# A Comparative Analysis of COVID Forecasting by Using Various Machine Learning Methods

Jamaluddin Mir[1*], Ahsan Zubair[2], Ayesha Aslam[2], Asim Shahzad[2], Muhammad Asshad[2], Aamer Khan[2]

[1] *Faculty of Computer Science & Information Technology (FSKTM), Universiti Tun Hussein Onn Malaysia (UTHM), 86400 Parit Raja, Batu Pahat, Johor, Malaysia*

[2]*Department of Computer Science, Abbottabad University of Science & Technology (AUST), Havelian, Pakistan*

*Corresponding author: mirjamal70@gmail.com*

***Abstract*— Covid-19 emerged as one of the most infectious diseases in the history of mankind, affecting nearly 250 million people all over the world in just a short period. The pandemic which started in China, has now spread all over the world, taking about 5 million lives globally. This has also severely affected the economies of countries and has proved to be a burden on health care systems. Due to these reasons, forecasting the spread of the disease has become critical so that concerned government authorities in countries can have the chance to mitigate the spread and plan health care resources efficiently and properly. This makes it more important to have a reliable forecast so that resources can be planned ahead of time. In the present work, linear regression is used for time forecasting the spread of Covid-19 in Pakistan. Statistical parameters and metrics have been used to evaluate and validate the model. The results show that linear regression results are highly reliable, time efficient and accurate.***

***Keywords*— Covid-19, forecasting, machine learning, Support Vector Machine, prediction**

## 1. INTRODUCTION

The world is facing SARS-CoV-2 diseases in the shape of a global pandemic. The origin of the virus still remains undetermined yet scientists believe that COVID-19 virus migrated from bats to humans. Although there are some conflicting reports but it is believed that the virus originated in Wuhan, China and later spread to more than 180 countries. Patient zero was reported on 8 December 2019 in Wuhan, China [1] while first death due to the virus was reported on 9 January 2020. Strict preventive measures were enforced by the Chinese authorities which included closures of airports and highways, suspending public gatherings, closure of all forms of public transport, shops and all such activities which were based on social gatherings of population. These prohibitions were an attempt to minimize the community level transmission of COVID-19. On 31 January 2020, WHO declared spread of COVID-19 an emergency for the whole world and was declared a pandemic by 11 March 2020. As of today, more than 250 million cases have been reported across the world and the death count is about million. Researchers have tried to come up with transmission models to efficiently implement control measures [2].

According to WHO, global mortality rate due to COVID-19 is around 3.4% and mortality rate in Pakistan is 2.1% which is much lower than some of the other countries. Also, the recovery rate in Pakistan is around 96% which is much higher than some of the other neighboring countries (77% in India).

Several factors play a very important role in spread, death, and recovery rate of Covid-19, including government lock down policies to stop the spread, enforcement of social distancing measure, closure of places of gathering and worships and timely responses according to changing situations, population density, climatology parameters like wind speed, humidity, precipitation, temperature, overall immunity of population and adherence to policy guidelines [3][4][5]. In [6], a detailed analysis on the performance of 142 countries for their response to COVID-19 outbreak has been presented. This research work uses k-means clustering, hierarchical clustering, decision tree and random forest algorithm for clustering analyses and analysis of several parameters. The results of this research work shows that certain factors which were believed to be important for response to pandemic did not prove to have an impact in this research. These include GDP of a country, rate of smoking among population and number of diabetic patients. In [7], researchers identified four most challenging factors that need to be given foremost importance in response to COVID-19 pandemic and have also highlighted the use of Artificial Intelligence/Machine Learning for overcoming these in issues. These identified issues include:

1. Developing effective ways for managing health care resources which have proved to be limited even for the most developed countries due to the extent of pandemic
2. Developing personalized plans for managing and treating patients as underlying health issues have been found out to cause different types of complications among patients
3. Informing policies on the basis data driven and analysis by AI/ML methods and promoting effective collaboration
4. Speeding up the clinical trials for developing vaccines

Researchers have also found that older people, children, and those with already prevailing health conditions are more susceptible to this new virus infection [8] which also a key factor to determine the infection recovery and death rate. This also plays an important role in determining infection, recovery, and death rate.

Researchers have implemented various methods to estimate and forecast the impact of COVID-19. Most noticeable of these methods include stochastic methods [9], Weibull distribution, exponential growth mode, lognormal distribution, ARIMA, NARNN and LSTM model [10], genetic programming [11] and several other methods. In [12], a hybrid model which is based in ANFIS (Adaptive Neuro Fuzzy Inference System) and MLP-ICA (multi-layered perceptron - imperialist competitive algorithm) is proposed for predicting the number of new infections and mortality rate. A comparative analysis of four machine learning methods, including Linear regression, Least absolute shrinkage and selection operator (LASSO), support vector machine (SVM) and exponential smoothing (ES) is presented in [13]. The results show that ES outperforms all the other methods for the dataset used in this study whereas SVM showed worst performance in this case. It was observed the proposed model works fine as long as there is no significant change in events like change in government policies. While several studies have predicted an incubation period of 5.1 for the virus and 14-day necessary quarantine for an infected person, no valid study has been published yet which could estimate the exact reproduction rate of the virus. Therefore, it still remains quite difficult to predict the behavior of virus in coming days. Also, China and European countries have been the focus of majority of the studies. A genetic programming-based model has been used in [11] for time series prediction of COVID-19 in India. This study uses confirmed cases and death count as two major parameters for analysis.

Corona virus is a large family of viruses which infects respiratory, hepatic and gastrointestinal systems in humans and causes neurological diseases. They are spread mainly through birds, livestock, mice and several other exotic and wild animals [14], [15], [16]. Two outbreaks of SARS-CoV and MERS-CoV in recent times have proved that the virus transmits from animal to animal as well as from human to human [17]. In [18], the researchers proved that the virus is likely to have originated in bats as it is quite similar to two of the bat-derived strains of the virus. On 30 January 2020, it was confirmed by the Center for Disease Control (CDC) that the virus can transmit from human to human. It was reported by CDC that the virus can spread through air, by close personal contact, touching surfaces or objects that have virus and in some very rare cases through fecal contamination. The 14 day incubation period of the virus remains one of the serious traits and during this time, the virus can spread from one individual to another [19]. The spread of such a virus that can be easily transmitted has proved to be quite dangerous and demands strict policies and guidelines to be enforced and followed. China and Malaysia are two excellent examples of such cases where the spread of the virus has been contained due to strict enforcement of policies on part of the authorities and strict adherence to those policies on part of the general population. The result of cooperation between the authorities and public in these countries is evident from the results and number of current infections. On the other hand, countries like USA and India are still failing to control the spread of the virus. Therefore, it is very critical to forecast the spread of virus so that necessary polices can be implemented to ensure that the spread of pandemic is controlled.

Since COVID-19 is a novel virus, hence lack of clinically established treatment, vaccine and the health care systems of developing countries are facing a huge strain which might increase with the onset of winters in most parts of the world. The current situation has also been the cause of several socio-economic and public health issues that reflect to emphasize the importance of forecasting such type of pandemic situations. Several mathematical models have been used to develop a better understanding of such pandemics and provided with useful insights for controlling the outbreaks. It must be noted that while the mathematical models might depict the conditions of outbreak and even provide forecasts, but in population-based models, the real-world complexities are very challenging. However, stochastic models have proved to be quite useful [20]. In any case, long-term forecasting of COVID-19 is arguable when using a simple mathematical model.

In [21], a mathematical model has been proposed to estimate the number of COVID-19 infections for the first 15 days of January 2020, which were not reported at the time. Moreover, it was predicted that after mid-January, the number of cases will increase to almost 21 times the current number of cases. In [22], an estimation model for the infection rate of COVID-19 in Wuhan, China was proposed which was based on the data of 565 Japanese nationals that were evacuated from Wuhan in January 2020. The research work reported that the estimated infection rate is 9.5% while the death rate ranged between 0.3%-0.6%. In [23], a mathematical model was proposed for the estimation of COVID-19 transmission risk. The research also made seven days forecast and predicted that the infection would reach its in peak in 2 weeks' time. In [24], a very limited data was used to calculate the human-to-human transmission rate of COVID-19 virus. It was proposed that the transmission rate according to the selected data was 0.4%. Moreover, they proposed that if the time from appearance of symptoms to hospitalization or isolation of the patients is reduced by half, the transmission rate will come down to 0.012%. In [25], an estimation model for the death rate due to COVID-19 was presented based on two different scenarios. A comparison of several time series forecasting models is proposed in [26] for forecasting Hepatitis A virus. The authors used the dataset for Turkey which spanned over 13 years. Four time series forecasting models were used in this study, namely multi-layer perceptron (MLP), time delay neural networks, radial basis function (RBF) and auto regressive integrated moving average (ARIMA). Comparison of these methods showed that MLP had better performance than other three models. A Kalman filter-based forecasting models was proposed in [27] for forecasting the seasonal outbreak of influenza. In another study [28], a dynamic model with Bayesian inference was used to forecast the outbreak of Ebola in three African countries, namely Liberia, Sierra

Leone and Guinea. In [29], a mathematical model for analyzing and forecasting the spread of SARS virus was proposed. It was found that the reproduction rate of the virus in Hong Kong and Toronto was 1.2 and 1.32 respectively. In [30], a forecasting model for type A influenza, most commonly known as H1N1 virus, was proposed. In [31], [32] a probability-based model for predicting the spread of MERS was proposed. Adaptive Neuro-Fuzzy Inference system (ANFIS) [33] is another model which has wide applications in forecasting problems. Several studies show that ANFIS has shown good performance for time series prediction problems. The combination of neural networks and fuzzy logic in the model offers flexibility in determining nonlinearity of the time series data. In [34] a forecasting method based on hybridization of ANFIS model with flower pollination algorithm (FPA) using Salp swarm algorithm (SSA) is used for forecasting the spread of COVID-19 virus over the next ten days. The study was based on the data from China.

Based on the above literature review, it can be fairly deduced that linear regression-based prediction models are simple to implement, robust and are quite capable to reduce overfitting. Also, since the disease is novel, not much data is available for application of some of the latest machine learning methods. However, the results show that Linear regression produces good enough results which can be the foundation of future research and policy makers. Rest of the paper is organized as: Section 2 presents the methodology of this research work, section 3 presents the results and discussion and section 4 discusses the conclusion and future work.

## 2. METHODOLOGY

Several studies show that time forecasting holds very high theoretical significance and importance for engineering applications. Due to its application in numerous fields of science, time series forecasting is a very popular research area. These applications include traffic planning, climate forecasting for transportation, agriculture and urban planners, financial forecasting and several more. In time series forecasting, some of the data can be utilized by linear models and some of it must be characterized by non-linear models by using different methods. For time series forecasting, several prediction models have been used in literature in which a standard linear regression (LR) model is either a linear auto-regression model (AR) or a linear auto-regressive integrated moving average model (ARIMA) [35]. LR model of time forecasting has diverse applications ranging from forecasting electricity demand [36] to stock market [37] and forecasting of wind speed [38].

This research work applies 4 machine learning methods on the dataset and compares their results afterwards on the basis of several basis statistical parameters. The four methods used in this study are: Linear regression, Support Vector Machine (SVM), Quadratic SVM and Complex Tree. Several research studies show that the four methods used in this study have established their supremacy for time series forecasting.

The motivation for using LR for prediction of COVID-19 lies in its simplicity, easy interpretability, and higher level of accuracy. Linear regression predicts numerical values by performing operations on a dataset where the target values have been defined and the results can be extended by adding new information to the already existing data. During this process the relationship established among the predictor and target values can form a pattern which can be used on other datasets where target values are unknown. The dataset is divided into two parts. First part is used for defining the model whereas the second part is used for testing. In this research work, 80% of the data is used for training and 20% for testing purpose.

Linear regression is one of the most commonly used machine learning models which belongs to supervised learning subset of machine learning. A predictive modelling method is concerned with minimizing the error in a model so that it becomes capable of making accurate predictions. Regression models are mostly used for finding relations between variables and making forecast. Linear regression predicts a value for dependent variable ‘y’ on the basis of an independent variable ‘x’. In other words, this model determines a relation between variable ‘x’ (input) and variable ‘y’ (output), hence named Linear Regression. The main idea behind Linear Regression is to create a trend line based on the dataset that best fits the data points on the plot so that the model could forecast the values for future. It is then used to determine a trend which can either confirm or deny the correlation between the attributes or variables involved in the study. If the dataset is large, it can produce much better results, however linear regression is quite capable for application on smaller datasets. In linear regression, such a trend line is considered to be best which has the minimum error values between the predicted and observed values. This is also called regression line and the errors are called residuals.

The strength of linear regression lies in its simplicity meaning that it can be applied on numerous problems belonging to different fields. In the first step, the dataset needs to be plotted along a line. If the difference between the predicted value and the results is small enough, linear regression can be applied on the problem. Once the model is successfully built, the difference between predicted and actual values is analyzed to determine if the model can make good predictions. If the difference is significantly small, the model is considered to be showing good performance. However, these values vary from problem to problem depending upon the variation that can be handled by data. For evaluation of predictive model, some of the metrics used in this study are:

### *2.1. $R^2$:*

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\gamma_i - \gamma P_i)^2}{\sum_{i=1}^{n}(\gamma_i - \bar{\gamma}_i)^2} \quad (1)$$

The value of $R^2$ lies between 0 and 1. An $R^2$ value of 0 means that there is no relation whatsoever between the dependent and independent variable. On the other hand, R2 value of 1 means that dependent variable can be successfully predicted from independent variable without any error. $R^2$ value lying between 0 and 1 shows the extent to which a dependent variable can be predicted. For example, $R^2$ of 0.3 means that there 30% variance in variable Y, $R^2$ value of 0.5 means that it is 50% predictable and so on.

### *2.2 Root Mean Square Error (RMSE):*

Root Mean Square Error (RMSE) is the standard deviation of the residuals or in other words, it is the standard deviation of prediction errors.

$$RMSE = \frac{1}{N_s}\sum_{i=1}^{N_s}(\gamma\gamma P_i - \gamma_i)^2 \quad (2)$$

### *2.3 Mean Absolute Percentage Error (MAPE):*

$$MAPE = \frac{1}{N_s}\sum_{i=1}^{N_s}\frac{|\gamma P_i - \gamma_i|}{\gamma P_i} \quad (3)$$

### *2.4 Mean Absolute Error (MAE):*

$$MAE = \frac{1}{N_s}\sum_{i=1}^{N_s}|\gamma\gamma P_i - \gamma_i| \quad (4)$$

### *2.5 Training Time:*

The total time it takes to train a model.

## 3. RESULTS AND DISCUSSION

The results of this research work are based on linear regression model and are capable of successfully predicting the future trends for Covid-19. The dataset used in this study is reported by Government of Pakistan and is compiled by the researchers involved in this study. As the data points are very limited in the dataset so this makes the dataset unsuitable for deep learning methods which might result in overfitting [39]. It must be emphasized that governmental policies will have great effect on the outcome of the outbreak. Model predictions in this research work were analyzed using RMSE, MSE, $R^2$ and MAPE values for four different models. Figure 1. shows plotted number of confirmed cases vs forecasted case, deaths and recovered cases in Pakistan as a function of time.

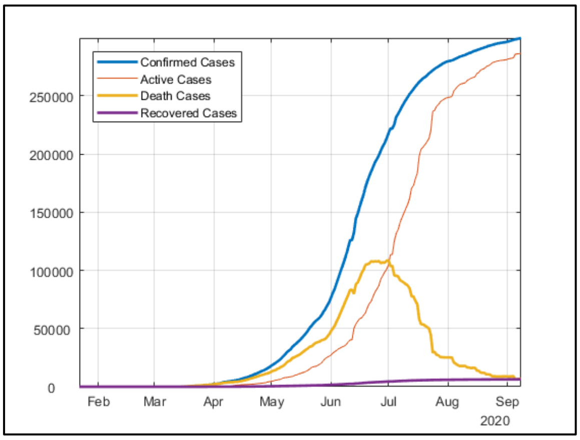


***Fig. 1 Operational model for proposed scheme***

The plot clearly shows that although the increasing curve straightened somewhere around August 2020 but once again it started to increase. On the other hand, graph for deaths due to Covid-19 decreased exponentially and shows no significant increase.

As mentioned above, linear regression model was used to forecast for the next 30 days based on the historical data from January 2020-Spetember 2020. The dataset was split into training and testing data according to 80%:20% ratio, respectively, to ensure that model does not face overfitting. Fig 2 shows performance of the trained model.

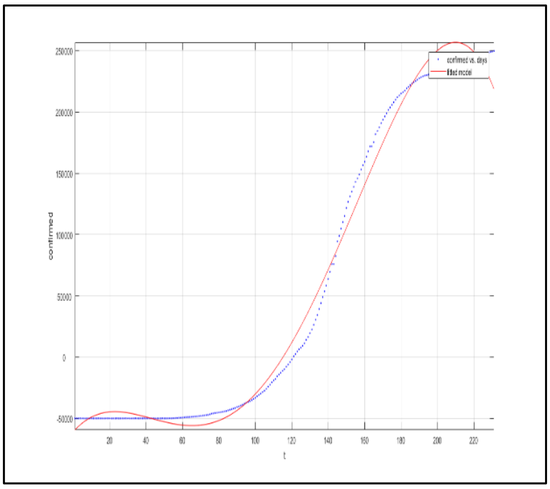

*Fig. 2 Confirmed vs fitted model*

Secondly, the trained model was then used to forecast the number of confirmed cases for the next 30 days as can be seen in the figure 3:

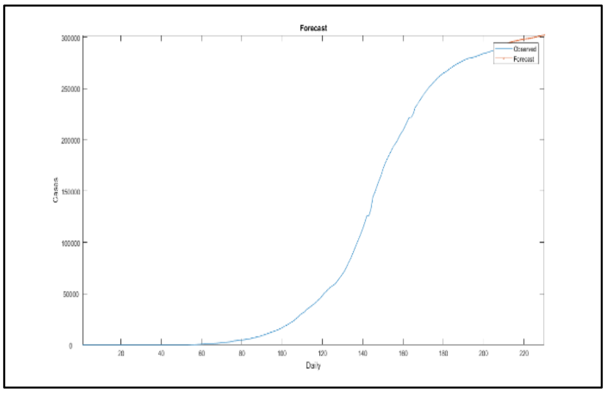

***Fig. 3 Confirmed case forecast***

Fig 4(a) shows the forecast for confirmed cases, fig 4(b) shows the forecast for unrecovered cases and fig 4(c) shows the forecast for recovered cases. These figures correlate the fact that

i. Number of recovered patients is increasing.

ii. Number of un-recovered cases is decreasing

iii. While the number of new confirmed cases is not showing any significant increase, it is progressing at a steady rate.

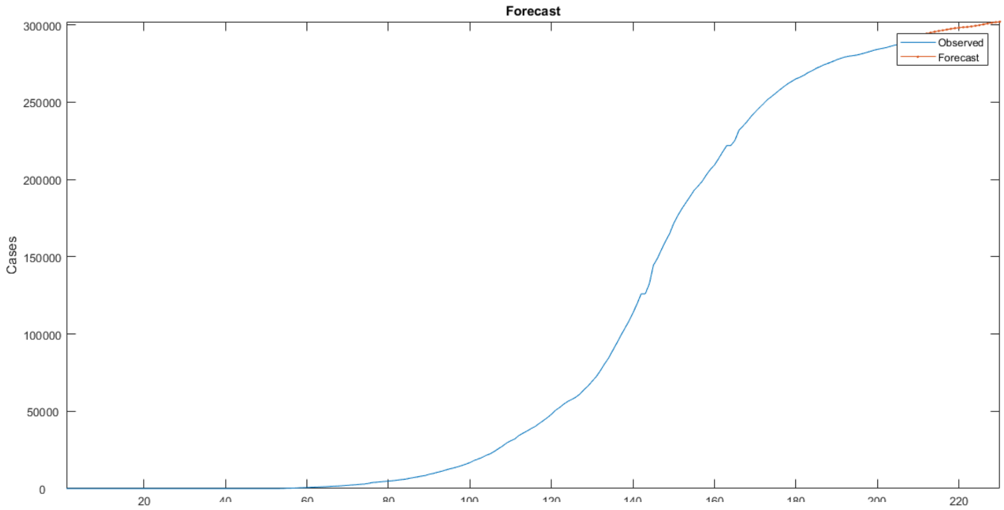


*Fig. 4(a)   Confirmed cases forecast*

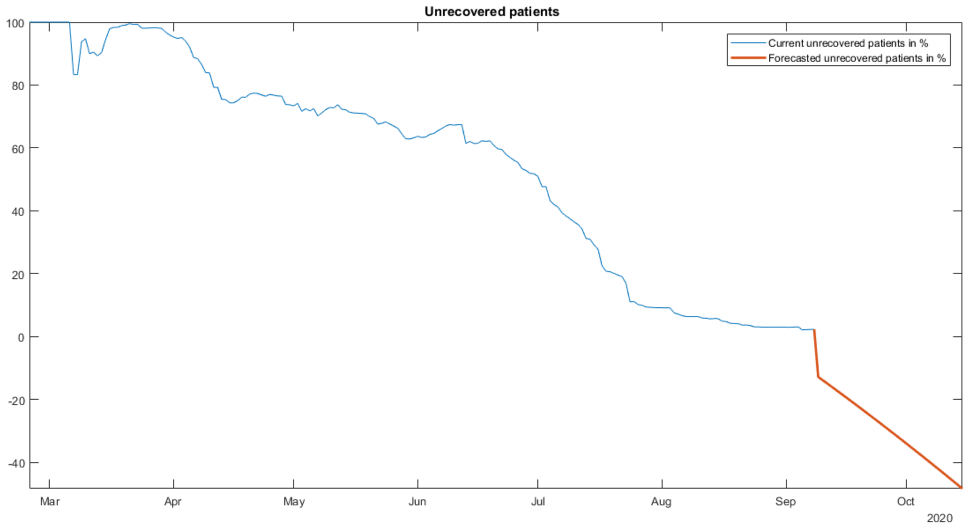


*Fig. 4(b)  Unrecovered cases forecast*

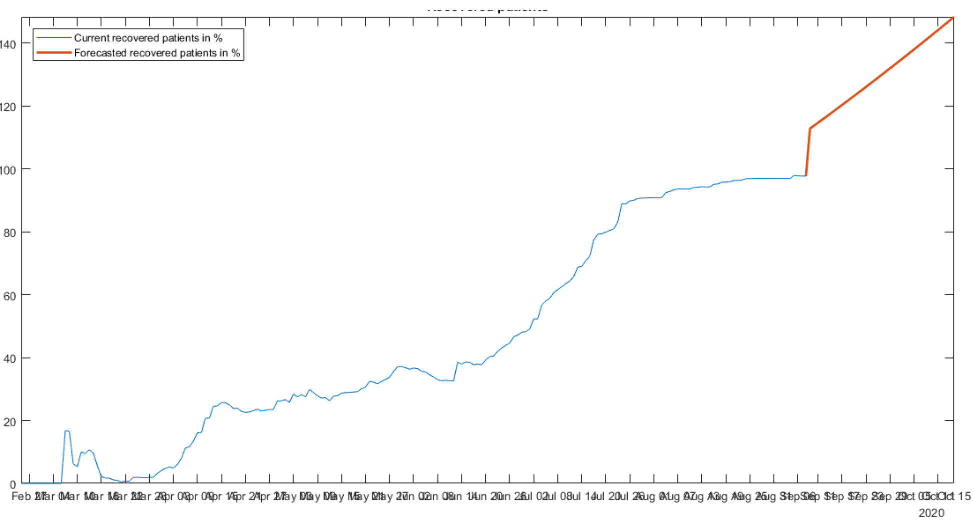


*Fig. 4(c) Recovered cases forecast*

Table 1 shows the comparison of results between the models used in this study to forecast the spread of COVID-19. The data shows that linear regression model (LR) gives far superior results in comparison to other models. By analyzing the parametric values, it can be observed that LR achieves much lower values for RMSE, $R^2$, MSE, MAPE as well as training time for the model. Moreover, the $R^2$ value, which refers to higher level of correlation between the prediction obtained by LR model and the original data set is equal to 1.

*Table 1. Statistical comparison of performance*

| Model | RMSE | R2 | MSE | MAPE | Training Time |
|---|---|---|---|---|---|
| Linear Regression | 1670.72 | 1 | 2.79x104 | 1201.81 | 1.14 |
| Complex Tree | 6229.72 | 1 | 3.88x105 | 3192.93 | 0.44 |
| SVM | 12472.57 | 0.99 | 1.55x106 | 11919.98 | 2.5 |
| Quadratic SVM | 11913.6 | 0.99 | 1.41x106 | 11221.73 | 1.98 |

Fig 5 shows the plot for prediction model developed in this research work. The blue dots represent the actual data whereas the black straight line represents the forecasted data. As both the blue dots and the prediction line overlap each other in most places, it can be deduced that the model is quite capable to predict and has very less number of residual points.

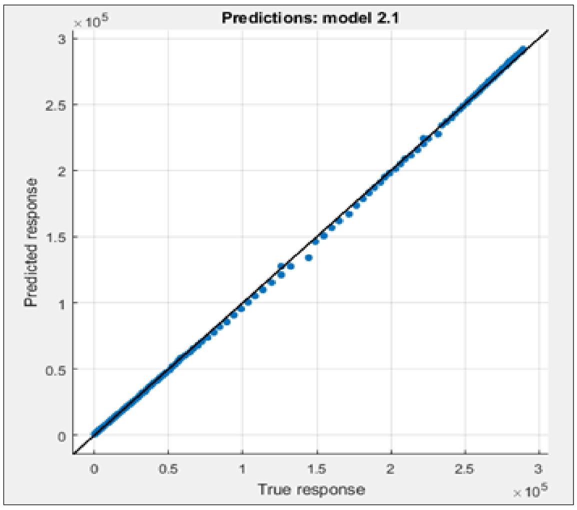


*Fig. 5 Prediction model*

Table 2 shows the forecasted values for confirmed new cases from 14 September 2020 to 23 September 2020. Data shows that although spread of the virus reached its peak somewhere around March and April, but the forecasted results show that the spread of the disease is increasing at a constant rate. The average percentage rise during the forecasted period for new daily cases and new recovered cases is 0.12%. Although the forecasted figures might not necessarily match the exact number of new cases and recoveries, yet the actual trend for increasing new cases is almost the same as forecasted by the LR model used in this study.

*Table 2. Forecasted increase in confirmed and recovered covid-19 cases*

| Date | % Increase in daily confirmed cases | % Increase in daily recovered cases |
|---|---|---|
| 9/14/2020 | 0.14 | 0.14 |
| 9/15/2020 | 0.14 | 0.14 |
| 9/16/2020 | 0.13 | 0.13 |
| 9/17/2020 | 0.13 | 0.13 |
| 9/18/2020 | 0.12 | 0.12 |
| 9/19/2020 | 0.12 | 0.12 |
| 9/20/2020 | 0.11 | 0.11 |
| 9/21/2020 | 0.11 | 0.11 |
| 9/22/2020 | 0.10 | 0.10 |
| 9/23/2020 | 0.10 | 0.10 |

It is observed that average for new daily confirmed cases from January 2020 – 13 September 2020 is 5.12% whereas average increase for daily new recovered cases during the same period is 5.82%. Table 3 shows the actual daily increase in confirmed and recovered cases during the last 10 days.

*Table 3. Percentage increase of confirmed and recovered cases from current data*

| Date | % Increase in daily confirmed cases | % Increase in daily recovered cases |
|---|---|---|
| 9/4/2020 | 0.17 | 0.10 |
| 9/5/2020 | 0.16 | 1.17 |
| 9/6/2020 | 0.13 | 0.04 |
| 9/7/2020 | 0.11 | 0.05 |
| 9/8/2020 | 0.06 | 0.12 |
| 9/9/2020 | 0.14 | 0.50 |
| 9/10/2020 | 0.18 | 0.09 |

| | | |
|---|---|---|
| 9/11/2020 | 0.19 | 0.11 |
| 9/12/2020 | 0.17 | 0.31 |
| 9/13/2020 | 0.18 | 0.13 |

## 4. CONCLUSION

Combatting the recent COVID-19 pandemic has been quite a challenge for computer science experts as limited data and still evolving epidemiological characteristics of the virus have not been fully explained. In this study, a Linear Regression model for time-forecasting the spread of COVID-19 virus is used that is based on the data from Pakistan provided by National Command and Operation Center (NCOC). The used model has a superior ability to predict the number of confirmed cases and deaths for next 30 days. The parametric comparison in table 2 shows that Linear Regression outperforms other models which were evaluated in the study. The non-linearity of the problem makes it very difficult to provide exact time-forecasting results. The used model has a superior ability to predict the number of confirmed cases and deaths for coming 30 days. Based on the results of the model used in this research work, following few suggestions are presented regarding the outbreak:

1. Since the results are a real-time forecast, these results must be taken as indicators of current policies enforced in the country rather than comparing the forecast to exact number of cases.

2. The forecast does not show any significant changes in the trend. However, this might change if different policies are implemented in the countries like strict or complete lockdown.

3 Based on these forecasts, the policymakers can visualize the impact of their policies and can change them accordingly.

4. This forecast can also be helpful for health care planning to mitigate any deteriorating situation in the countries.

In future work, as the size of data increases, the quality of results will also improve. Forecasting of COVID-19 cases can be improved by using Deep Learning model and other ensemble methods when the size of dataset is large enough. Since the pandemic is ongoing, mortality rate cannot be effectively calculated at this stage. But once we have large enough historical data, an effective approach for determining mortality rate can be laid out as it involves actual number of infections and actual number of deaths instead of reported numbers.

## REFERENCES


[1] C. Huang *et al.*, “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,” *Lancet*, vol. 395, no. 10223, pp. 497–506, 2020, doi: 10.1016/S0140-6736(20)30183-5.

[2] J. A. Backer, D. Klinkenberg, and J. Wallinga, “Incubation period of 2019 novel coronavirus (2019- nCoV) infections among travellers from Wuhan, China, 20 28 January 2020,” *Eurosurveillance*, vol. 25, no. 5, pp. 1–6, 2020, doi: 10.2807/1560-7917.ES.2020.25.5.2000062.

[3] J. Wang *et al.*, “High Temperature and High Humidity Reduce the Transmission of COVID-19”.

[4] M. Ahmadi, A. Shari, S. Dorosti, S. Jafarzadeh, and N. Ghanbari, “Science of the Total Environment Investigation of effective climatology parameters on COVID-19 outbreak in Iran,” vol. 729, 2020, doi: 10.1016/j.scitotenv.2020.138705.

[5] B. Pirouz, S. Sha, and B. Pirouz, “Development of an Assessment Method for Investigating the Impact of Climate and Urban Parameters in Confirmed Cases of COVID-19 : A New Challenge in Sustainable Development”.

[6] N. Aydin and G. Yurdakul, “Assessing countries’ performances against COVID-19 via WSIDEA and machine learning algorithms,” *Appl. Soft Comput.*, vol. 97, p. 106792, 2020.

[7] M. van der Schaar *et al.*, “How artificial intelligence and machine learning can help healthcare systems respond to COVID-19,” *Mach. Learn.*, vol. 110, no. 1, pp. 1–14, 2021.

[8] C. P. J. JM Read, JRE Bridgen, DA Cummings, Antonia Ho, “Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions,” vol. 2020, no. December 2019, 2020.

[9] P. Boldog, T. Tekeli, Z. Vizi, A. Dénes, F. A. Bartha, and G. Röst, “Risk Assessment of Novel Coronavirus COVID-19 Outbreaks Outside China,” 2020, doi: 10.3390/jcm9020571.

[10] İ. Kırbaş, A. Sözen, A. D. Tuncer, and F. Ş. Kazancıoğlu, “Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches,” *Chaos, Solitons and Fractals*, vol. 138, 2020, doi: 10.1016/j.chaos.2020.110015.

[11] R. Salgotra, M. Gandomi, and A. H. Gandomi, “Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming,” *Chaos, Solitons and Fractals*, vol. 138, 2020, doi: 10.1016/j.chaos.2020.109945.

[12] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, and R. Gloaguen, “COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach,” *Mathematics*, vol. 8, no. 6, p. 890, 2020.

[13] F. Rustam *et al.*, “COVID-19 future forecasting using supervised machine learning models,” *IEEE access*, vol. 8, pp. 101489–101499, 2020.

[14] K. N. Nabi, “Forecasting COVID-19 pandemic: A data-driven analysis,” *Chaos, Solitons and Fractals*, vol. 139, p. 110046, 2020, doi: 10.1016/j.chaos.2020.110046.

[15] L.-F. Wang, Z. Shi, S. Zhang, H. Field, P. Daszak, and B. T. Eaton, “Review of bats and SARS,” *Emerg. Infect. Dis.*, vol. 12, no. 12, p. 1834, 2006.

[16] Y. Chen, Q. Liu, and D. Guo, “Emerging coronaviruses: genome structure, replication, and pathogenesis,” *J. Med. Virol.*, vol. 92, no. 4, pp. 418–423, 2020.

[17] X.-Y. Ge *et al.*, “Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor,” *Nature*, vol. 503, no. 7477, pp. 535–538, 2013.

[18] S. Cauchemez, M. D. Van Kerkhove, S. Riley, C. A. Donnelly, C. Fraser, and N. M. Ferguson, “Transmission scenarios for Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and how to tell them apart,” *Eurosurveillance*, vol. 18, no. 24, p. 20503, 2013.

[19] R. Lu *et al.*, “Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding,” *Lancet*, vol. 395, no. 10224, pp. 565–574, 2020.

[20] Z. J. Cheng and J. Shan, “2019 Novel coronavirus: where we are and what we know,” *Infection*, vol. 48, no. 2, pp. 155–163, 2020, doi: 10.1007/s15010-020-01401-y.

[21] S. Zhao *et al.*, “Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak,” *J. Clin. Med.*, vol. 9, no. 2, p. 388, 2020.

[22] H. Nishiura *et al.*, “The rate of underascertainment of novel coronavirus (2019-nCoV) infection: estimation using Japanese passengers data on evacuation flights.” Multidisciplinary Digital Publishing Institute, 2020.

[23] B. Tang *et al.*, “Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions,” *J. Clin. Med.*, vol. 9, no. 2, p. 462, 2020.

[24] D. N. Vinod and S. R. S. Prabaharan, “Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19,” *Chaos, Solitons and Fractals*, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110182.

[25] S. Jung *et al.*, “Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: inference using exported cases,” *J. Clin. Med.*, vol. 9, no. 2, p. 523, 2020.

[26] M. Ture and I. Kurt, “Comparison of four different time series methods to forecast hepatitis A virus infection,” *Expert Syst. Appl.*, vol. 31, no. 1, pp. 41–46, 2006.

[27] J. Shaman and A. Karspeck, “Forecasting seasonal outbreaks of influenza,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 50, pp. 20425–20430, 2012.

[28] J. Shaman, W. Yang, and S. Kandula, “Inference and forecast of the current West African Ebola outbreak in Guinea, Sierra Leone and Liberia,” *PLoS Curr.*, vol. 6, 2014.

[29] E. Massad, M. N. Burattini, L. F. Lopez, and F. A. B. Coutinho, “Forecasting versus projection models in epidemiology: the case of the SARS epidemics,” *Med. Hypotheses*, vol. 65, no. 1, pp. 17–22, 2005.

[30] J. B. S. Ong *et al.*, “Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore,” *PLoS One*, vol. 5, no. 4, p. e10036, 2010.

[31] Z. Malki *et al.*, “The COVID-19 pandemic: prediction study based on machine learning models,” *Environ. Sci. Pollut. Res.*, vol. 28, no. 30, pp. 40496–40506, 2021.

[32] K. Nah, S. Otsuki, G. Chowell, and H. Nishiura, “Predicting the international spread of Middle East respiratory syndrome (MERS),” *BMC Infect. Dis.*, vol. 16, no. 1, pp. 1–9, 2016.

[33] J.-S. Jang, “ANFIS: adaptive-network-based fuzzy inference system,” *IEEE Trans. Syst.*

*Man. Cybern.*, vol. 23, no. 3, pp. 665–685, 1993.

[34] M. A. A. Al-Qaness, A. A. Ewees, H. Fan, and M. Abd El Aziz, “Optimization method for forecasting confirmed cases of COVID-19 in China,” *J. Clin. Med.*, vol. 9, no. 3, p. 674, 2020.

[35] W. Xu, H. Peng, X. Zeng, F. Zhou, X. Tian, and X. Peng, “A hybrid modelling method for time series forecasting based on a linear regression model and deep learning,” *Appl. Intell.*, vol. 49, no. 8, pp. 3002–3015, 2019.

[36] D. H. Vu, K. M. Muttaqi, A. P. Agalgaonkar, and A. Bouzerdoum, “Short-term electricity demand forecasting using autoregressive based time varying model incorporating representative data adjustment,” *Appl. Energy*, vol. 205, pp. 790–801, 2017.

[37] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, and S.-P. Guo, “Stock index forecasting based on a hybrid model,” *Omega*, vol. 40, no. 6, pp. 758–766, 2012.

[38] M. Lydia, S. S. Kumar, A. I. Selvakumar, and G. E. P. Kumar, “Linear and non-linear autoregressive models for short-term wind speed forecasting,” *Energy Convers. Manag.*, vol. 112, pp. 115–124, 2016.

[39] T. Chakraborty and I. Ghosh, “Real-time forecasts and risk assessment of novel coronavirus ( COVID-19 ) cases : A data-driven analysis,” vol. 135, 2020, doi: 10.1016/j.chaos.2020.109850.