Forecasting & Severity Analysis of COVID-19 Using Machine Learning Approach with Advanced Data Visualization

Ovi Sarkar
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh
ovisarkar@ieee.org

Md Faysal Ahamed
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh
faysalahamed@ieee.org

Pallab Chowdhury
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh
chowdhurypall95@gmail.com

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392704

***Abstract*—SARS-CoV-2 (n-coronavirus) is a global pandemic that causes the deaths of millions of people worldwide. It can cause Pneumonia and severe acute respiratory syndrome (SARS) and lead to death in severe cases. It is an asymptomatic disease that hardens our life and work conditions. As there is no effective treatment available, many scientists and researchers are trying their best to fight the pandemic. This paper focused on the coronavirus pandemic situation in the global and Bangladesh region and its related effects and future status. We have utilized different information representation and machine learning calculations to recreate the affirmed, recuperated, and passing cases. We believe the research will help scientists, researchers, and ordinary people predict and analyze this pandemic's impact. Finally, the comparison and analysis of different models and algorithms successfully showed our visualization and prediction success.**

***Index Terms*—COVID-19, SARS-CoV-2, Visualization, Polynomial Regression, ARIMA model, Facebook prophet model, Forecasting, Clustering, K - means.**

I. Introduction

COVID illness (COVID-19) is an irresistible infection brought about by a newfound Coronavirus. Coronavirus is a group of infections that causes the normal chilly, serious intense respiratory condition (SARS), and the Middle East respiratory syndrome (MERS). It is otherwise called the extreme intense respiratory disorder COVID 2 (SARS-CoV-2). SARS-CoV-2 (n-COVID) is the new infection of the COVID family that was first recorded in December 2019 in Wuhan, Hubei region of China and was already unidentified in people. It is an infectious infection that can cause respiratory disease and even lethal respiratory pain disorder (ARDS) [1]. COVID was later announced a pandemic by WHO in light of its high spread overall. The signs and indications of seriousness with COVID-19 can change from mellow to extreme. Numerous people may just have less side effects, and others may have no manifestations by any means. People who are more adult or have existing tenacious clinical issues, for instance, coronary disease, lung contamination, diabetes, outrageous strength, consistent kidney or liver disorder, or bartered safe structures, may be at higher threat of certifiable infirmity. That resembles what see with other respiratory ailments, for instance, influenza.

This paper is a push to dissect the aggregate information of affirmed passings and recouped cases over the long run, which is examined in the information investigation area. The primary center is to investigate the spread pattern of this infection around the world, including Bangladesh. In segment (V-VIII), we have used several algorithms, i.e., Polynomial Regression algorithm, Arima, and Facebook prophet time series forecasting algorithms to measure the daily increase in confirmed, recovered, death cases and growth factor. In section IX, we have analyzed the results of different models and algorithms and compared them with the least RMSE value. Finally, in section X, our research inspection's successful completion is introduced, indicating our study's summary and fruitful experiment reviews of these improved algorithms.

II. Related Work

T. Turki and Z. Wei analyzed and Predicted the death of the Middle East respiratory diagnostic victims [2]. They also identified which patient needs hospitalization at affecting the virus. They also predicted the result of local area people by classifying them in some category based on a real data-set. E. Kim and T. Lee introduced an algorithm to predict the path of transmissible diseases [3]. They also calculated the data on how to monitor and predict the growth factor and calculated the error rate. N. Zheng et al. focus on the plague model that can forecast the power of irresistible sicknesses well however don't think about different elements, for example, counteraction and control gauges—their expectations in the China area based [4].

III. Methodology

An efficient prediction algorithm has built using the ANACONDA3 Jupyter Notebook platform, where we utilize the data processing technique. The multi-stage data analysis like table and the graphical view has used to analyze the pandemic

978-1-6654-2244-4/20/$31.00 ©2020 IEEE

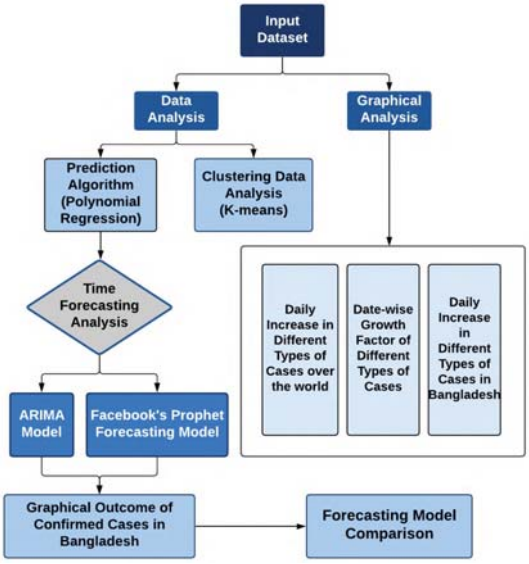

Fig. 1: Overall working process.

of COVID-19. The next week's prediction of confirming cases have done by using the Polynomial Regression algorithm, Arima model, and Facebook prophet time series forecasting. Before that, we classify our region into classification depending on the confirms cases, death rates, recovery rates. After completing the research and applying the algorithms, we find and declaring the best algorithm for the COVID-19 forecasting method by comparing the root mean square error value (RMSE). In Fig. 1 the overall flow chart of the introduced system represented.

IV. Data Analysis

Coronavirus is a respiratory infection that was distinguished in December 2019 in China and has quickly extended everywhere on the world. Ordinary numerous individuals have influenced by this infection due to restricted information on the new COVID by the clinical affiliation and the deficiency of restorative backings. The COVID-19 is an irresistible ailment that can be spread from individual to individual quickly, as endorsed on January 20, 2020. That is the reason we have gathered all the information over the world for imagining the episode of this pestilence from January 22, 2020, to August 22, 2020, from World Health Organization(WHO) and human information association [5]. In our informational index, we have utilized the perception time, region/state, nation/locale, last perception date, affirmed cases, recuperated cases and, passing cases on date-wise. We have measured the daily increase in confirmed, recovered, and death cases with python on Jupyter notebook with these data. We have used the date(observation) in the x-axis and the number of cases (confirmed, recovered, deaths) on the y-axis. The daily increase in different types of cases is shown in Fig. 2.

At that point we have estimated the development factor, which is an amount increases itself after some time. The

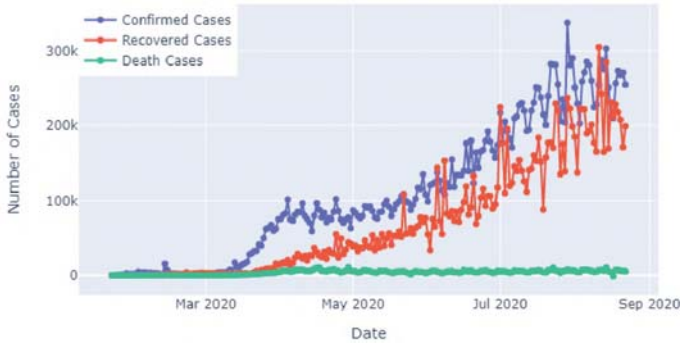

Fig. 2: Daily increase in different types of cases.

formula we have applied:

$$G = \frac{E}{P} \tag{1}$$

Here, G = Growth Factor, E = Consistently's new (affirmed, recuperated, passing) cases, and P = New (affirmed, recuperated, demise) cases on the earlier day. A growth factor over one shows an increase in confirmed cases. A development factor more than one however moving downwards is a positive sign, yet a development factor consistently more than one is the kind of remarkable development. A development factor consistent at one assigns there is no variety in any cases. We have used the date(observation) in the x-axis and the growth factor in the y-axis. The date-wise growth factor of different cases is shown in Fig. 3.

Similarly, we have estimated the day by day increment in various sorts of cases in Bangladesh. The every day increment in various kinds of cases in Bangladesh is appeared in Fig. 4.

V. Clustering Algorithm (K-means)

K-means clustering is a method of quantization vectors, which is common in data harvesting of cluster analysis. Unpredictable algorithms usually deduct data sets using only input vectors without referring to displayed or specified consequences. The approach presents a hierarchical and fast way to evaluate a given collection of data across a clustering algorithm (implying k clusters) concluded by the centering of data [6].

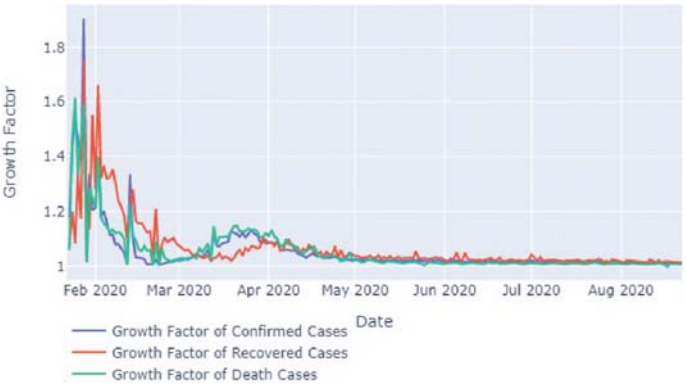

Fig. 3: Date-wise Growth Factor of different types of cases.

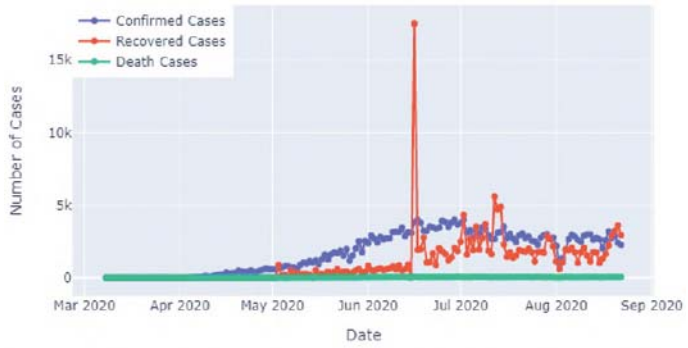

Fig. 4: Daily increase in different types of cases in Bangladesh.

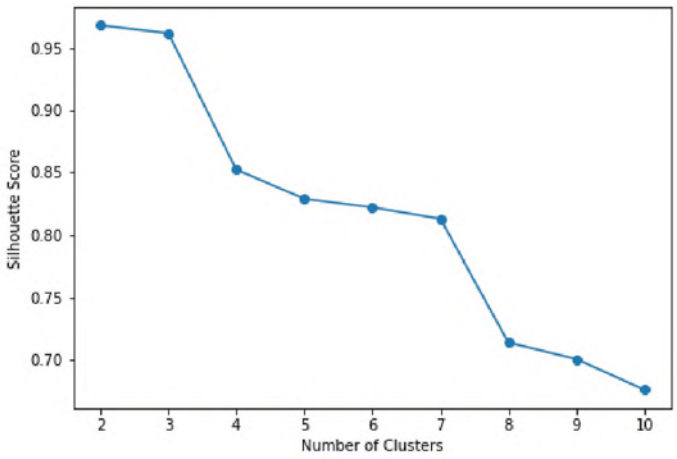

Fig. 5: Silhouette Score Method.

These centers should be located cunningly because of separate location makes a different outcome. The trustworthy solution is to make others as satisfactory and remote as feasible. The next step is to define and connect any argument with a given knowledge through a new emphasis. After finishing the initial step and an early gathering period does. Now, we have to recalculate k current centroids as the clusters rising out of the first step. Then, the point from k current centroids, another adhesive should do inside similar informational index focuses and the most compromising new focus. This progress will repeat. So the end of k areas changes their position bit by bit until there are no more differences found. Finally, the aim of using k means clustering for decreasing an outside function acknowledged as the squared error function, which is given:

$$M = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - y_j\|^2 \tag{2}$$

Where, $||x_i^j - y_j||^2$ Would be a preference test among the point of data $x_i{}^j$ and center of the group y_j would be a reference for the interval between n data points and their aggregation centers.

1) Distance Measure: To figure out the similarities between the alternatives, this distance between the points takes as a standard metric. The basic separation computation utilizes the Euclidean measurement, that characterizes the span among a^i and b^i, where [7]:

$$d = [\Sigma(a_i - b_i)^2]^{1/2} \tag{3}$$

A. Algorithm Steps

The following approaches to this algorithm:
(1) Here K points are objecting state effective internal control, and the initial position is labeled.
(2) Each group object assigns to the nearest centroid.
(3) Recalculate the position of the K centroids after assigning all the objects.
(4) Phases 2 and 3 likely proceed. And why the category should differ from the one that minimizes the metric.

B. K-means in Data Analysis

In my Data-set analysis, we apply the clustering algorithm (K-means) to classify the dangerous region in this world. Here, we take the mortal rate, recovery rate, COVID-19 positive confirmation rate as a feature for the clustering algorithm.

Analysis of the silhouette finds out the distance between the number of observations. This plot of silhouettes indicates how similar each point from group to group. And that measure has a [-1,1] range. If the silhouette coefficients are close to +1, the sample is distant from the nearest neighbor. If this coefficient value is 0, the unit is on or somewhat similar to the boundary of judgment between two neighboring clusters. In Fig. 5 negative values intimately assigned to the wrong group by those representations.

We can empower each of these features to a separate cluster. Based on the comparison of these clusters, we can consolidate the most related groups mutually; after doing that, we repeat this method until only a single bunch of drops. We are developing hierarchy clusters.

We got the chance that the community success in our knowledge code of COVID-19 could be breached and that it is more proficient to approximate k = 3. Where the plotting base depends on the mortality and recovery rate is classified. Thus why the classification does the plot shows in Fig. 6. After the classification, we categorize our result in world countries as below:

- Group 0 is a group of nations where high probability of infection and a remarkably good level of recovery. Almost no country in those packages has seen the most discernible of this pandemic, but it is now recovering at a high rate of recovery. e.g. ‘Germany’, ‘Turkey’, ‘Iran’, ‘Mainland China’, ‘Saudi Arabia’ , etc.
- Cluster, 1 arrangement of nations, have Low Mortality Rate and High Recovery Rate. E.g. ’Nigeria’, ’Russia’ etc.
- Cluster 2 is a bunch of nations that have Low Mortality Rate and Low Recovery Rate. Such states ought to raise their survival rate, others have a significant number

TABLE I: Clustering region possibility

Sl.No.	Cluster Value	Mortality Rate	Recovery Rate
1	K = 0	2.76 %	80.87 %
2	K = 1	2.45 %	31.43 %
3	K = 2	12.83 %	37.55 %

of contaminated cases, but low mortality is a good indication. E.g. 'Mexico', 'Belgium, 'Netherlands', and 'Ecuador' etc.

After building the clustering scheme in Fig. 6, we can determine the country's number present in which cluster (k value). We ascertain the dangerous area rate based on mortality and recovery progresses in TABLE I.

VI. Polynomial Regression Algorithm

There are many machine learning algorithms, and polynomial regression is a specific case of a linear regression algorithm. Here, the equation is formed by polynomial construction on the data with a direct curve correlation. This relationship is created between the objective variable and the free factors. The method is to acquire a regression model from a set of input values (calculated values). This process is a very straightforward machine learning approach, where the dependent variable is modeled as a linear order of predictors. More high-level regression methods and models include regression and multivariable regression, which is an acknowledged polynomial regression algorithm [8].

At a simple linear regression method, one feature or single independent variable. So, the formula for linear regression is:

$$\mathrm{y} = \theta_0 + \theta_1 \mathrm{X}_1 \tag{4}$$

here, Y is known objective, X means indicator, θ_0 calls bias, and θ_1 is known as the weight in condition 4.

Several machine learning techniques use high-level statistical methods to achieve regression. So there were one or several input parameters, but only one (dependent) output variable. We are using the polynomial regression method to reduce the positioning error and find out the actual positioning value from the curve. And which proposes a multiple regression to define the relevance of the input variables and the output variable

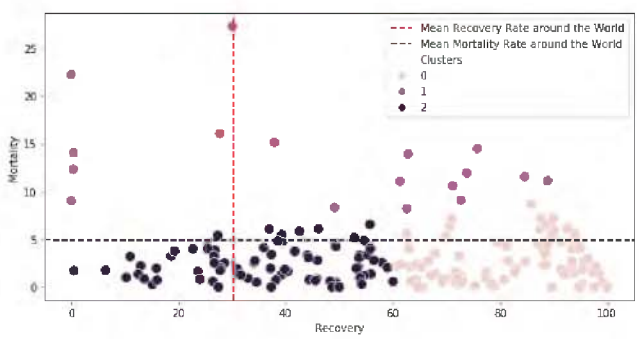

Fig. 6: Recovery and mortality rate clusters.

rendering the positioning prediction point. It expects that the output variable y (dependent), which linearly correlates to the input variables at the equation X. Multiple regression helps to determine the positioning value, thus why the output error is remaining low. Suppose, there is only one variable, the model represented relationship formula is:

$$\mathrm{Y} = \theta_0 + \theta_1 \mathrm{X} + \theta_2 \mathrm{X} + \theta_3 \mathrm{X}^2 + \theta_4 \mathrm{X}^2 + \ldots + \theta_\mathrm{m} \mathrm{X}^\mathrm{n} \tag{5}$$

Here, n is the polynomial equation degree. As n (degree) increases, the feature complexity increases in equation 4. The stated polynomial regression assessment used to forecast each number of reported case scenarios. The θ_m is the polynomial achieved by applying a sci-kit-learn Machine Learning algorithm in Python.

VII. ARIMA Model

ARIMA, a model set up by Box and Jenkins, is furthermore called the B-J model, and it's a mix of the AR (Auto-Regressive) model and the MA (Moving Average) model. Econometricians normally use ARIMA models applied to figure time arrangement, and the ARIMA model is known as the most unpredictable and progressed time arrangement determining technique in the global. It is quite familiar amongst time series approaches for the adaptive characteristics to control linear models [9]. The generalized equation of the ARIMA model is:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \mu_t - \beta_1 \mu_{t-1} - \beta_2 \mu_{t-2} - \cdots - \beta_q \mu_{t-q} \tag{6}$$

$$X_t = \nabla^d y_t \tag{7}$$

Here, X_t= new time series after steps difference of d, α_1, α_2,$\cdots$, α_p= auto regressive coefficient, β_1, β_2,$\cdots$, β_q = moving average coefficient, ∇ = backward difference operators, μ_t = random disturbance, and y_t= time series. The above model is the combination of AR(p) and MA(q) after making a d steps difference of the raw time series represented as ARIMA (p, d, q). Auto ARIMA model is used to fit the data as it is a stepwise strategy to examine various combinations of p,d,q parameters, and determines the most suitable model.

VIII. Facebook's Prophet Forecasting Model

The prophet is a decomposable time course of action model of Facebook. It utilized for anticipating time with three fundamental model parts: pattern, irregularity, and occasions. They consolidated in the accompanying condition:

$$Z(a) = P(a) + Q(a) + R(a) + \xi(a) \tag{8}$$

Here, P(a)= piecewise straight or strategic development bend shaping non-intermittent time arrangement shifts, Q(a)= periodical variations(e.g. week after week/yearly irregularity), R(a)= effects of occasions (client gave) with unpredictable timetables, $\xi(a)$= blunder term considering any uncommon varieties not upheld by the model. Irregularity can present

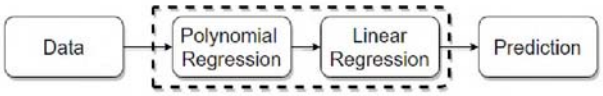

Fig. 7: Prediction approach in polynomial regression.

with numerous periods, and the investigator can get a few theories in regards to patterns. The estimations need not be dispersed equitably, and exceptions need not be barred. The Prophet is endeavoring, as a regressor, to fit various direct and non-straight elements of time as sections. In the Prophet, there are two potential pattern models for P(a). The underlying one is called Nonlinear, Saturating Growth.Seasonality can present with different periods, and the investigator can get a few theories with respect to patterns. The estimations need not be circulated equitably, and anomalies need not be avoided. The Prophet is endeavoring, as a regressor, to fit various direct and non-straight elements of time as fragments. In the Prophet, there are two potential pattern models for P(a). The underlying one is called Nonlinear, Saturating Growth. It expressed in the form of the logistic growth model:

$$P(a) = \mathrm{C}/(1 + \mathrm{e}^{-k(a-m)a}) \tag{9}$$

Here, C= carrying capacity which is the maximum value of the curve, k= growth rate that reflects curve steepness, and m= a variable offset. This calculated condition empowers non-straight displaying development with immersion when the development pace of significant worth decreases with its development.

IX. Obtained Results & Comparison in the Perspective of Bangladesh

Training of the model to be carried out offline. Our goal is to predict the confirming situation of Bangladesh's COVID-19. Therefore, we have only considered data from Bangladesh and we have verified cases concerning time on the basis of the characteristic values.

A. Polynomial regression prediction

In Fig. 7, these coefficients are related instantly to the control algorithm for predicting the number of positive cases. The input data derived as polynomial and every point of graph for each data is represented via linear regression approach. All coefficients instantly link to the control algorithm for predicting the number of positive cases. We take the value of degree = 2 in our polynomial model framework. The preparation and testing collection in the x-axis determines the number of days, and for the y-axis, the positive cases show appropriately. Once the sample data has been correctly compressed, and the scattering is completed. Given the dataset values of days and positive cases by calculating the root mean square error, the value prediction is provided. Root implies that a square error for polynomial regression is observed in which meaning is more right than a linear regression method.

If the degree value more than 2, then the complexity rises much more, and processing of model training time becoming

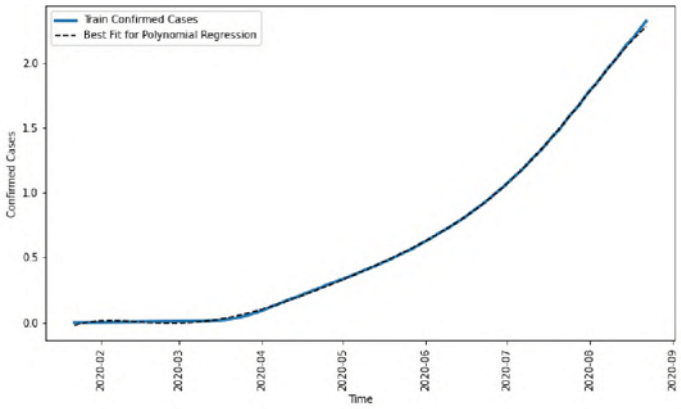

Fig. 8: Confirmed Cases in Polynomial Regression Model Prediction

higher than as usual. Our prediction is estimated by time, from previous data observing the confirmed cases in Fig. 8. In our model, primarily, we are predicting the following week's positive cases.

B. ARIMA model analysis

ARIMA model used for forecasting the time series. AR(p), MA(q) and, d are the parameters used in this model. Auto ARIMA approach widely used for searching multiple combinations of these parameters, and finally, it chooses the best model. All the COVID-19 confirmed cases in Bangladesh are used here as a train set and using auto ARIMA method it is predicting all the new confirmed cases for the following week. In Fig. 9, the ARIMA model prediction result shows, the orange line is the validation set, the blue line is the train set, and the green line is the prediction set. It shows that the prediction set is going higher.

C. Facebook's prophet forecasting model analysis

Facebook's Prophet is a technique utilized for estimating time arrangement information. Here three parts: pattern, irregularity, and occasions are utilized. As opposed to expressly taking a gander at the time sensitive succession of every perception inside a period arrangement, the Prophet drafts the

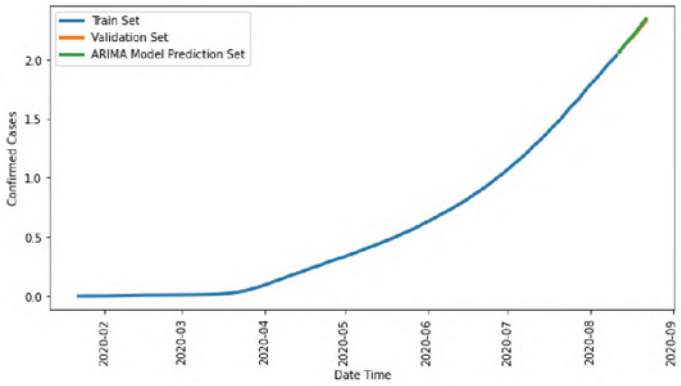

Fig. 9: Confirmed Cases in ARIMA Model Prediction.

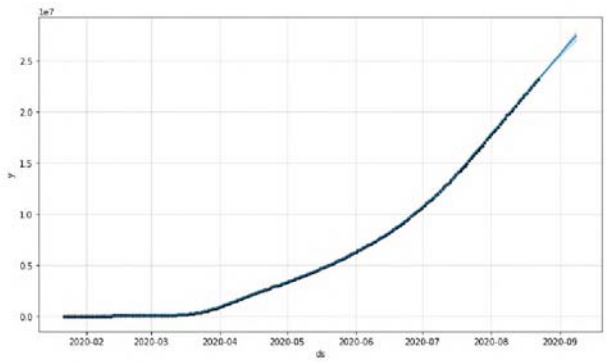

Fig. 10: Confirmed Cases in Facebook's Prophet Forecasting Model Prediction.

TABLE II: Error comparison among the models

Sl.No.	Model Name	RMSE
1	Facebook's Prophet Forecasting Model	33785.550426
2	ARIMA Model	113560.369224
3	Polynomial Regression	222589.392931

estimating trouble as a bend fitting activity. Q(a) is an occasional component speaking to an adaptable model of periodical changes because of week by week and yearly irregularity.The segment R(a) addresses unprecedented obvious days of the year, including those on eccentric plans. The $\xi(a)$ botch term depicts the information which isn't considered in the model. The dataset we have used, all the COVID-19 confirmed cases in Bangladesh, are used as a train set, and using this method, it predicted all the new confirmed cases for the following week shown in Fig. 10. Here we have utilized the date(observation) in the x-axis and the quantity of affirmed cases in the y-axis. The dark blue dab line shows the train set, and the light blue straight line shows the expectation set and the sky blue line demonstrates the approval set.

D. Comparison of different types of models

Finally, we measured the values of these models in RMSE (Root Mean Squared Error). From TABLE II, it is apparent that Facebook's prophet model has the base RMSE esteem that we have. The other two models have a higher RMSE value than this one. Facebook's prophet is much faster and more accurate than any different model. It supports a significant amount of data for making forecasts in time series methods without preparation.

X. Conclusion

In this research, we evaluate the state of COVID-19 as a result of its rapid growth in the affected cases and forecast new numbers for the following week with different models. The various forecasting techniques clearly show the increasing rate of the newly confirmed cases. With all the researched methods, we have found Facebook's Prophet model to be the perfect one to show the accurate forecasting with the least RMSE value. With the increasing number of confirmed cases, it is quite impossible to bring everything under control for a country. So it is essential to create public awareness, follow them strictly with the help of laws, and provide necessary medical resources for the general people.

References

[1] X. Wang et al., "A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization From Chest CT," in IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2615-2625, Aug. 2020, doi: 10.1109/TMI.2020.2995965.

[2] T. Turki and Z. Wei, "A greedy-based oversampling approach to improve the prediction of mortality in MERS patients," 2016 Annual IEEE Systems Conference (SysCon), Orlando, FL, 2016, pp. 1-5.

[3] E. Kim, S. Lee, J. H. Kim, Y. T. Byun, H. Lee and T. Lee, "Implementation of novel model based on Genetic Algorithm and TSP for path prediction of pandemic," 2013 International Conference on Computing, Management and Telecommunications (ComManTel), Ho Chi Minh City, 2013, pp. 392-396.

[4] N. Zheng et al., "Predicting COVID-19 in China Using Hybrid AI Model," in IEEE Transactions on Cybernetics.

[5] "Home," Humanitarian Data Exchange. [Online]. Available: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases. [Accessed: 24-Aug-2020].

[6] A. Jain, A. Rajavat and R. Bhartiya, "Design, Analysis and Implementation of Modified K-Mean Algorithm for Large Data-Set to Increase Scalability and Efficiency," 2012 Fourth International Conference on Computational Intelligence and Communication Networks, Mathura, 2012, pp. 627-631, doi: 10.1109/CICN.2012.95.

[7] Pham, D. & Dimov, Stefan & Nguyen, Cuong. (2005). Selection of K in K -means clustering. Proceedings of The Institution of Mechanical Engineers Part C-journal of Mechanical Engineering Science - PROC INST MECH ENG C-J MECH E. 219.

[8] H. Li and S. Yamamoto, "Polynomial regression based model-free predictive control for nonlinear systems," 2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Tsukuba, 2016, pp. 578-582, doi: 10.1109/SICE.2016.7749264.

[9] S. Karthika, V. Margaret and K. Balaraman, "Hybrid short term load forecasting using ARIMA-SVM," 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, 2017, pp. 1-7, doi: 10.1109/IPACT.2017.8245060.