks.org/ml-linear-regression/, Last Accessed: 14-06-2020.

[12] ML Polynomial Regression
Link:https://www.javatpoint.com/machine-learning-polynomial-regression, Last Accessed: 14-06-2020.

[13] Abhishek Sharma. (2020, March 16). Introduction to the Polynomial Regression(with Python Implementation). Link:https://www.analyticsvidhya.com/blog/2020/03/polynomial-regression-python/ , Last Accessed: 14-06-2020.

[14] RMSE:Root Mean Square Error. Link:https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error, Last Accessed: 14-06-2020.

[15] JJ. (2016, March 23). MAE and RMSE – Which metric is better?. Link:https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d, Last Accessed: 14-06-2020.

[16] R-Squared.,Link: https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/, Last Accessed: 14-06-2020.

[17] Jim Frost. How To Interpret R-squared in Regression Analysis. Link:https://statisticsbyjim.com/regression/interpret-r-squared-regression/ ,Last Accessed: 14-06-2020.

[18] Hannah Ritchie. Coronavirus Source data. Link:https://ourworldindata.org/coronavirus-source-data, Last Accessed: 14-06-2020.