Novel Corona Virus Prediction and Transmission Analysis using Machine Learning Models

G. Karuna^{1*}, K. Pravallika¹, Karanam Madhavi¹, V. Srilakshmi¹, K. Swaraja¹, G. Kalpana²

¹Computer Science and Engineering, GRIET, Hyderabad, Telangana, India.

²Computer Science and Engineering, VJIT, Hyderabad, Telangana, India.

Abstract. Today we all are suffering from Covid-19, a novel virus and it is the most harmful disease across the world which mainly comes under the domain of health care research. Healthcare system gives importance to health states of the population or individual. Healthcare plays a vital role in promoting physical and mental health and well-being of people around the world. Efficient health care system leads to country's economy, industrialization and development. Corona virus is dangerous animal and human pathogens and it is threatening people by spreading all over the world. Corona virus patients mostly suffer from lung infection studies have shown it clinically. We proposed detailed analysis on how to predict the expected death, recovered and confirmed cases based on the available data across the world using various machine learning models. Especially we constructed linear regression model (LRM), support vector machine model (SVMM) and polynomial regression models (PRM) and predicted future expected cases over a period of next 15 days. The error between the predicted model and official data curve is quite small in the process of transmission in data modeling. Compare to other models Polynomial regression model performs best prediction of corona positive cases. Forward prediction and backward inference of the epidemic helps to take decisions for necessary actions during Covid-19 propagation.

1. Introduction

Healthcare techniques, practices, programs and policies are the set of variables comes under health care research. Every individual must take care of their own health in order to achieve best possible outcomes. It will analyze the factors that affect the health systems and interventions.

Corona virus mainly caused for animals and human beings [2]. Mild East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) are to blame [1]. There are various symptoms for this corona virus. Each and every person may or may not have all the symptoms. Sometimes the person may not have any symptoms also. Transmission of the corona virus from animals to humans is very rare [2]. It may take up to 5-6 days to identify the symptoms or it may take up to 14 days also it varies from person to person. Corona infection has a pH ranging from 5.5 to 8.5. We ought to eat alkaline foods that are higher in PH.

1.1 Background

It first originated in food market, Wuhan which is the seventh largest place in china. Corona virus can be tested through Swab test, blood test etc [4]. A swab is inserted in nose or mouth to take the sample for testing purpose [4]. In December 2019 [5], Corona virus spread rapidly and widely in china so many people are affected by the virus [3]. Europe has become the central hub of the pneumonia epidemic, and on March 11, a new pneumonia infestation known as a global pandemic was declared [7]. Viruses have become a major threat to human health and safety all over the world [28], and as a result, many individuals have been infected and harmed [23]. Whenever the person coughs or sneezes the droplets are going to fall on earth as they cannot remain in the air for long time. So, people should take necessary precautions while coughing or sneezing. People should follow all the steps implemented by the government in order to protect from this pandemic and one should not step out of the house if it is an emergency case then only should go out by taking steps like Covering the face with mask, sanitizing hands and washing the

*Corresponding author: karunavenkatg@gmail.com

hands frequently.

In this article, we developed and built a new model of corona virus transmission using linear regression [21], support vector machines (SVMs) [11, 12, 14], and polynomial regression Model [8]. In propagation process the simulating of COVID-19 studies found that curves of proposed will get simulated by the official data curves for all the countries [8]. Major affecting factors are the virus spreading, such as confirmed cases, death cases and recovered cases. We can predict epidemics development in various regions and they are inferred at the time of initial cases. Some of the symptoms of the Corona virus are loss of smell and taste, fever, conjunctivitis, headache, dry cough, High temperature, tiredness, diarrhea, rashes on skin and body, Finger and toe discoloration [15] Severe symptoms include pain breathing or breathlessness, chest pain or tenderness, and decline of speech or action. If any individual have serious symptoms they need to seek for immediate medical attention. Mild symptom people can take their medication at home. It may take up to 5-6 days or 14 days to identify the virus symptoms.

Viral test and anti-body test are the two kinds of tests for corona virus. If you are suffering from infection currently then viral test is applicable. If you have past infection then you go for antibody test [18]. These anti body test cannot be shown if you are suffering from infection currently your body takes 1-3 weeks to develop antibodies. Having antibodies might prevent from getting infected again to the virus. A swab test [4] is also done to the virus a swab is inserted in the nose or mouth to collect the sample. Results may take some days .Some type of kits are also available to test the virus .The result must be positive or negative. If it is positive the person is suffering from virus. If it is negative he is not affected to the virus. Some tests may be accurate or not. Some of the tests may not getting accurate results even though the person is suffering from different symptoms like breathing. These types of cases can be detected by using MRI and CT scan. In this the project's main purpose is predicting the future covid-19 cases by taking the datasets and applying the machine learning algorithms such as SVM , linear[21] and polynomial regression models. If we take the dataset accurately then we can predict for the next 10 to 20 days. The key factors that influence the spread of COVID-19 are death instances, recovered incidents, and confirmed cases. We can find out easily by using the existing data and can predict the evolution of existing epidemic data. We make bold predictions for the data which exists abroad and trends in development of epidemic in different regions, associated control time and the best early transmission trying to trace of nations with diverse

dates

Corona virus became threat to people safety and health due to its harmful and spreading power. Presently, outbreak is effectively controlling and spreading rapidly is done by other areas due to corona virus. It is hard to determine the virus transmitted and the time it takes to spread through the areas. At different stages the epidemic has the transmission characteristics and uses SVM, linear[21] and polynomial regression models to create a virus estimation procedure Instances that have been confirmed, cases that have starved to death, and situations that have been recovered are the factors majorly that will affect the spread of virus. Prediction is done in development and trend of epidemics in different regions at initial cases at the time of inferring. We analyze further the control time's impact in the spread of epidemic. Related countries epidemic prevention is controlled by applicable models and analysis of data that might provide foundation and direction. Simulation results are getting compared with real data, the propagation process and influencing factors can be analyzed. There are datasets for death cases, confirmed cases, recovered cases [16] for different states, countries and regions. If the dataset is most accurate and each and every region of the data is included then we can find the predictions for villages and rural areas if the data is available most accurately. By prediction we can prevent the spreading of the virus by taking the necessary precautions timely.

2. Related Work

Using enhanced epidemiological models for prediction and forecasting COVID-19 epidemic in India [24]. By using SARS (Severe Acute Respiratory Syndrome) we come to know that corona virus is an infectious disease and thus analyzing the future predictions and forecasts in India. Here in this as they are insufficient tests conducted and it is highlighted using a relevant mathematical formula and the relationships are establishing for number of people infected and the death counts.

Based on the tracking of social interaction of people and patient prediction system can predict the possibility of infections[25]. They used GPS and BLE (Bluetooth Low Energy) tracking system[30] based on social Interactions and proposed a graph model in order to analyses the proposed algorithm behavior. Embedded sensors and Smartphone used for diagnosing Novel Coronavirus in AI [26]. It is less commercial if we use smart phones and sensors to detect the virus compare with the medical kits. Based on the memory, wireless sensors including cameras are

used. Sensors can detect the severity of the disease through prediction[27]. Flattening the curve by using Bhilwara Model for COVID-19 Outbreak in India specified in [29]. Polynomial 3rd Degree Regression is used in Bhilwara Model and calculated the growth rate and mean by using all the datasets of all states in India.

3. Proposed Work

3.1 System Architecture

There are various steps used in this corona virus prediction model and they are Data collection, Feature Extraction, Data preprocessing, Model Selection, Data visualization [19]. Initially you need to collect all the data[6] and then need to check which elements are needed for the proposed work then need to extract those particular features. While implementing the code first you need to load all the libraries needed there after load the datasets and preprocessing is done to all the data means comparing the data by using plots and graphs, scatter plots can also be used. Model selection is done after preprocessing which algorithm gives the best and accurate results needs to be selected, can be visualized by using various charts and graphs.

Fig.1. System Architecture

3.1.1 Data Collection

Data collection means the collection of data by different means such as groups, surveys, observations, questionnaires and various organizations. Data such as confirmed cases, deaths reported and recovered cases needs to be collected for some days or months.

3.1.2 Feature Extraction

Amount of data for the accurate and complete data set which is original must be processed effectively by select and/or combine variables into features for Feature Extraction[10].

3.1.3 Data Preprocessing

Data preprocessing can be done by data cleaning, integration, reduction and transmission. We need to remove the unnecessary data which is not required[20]. Data cleaning means identifying inaccurate or incorrect records from the dataset and removing those elements from the dataset. Data integration means combining the single unified view of data. Integration includes cleaning, mapping which is ETL and transformation. Data integrations are residing in different types of sources.

3.1.4 Model Selection

We used linear, SVM and polynomial regression methods to show the difference between the official data collected and the simulated data[16]. There is a slight difference between the two curves.

3.1.5 Data Visualization

Data Visualization means in which form we can see the output like bar graphs, pie charts, pictures, curves etc. It is the graphical representation of data and information. It provides tools[13,17] to accessible and understands outliers, patterns and data.

3.2 Methodology

Covid-19 prediction model can be implemented by using linear regression, support vector machine and polynomial regression models. Linear regression is a linear approach. Relationship between the independent and dependent variables gives the prediction of relationship between the two

variables.

Fig.2. Linear Regression graph

A linear regression equation is of the form

$$Y = a + bX$$

where X is the explanatory variable Y is the dependent variable

Slope of line is the b, a is the intercept

Support vector machine (SVM) algorithm used in

finding hyper plane in N-dimensional space where N is the number of features that classify the data points distinctly. Data points can be separated by using two classes and chosen by many hyper planes which are possible. Maximum margin finding is the main objective which means the distance between data points of both classes which is maximum. Margin distance maximization gives reinforcement to future data points can be classified with more

confidence.

Fig.3. SVM graphs

Fig.4. Margin separating hyper plane

Regression is a type of algorithm that gives and models the dependent(y) and independent variable(x) relationship as polynomial of nth degree. The equation for Polynomial Regression is given below:

$$y = b_0 + b_1x^1 + b_1x^2 + b_1x^3 + \dots + b_nx_1^n$$

Application of linear dataset or model gives good result for Simple Linear Regression and application of same model for non-linear dataset without any modification will leaves a drastic output. This leads to increase of loss function and high error rate and decrease in accuracy. In non-linear fashion, the data points are arranged for such cases for that we need regression type of model.

Nonlinear and linear dataset can be understood in a better way by the comparison diagram below. Application of linear dataset or model gives good result for Simple Linear Regression and application of same model for non-linear dataset

without any modification will leaves a drastic output. This leads to increase of loss function and high error rate and decrease in accuracy. In non-linear fashion, the data points are arranged for such cases for that we need regression type of model. Nonlinear and linear dataset can be understood in a better way by the comparison diagrams.

4. Machine Learning Models

For covid-19 propagation analysis, three models are using for prediction purpose. They are SVM model, linear regression model and polynomial regression model.

4.1 Support Vector Machine (SVM) Model

SVM is nothing but support vector machine the main aim is finding data points in a distinctly classified method to find the hyper plane of this model. To separate data points hyper planes are chosen. SVM is used for classification and regression tasks. Maximum margin finding hyper plane is the main objective i.e data point's distance is maximum. It is a linear model used for classification and regression problems. It solves linear problems and non-linear problems and works for practical problems too. SVM creates a line or hyper plane which separates the data into

classes.

Fig.5. Hyper plane separating the data into classes

Fig.6. Separation of data points in 3-dimensional view

4.2 Linear Regression Model

Linear Regression model uses linear approach and shows the relationship between the dependent variable and independent variables. It is used in predicting the two variable relationship. We can use this model to predict output values for inputs which are not present in the dataset. We may believe that those data points will fall under that line.

Fig.7. Random error for Linear Regression

Fig. 8. Training and Testing samples

4.3 Polynomial Regression Model

Relationship between dependent variable(y) and independent variable(x) gives the Polynomial Regression for the nth degree. It is a special case of linear regression. Curvilinear relationship between the target and the independent variables is plotted.

Polynomial equation of n degree is represented as

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + \dots + b_n x_n^n$$

where b_0 is the bias, $b_1, b_2, \dots, b_n$ are the weights in equation of polynomial regression, n is the degree of the polynomial. Increasing value of n increases higher order terms, therefore the equation becomes complicated.

Fig. 9. Polynomial Regression

Fig. 10. Polynomial for 3rd, 5th, 9th order Polynomial fit

5. Result Analysis

Initially we imported all required libraries and loaded data set of covid-19 death, recovered and confirmed dataset of two months from the specific location and preprocessed the data to remove outliers then analyzed the data by making predictions with construction of linear regression, support vector machines and polynomial regression models. Finally the results are visualized using line graphs shown in below figures 11, 12, 13.

Data collection, Feature Extraction, Data preprocessing, Model Selection, Data visualization is done as step by step process. The results are clearly shown that the expected deaths, recovered and confirmed cases in next 15 days from 22nd January 2020.

5.1 Resultant graph for SVM Model

Fig.11. Prediction of Corona cases using SVM

5.2 Resultant graph for Linear Regression Model

Fig.12. Prediction of Corona cases using Linear Regression

5.3 Resultant graph for polynomial Regression

Fig.13. Prediction of Corona cases using Polynomial Regression

Code snippet:

Linear regression:

```
def plot_predictions(x, y, pred, algo_name, color):
    plt.figure(figsize=(16, 9))
    plt.plot(x, y)
    plt.plot(future_forecast, pred, linestyle='dashed',
             color=color)
    plt.title('Number of Coronavirus Cases Over
    Time', size=30)
    plt.xlabel('Days Since 1/22/2020', size=30)
    plt.ylabel('Number of Cases', size=30)
    plt.legend(['Confirmed Cases', algo_name],
               prop={'size': 20})
    plt.xticks(size=20)
    plt.yticks(size=20)
    plt.show()
```

SVM:

```
vm_df = pd.DataFrame({'Date':
    future_forecast_dates[-20:], 'SVM Predicted # of
    Confirmed Cases Worldwide':
    np.round(svm_pred[-20:]))
svm_df
```

Polynomial Regression:

```
linear_pred = linear_pred.reshape(1,-1)[0]
poly_df = pd.DataFrame({'Date':
    future_forecast_dates[-20:], 'Predicted number of
    Confirmed Cases Worldwide':
    np.round(linear_pred[-20:]))
poly_df
```

Mean Absolute Error (MAE) and Mean Square Error (MSE) values for three Models specified in below table1. The resultant graphs in fig 10, 11 and 12 are depicting that the actual data curve (available data) and predicted data curve for next 15 days and also the MAE and MSE errors calculated for all three models as given in table1. It is clearly shown that the polynomial regression model got less error rate compared to linear regression and SVM models. Hence it is correctly producing the expected death cases, recovered cases and confirmed cases.

Table1. Error values of MAE and MSE methods

Model	Mean Absolute Error (MAE)	Mean Square Error (MSE)
Linear Regression	11965.53	307996364.01
SVM	16262.29	284685673.18
Polynomial Regression	1327463.79	2775851143565.70

As per observations of Mean Absolute Error (MAE) and Mean Square Error (MSE) values it is clearly identified that polynomial regression model is well predicted the Corona positive cases

accurately and can recognize easily the expected future growth in positive cases.

6. Conclusion

Covid-19 is novel dangerous virus which causes human death and created a pandemic situation for every one day to day life. In this proposed work, propagation analysis and prediction of covid-19 done by taking the datasets from various organizations. While doing some preprocessing techniques, visualizing the data and by handling the missing values and redundancy, there is a slight difference between the predicted and the actual curves. We make bold predictions using linear regression, SVM model and polynomial regressions accurate predictions can be done by these models. By these predictions we can be aware of the pandemic and take necessary precautions to safe ourselves from covid-19 and Polynomial regression model predicted better than linear and SVM models. For future enhancement of this work, non-linear models may be helpful for further accurate and better prediction of Covid-19 cases.

7. References

1. Jie Liu, Wanli Xie, Yanting Wang, Yue Xiong, Shiqiang Chen, Jingjing Han, Review article, *International Journal of Surgery*, 81, pages 1-8, (2020).
2. Lam, T.TY., Jia, N., Zhang, YW., et al. 583,282–285 *Nature* (2020).
3. V. Chamola, V. Hassija, V. Gupta, and M. Guizani, *IEEE Access*, vol. 8, pp. 90225-90265, (2020)
4. D. Brinati, A. Campagner, D. Ferrari, et al. *J Med Syst* 44, 135–137. (2020).
5. Buddhisha Udugama, Pranav Kadhiresan, Hannah N. Kozlowski, Ayden Malekjahani, Matthew Osborne, Vanessa Y. C. Li, Vanessa Y. C. Li, Hongmin Chen, Samira Mubareka, Jonathan B. Gubbay, Warren C. W. Chan, *ACS Nano* 2020, 14, 4, pp. 3822–3835, (2020).
6. Dhanalaxmi, B., Apparao Naidu, G., Anuradha, K. *Procedia Computer Science*, 46, pp. 432-442, (2015).
7. Lixiang Li, Zihang Yang, Zhongkai Dang, Cui Meng, Jingze Huang, Haotian Meng, Deyu Wang, Guanhua Chen, Jiaxuan Zhang, Haipeng Peng, Yiming Shao, *Infectious Disease Modelling*, Volume 5, Pages 282-292, (2020).
8. K.-R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. *International conference on artificial neural networks*. Springer-Verlag, pp. 999–1004, (1997).
9. Ostertagová E. Polynomial regression modelling. *Procedia Eng.*, 48:500–506 (2020).
10. Usha Kumari, Ch., Jeevan Prasad, S., Mounika, G. *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, art. no. 8819750, pp. 1095-1098, (2019).
11. M. Awad and R. Khanna. *Support vector regression. Efficient Learning machines*, (2015).
12. Travel-time prediction using support vector regression, Wu C.-H., Ho J.-M., Lee D.-T. *IEEE Trans on Intelligent Transportation Syst.*, 5(4), pp. 276–281, (2004).
13. Kumar, S.K., Reddy, P.D.K., Ramesh, G., Maddumala, V.R., *Traitement du Signal*, 36 (3), pp. 233-237, (2019).
14. G.F. Smits and E.M. Jordaan, *Proceedings of the international joint conference on neural networks. IJCNN'02*, vol. 3. pp. 2785–2790, IEEE, (2002).
15. WJ Wiersinga, A Rhodes, AC Cheng, SJ Peacock, and HC Prescott. *JAMA*, online first on July 10, (2020).
16. P Nayak, B Vathasavai, *IEEE Sensors Journal* **17(14)**, 4492-4499, (2017).
17. Swaraja K , *Multimedia Tools and Applications*, 77 (21), pp. 28249-28280, (2018).
18. Deeks JJ, Dinnes J, Takwoingi Y, Davenport C, Spijker R, Taylor-Phillips S, Adriano A, Beese S, Dretzke J, Ferrante di Ruffano L, Harris IM, Price MJ, Dittrich S, Emperador D, Hooft L, Leeftang MM, Van den Bruel A *Cochrane Database Syst Rev.* Jun 25;6(6):CD013652, (2020).
19. J. Shuja, E. Alanazi, W. Alasmay, and others. *Intelligence Appl* (2020).
20. Prasanna Lakshmi, K., Reddy, C.R.K. *ICNIT 2010 - International Conference on Networking and Information Technology*, art. (2010).
21. Samit Ghosal, Sumit Sengupta, Milan Majumder, Binayak Sinha, *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* Volume 14, Issue 4, Pages 311-315, ISSN 1871-4021, (2020).
22. Melina Hosseiny, Soheil Kooraki, Ali Gholamrezanezhad, Sravanthi Reddy, and Lee Myers, *American Journal of Roentgenology*, 214:5, 1078-1082, (2020).
23. Kora, P., Kalva, S.R. *SpringerPlus*, 4 (1), art. no. 481, 19 p. (2015).
24. S. Singh, P. Raj, R. Kumar, and R. Chaujar, *Second International Conference on Innovative*

Research in Computing Applications (ICIRCA), Coimbatore, India, pp.93-97, (2020).

25. C. Sandeepa, C. Moremada, N. Dissanayaka, T. Gamage, and M. Liyanage, IEEE 3rd 5G World Forum (5GWF), Bangalore, India, pp. 13-18, (2020).

26. H. S. Maghded, K. Z. Ghafoor, A. S. Sadiq, K. Curran, D. B. Rawat, and K. Rabie, IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, (2020).

27. P. Nayak, B. Vathasavai, 7th International conference on cloud computing, Data Science, (2017).

28. SB Babu, A. Suneetha, GC Babu, YJN Kumar, G Karuna, Periodicals of Engineering and Natural Sciences (PEN) 6 (1), 229-240 (2018).

29. A. Bhati and A. Jagetiya, 5th International Conference on Communication and Electronics Systems (ICCES), COIMBATORE, India, pp. 951-956, (2020).