Experimental Analysis of Covid-19 Spread Predictor using Linear Regression Algorithm

Rakesh Kumar Yadav, Abhay Pratap Mishra, Aman Singh

Abstract: In the past few years, people’s life is affecting badly by the spread of coronavirus due to a lack of information about the spread of the virus and proper management to control it. The government is also looking for ways to get information that how beneficial is their preventive measures. So, that they can know that whether their preventive measures need to be modified or not. The effect of coronavirus can be seen by the number of people affected, the number of people being treated, and the number of people dead. These are the data based on which our application will make a prediction. The goal of this paper is to make a model that will give us a good prediction based on other variables. In most cases, we use linear regression for data because linear regression gives good accuracy. This paper will be helpful for both people and the government, they will be able to predict the number of cases in the next month so that they can prepare themselves to face the problem and control it from further spreading.

Keywords: Covid, Spread, Linear, Regression, Algorithm, Predictor

I. INTRODUCTION

Here in this chapter, we have given a brief introduction of the project like the Machine Learning Algorithm used and how it will be helpful in tackling Covid. As you can see nowadays Covid-19 cases are increasing at an alarming rate and the main causes of this is wrong information, no coordination between the people and government until or unless these problems will be solved we will not be able to tackle this problem of Covid-19. To do so we need proper and accurate information about the functioning of policies introduced to control Corona virus if the policies good enough we will see less number of new cases but if policies are not good enough or if they are not implemented properly we will face more problem that’s why we need to model a tool to predict the number of cases in coming days or months so that government can impose those policies which really needed and will control spread of Covid-19 eventually. Here comes our Project of Covid-19 Spread Predictor which uses the Linear Regression Algorithm (LRA) Model to predict new cases and will provide accurate info.

II. LITERATURE SURVEY

Researchers worldwide have had since the COVID-19 outbreak in January 2020 tried advanced data science and Prediction methods and used it on the present available data from various sources to do the prediction and understand the pattern of Spread of Infection in all other counties or Region in table 1. Attempts to focus include publicly available and frequent updates University Predicted by the many International Universities. Some Predictions focus on the deaths, hospitality needs, infection cases and peaks that they will face in future, whereas others Forecasts focus on the Benefits of travel restrictions, social distancing, and suppression strategies. Restrictions, social distancing, and suppression strategies.

Table 1. Public Source to know the Forecasting of Covid-19 around the world

| Institution | URL |
|-------------|-----|
| University of Geneva, ETH Zürich & EPFL | https://renkulab.shinyapps.io/COVID-19-Epidemic/ |
| Northeastern University | https://covid19.gleamproject.org/ |
| Imperial London College | https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/ |
| University of California, Los | https://covid19.uclaml.org/ |
| Massachusetts Institute of Technology | https://www.covidanalytics.io/projections |
| Los Alamos National Laboratories | https://covid-19.bsvgateway.org/ |
| The University of Texas, Austin | https://covid-19.tacc.utexas.edu/projections |

Here, we have given a summary of the Research papers with their advantages and disadvantages, which we have studied to understand the effect of Corona virus properly and get the idea to do our project efficiently. In ref. [1], authors presented research paper on mental health of scientists in the time of COVID which shows that the world right now need more and more frontline health workers and they have to put their life in danger which does affect their mental health. In ref. [2], Can Hou (2020) Jiaxin Chen presented research paper on the effectiveness of quarantine of Wuhan city against the Corona Virus Disease which shows that how effective was quarantine in wuhan city and how people where affected by it.
In ref. [3], Chengdi Wang presented a research paper on COVID-19 in early 2021: Based on current condition it is expected that during early 2021 more that hundred millions people will be infected and it will cause great problem in people’s life from economic to health. It shows the cases increased since last 12 months and result of our preventive measures. In ref. [4], Precision Immunology Institute at the Icahn School of Medicine at Mount Sinai, USA presented research paper on Immunology of COVID: Current State of the Science. It shows the outcome medical techniques to cope up with the present problem which are proved to be effective in facing covid.

In ref. [5], Einstein and Rosen presented a research paper on the Evolution of vaccine production will be experienced, the vaccine works by boosting our immune system which means it has part of the virus through which our immune system is made familiar to face it. Improvement in our immune response will be enhanced by vaccine. In ref. [6], Farid Rahimi presented a research paper on tackling the COVID Pandemic which is about the Outbreak of Covid in China that needed to be control and strong decisions taken like supporting expediting projects to develop vaccines, immediate quarantine of the involved regions. In ref. [8], Marcus Renato Castro Ribeiro presented a research paper on the spirituality help lines and website is to listen to the problems faced by frontline workers and understand that and help them to cope with it. In ref. [9], Gerard M O’Reilly presented a research paper on The COVID Emergency Department Quality Improvement Project protocol It Includes all patients present to participate in emergency department meeting testing criteria for Covid. The registry has been developed based on which respiratory support will be given to the patients in need. In ref. [10] Tung Thanh Le and Stephen Mayhew presented the research paper on The COVID vaccine development landscape The chances of vaccine being available is in early 2021 totally changes the process of vaccine development as It mostly takes 5-10 years to develop and due to the accelerated rate of vaccine development there could be some side effects.

| Ref. No. | Topic                                                                 | Introduction                                                                 | Advantage                                                                                                                                  | Disadvantage                                                                 |
|---------|----------------------------------------------------------------------|-------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| 1.      | Mental health of scientists in the time of COVID                      | The world right now need more frontline health workers and they have to put their life in danger which does affect their mental health. | It shows that how mental health of any frontline working is affected and what are the problems they are facing.                         | Does not guarantee the safety of the frontline workers.                        |
| 2.      | The effectiveness of quarantine of Wuhan city against the Corona Virus Disease | It shows effectiveness of quarantine in Wuhan city and how people were affected by it.                                                   | Due to quarantine, contact between people was less and It is helpful in controlling Covid. Spread                                         | People cannot be quarantined forever as they will have to go out to fulfil their daily needs. |
| 3.      | COVID-19 in early 2021                                               | Based on current condition it is expected that during early 2021 more that hundred millions people will be infected and it will cause great problem in peoples life from economic to health | It shows how much Cases increased since last 12 months. How much our preventive measures were successful and unsuccessful.            | It only shows past data comparison and how much our preventive measures have been helpful. |
| 4.      | Immunology of COVID: Current State of the Science                    | It shows how Immune is affected and how making our immune system strong enough to tolerate it will help in facing covid is given in this project. | The outcome medical techniques to cope up with the present problem which are proved to be effective in facing Covid.                  | The corona virus’s proper cure is not found yet, there is a chance that even the current vaccine will not work if the new strain will have a change in its chain of the protein |
| 5.      | The Evolution of vaccine production                                   | Evolution of vaccine production will be experienced but the vaccine works by boosting our immune system which means it have part of virus through which our immune system is made familiar to face it. | It shows the new era of vaccine development and immune enhancement.                                                                    | Many uncertainties remain given the lack of robust clinical data so far         |
| 6.      | Tackling the COVID Pandemic                                          | Outbreak of Covid in china needed to be control and strong decisions taken like supporting and expediting projects to develop vaccines, immediate quarantine of the involved regions. | These changes in rules do control the spread of coronavirus cases. Environment is also benefited                                  | Without exact data about the condition in a particular place imposing rules and regulations will not help. |
| No. | Topic                                | Description                                                                                                                                                                                                                                                                                                                                 |
|-----|--------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 7   | Predictive Monitoring of COVID       | The main assumption of prediction based monitoring is that the parameter is dynamic, means the prediction is expected to change over time. These changes in prediction should be viewed as important signals which shows the change in real-world parameter.                                                                                                           |
|     |                                      | The advantages of Prediction Based Monitoring system is that the work on facts and data and there are many techniques devised to help achieve our goal. Data Processing is also important and understanding the problem and sub problem in data and processing it properly is difficult task.                       |
| 8   | The role of spirituality in the COVID pandemic: a spiritual hotline project | These spirituality helplines and website is to listen to the problems faced by frontline workers and understand that and help them to cope with it.                                                                                                                                                                                                 |
|     |                                      | These helplines and websites are helping people from different parts of world to give the mental comfort during crisis.                                                                                                                                                                                                                      |
| 9   | The COVID Emergency Department Quality Improvement Project protocol | Includes all patients present to participate in emergency department and meeting testing Criteria for covid.                                                                                                                                                                                                                         |
|     |                                      | Registry has been developed based on which respiratory support will be given to the patients in need.                                                                                                                                                                                                                                   |
| 10  | The COVID vaccine development landscape | The chances of vaccine to be available is in early 2021 which totally changes the process of vaccine development as it takes so much time to develop vaccine and here we are making it in year                                                                                                                                  |
|     |                                      | It mostly takes 5-10 years to develop certain vaccines and here we are developing it even faster than accelerated rate which is 5 years.                                                                                                                                                       |
|     |                                      | Due to accelerated rate of vaccine development there could be some side effects.                                                                                                                                                                                                                                                          |

III. EXISTING SYSTEM

Some of the studies have tried to validate how accurate any specific prediction method is, some used Data from present covid facing countries. But, most helpful and accurate prediction is introduced by an International University named IHME the number of deaths it predicted was only 50% accurate in ref. [7]. The team later updated the model but the difference in actual and predicted data remains high. The researchers are trying to understand the problem and improve it. Even though it was not much accurate prediction it had already influenced the policymakers in certain ways. The accurate prediction it is a much more valuable thing but it is also difficult to do it under such an uncertain and wicked problem like the Covid, we are focusing more on understanding the potential of predictive monitoring to know more about the uncertainty of the prediction. Whenever we do the continuous monitoring of prediction of crucial events that are going to occur in the future, like when the pandemic life cycle curve will bend and end, with the help of past data we collected. Common monitoring practices report past cases of infection, recovery, and death, and hope that it may stimulate reactive and responsive actions but we don’t that in predictive monitoring. In contrast, predictive monitoring will be beneficial to know more about the effect of Covid in the future so that we will be able to do the planning, implement policies properly and take actions that are needed, and provided resources where it needed the most.

due to government policies, states, Number of days, Number of Waves we need to choose accurate and less dynamic parameter. In this paper, two parameters taken that are the number of days and the number of confirmed cases. Dataset of Covid-19 testing is taken as input and then the data is processed all the missing values are first dropped and then test split of data is done and then linear regression algorithm is applied to dataset and we get the predicted value based on the efficiency of the model we made.

V. ARCHITECTURAL DESIGN

Here, we are going to explore the architectural design of the proposed system and explore, we see the Flow Chart to Predict future Covid-19 Cases in figure 1., which we follow to predict the number of future active cases first comes Problem statement in which we have a precise explanation of the problem. Then we have problem conceptualization in which we understand the subproblems that we will have to solve to get the output we need. In Solution Conceptualization data which is collected from online sources are first processed, missing data is dropped, categorizing data on basis of its correlation. Method Identification is to choose ML algorithms like linear regression Algorithm, Random Forest Algorithm etc., and split the data into two parts 70% to train model and 30% to test model. Then we will fit the model with train data and do the prediction. In Realization of Solution, we visualize the difference between Actual data and Predicted data by plotting two linear graphs and see the difference between them and if the output is satisfactory then we will proceed to get output otherwise repeat from the Solution conceptualization step and try different way as shown in figure 1.
Experimental Analysis of Covid-19 Spread Predictor using Linear Regression Algorithm

Here in below figure 2 shown, we have filtered data and we will split it into two parts that are 70% to train the model and 30% to test the model. Classification of Problem is done which is supervised and distributive. Then we will fit the model with the train data and after that when the model is trained we will start our prediction with the test data which is known as validating the model and after that, we will try to do the prediction with reduced data like by excluding all the useless data.

As shown in figure 3, how a user can use our application. The model is trained, validation of the model is done, and accuracy that is difference between Actual and Predicted value is less, now only left part is to provide the input and get the predicted value. In this User will just have to provide the state and number of days in the future as an output and it will be fitted in the equation and we will get our Prediction of the number of Confirmed cases of the number of days in the future we have filled-in input.

VI. SPREAD PREDICTION USING PREDICTIVE MONITORING EXPERIMENT

In this paper, we used predictive monitoring in real-life circumstances of the current Covid pandemic and to understand its benefits and make particular guidelines to properly use it. To do so first thing that we needed to do was to choose a particular prediction model and a Data Source to get an accurate, recent and large amount of data before doing any prediction and store Future predicted data. The spread of viral disease often follows a pattern, that is from the first few cases or outbreaks to the sudden increase in the spread of infection that is the accelerating of the spread of disease to going to its peak and then eventually start decreasing that is deaccelerating of the spread of disease. These kinds of the pattern of spread of the virus are the result of how are people getting infected, what are the properties of the virus, How are people responding to the spread of infection, their adaptive behaviours and how they are countering it But these patterns of spread of infection are different for different countries and maybe many countries be in a different phase of spread of infection at the same time. Let’s see an example like on 14th April, India had approx. 10,363 confirmed cases, and lockdown got extended for months to control it, whereas on the same day England had approx. 700 cases and there was no lockdown but peoples were taking precautions.
In the best possible way, we can say that these kinds of planning can be done more efficiently when we know where is our country is in pattern or phase of spread of infection as compared to the other countries. Pattern for the spread of infection is the only basis for such estimations.

6.1 Analysis of Models

When we plot the daily cases or months cases in the graph we will notice that the pattern that a spread of virus follows is of "S-shape" or "bell-shape". It is not expected to be symmetrical and also we don't expect both sides of the peak will follow some normal distribution it may have many peaks in between or high or low cases on daily basis but the overall diagram will be similar to the "bell-shape". These kinds of patterns have been seen and well-studied in various domains, like the spread of popularity of new technology or any kind of Software application in society, population growth spread of other infectious diseases and it has a hypothetical mathematics model. Similar to the logistic model that tells the general lifecycle phenomenon which describes the outbreak process of the spread of these infectious diseases. As we can see from below given figure, we can see that same phase which shows that how dynamic are the declared phase but at the same time, they were never in the same country just different states even though they are showing a different pattern that does prove traditional predictive methods will not work.

![Uttar Pradesh](image4)

**Figure 4. Comparison between Predicted and actual value of Uttar Pradesh**

Here in Figure 4 and Figure 5 we can see the Actual Confirmed Cases plot shows similar to “Bell-shaped” pattern for both states Uttar Pradesh and Tamil Nadu and it went to those all phases i.e outbreak, accelerated, peak, and declared phase but at the same time, they were never in the same phase which shows that how dynamic are the parameters we are going to use for predictions. Similarly Predicted Confirmed Cases is a straight line for both States as we are using linear Regression for our prediction but slope of the graph is different due to the difference in spread of the coronavirus. We can also see that there is very small difference between Actual and predicted values and the correlation between the Actual and predicted value is nearly 0.9 which shows the prediction is good and our Model will help predict future cases efficiently.

![Tamil Nadu](image5)

**Figure 5: Comparison between Predicted and actual value of Uttar Pradesh**

We employed the Linear regression model in this project for few reasons as we know that when there are two sets of data and it shows a linear relationship then we can employ linear regression, we can check whether the data is showing the linear relationship or not with the help of Scatter plot or finding the correlation between the two variables, Scatter plot is the way of visualizing data whereas in correlation if it is more than 0.7 then we can say two variables or two sets of data have linear relation.

6.2 Implementation

These two sets of data that is Total Number of days and the Number of Confirmed or Positive cases tells us about the shape of the Curve or life Cycle of the Infection spreading when we will plot the graph. Number of Days and the Number of Confirmed or Positive Cases can be used to find the Slope of a linear equation and mean of two sets of data individually and slope will be used to find the intercept i.e. B0=m_y-B1*m_x here B0 is intercept and m_y and m_x are mean of confirmed cases and Number of days data set and B1 is the slope. The dataset that we collected from various sources does have much more information like states, date, Active cases but we will only need two data sets i.e Confirmed cases and Number of days to train our model, Compare the actual and predicted values, and do our Prediction of Future Cases. The code is written by ourselves for this Project but to get an idea about the use of Data Science using Python we took help from Online Courses. In our Project we are using data of India i.e. we are doing future Confirmed Cases Prediction for overall India and also for all the States of India individually. There have been more efficient models which were used for forecasting the Covid-19 spread like the SEIR model. But these models have more equations and the parameters of the equation also required much more sophisticated data as an input which we can't get easily or you can say that we don't have it right now. Here in this project, Linear Regression is for individual Country or State and is being updated daily with the new data that we get from various sources.

Not all countries or States data will be able to produce the statistically meaningful result from regression. Countries that have satisfactory data which will fit perfectly in the model and will produce statistically meaningful results. The data sets which show the correlation between them more than 0.7 are accepted, processed, and used to train the model and do Future Case Prediction As you can see from the above Graphs.
VII. MAKE SENSE OF PREDICTION CHANGES

As we can see in the above Graphs that the predicted value is showing quite accuracy as they are mostly near to the Actual Values. With the help of the Life Cycle Curve of the spread of Covid-19 one can easily visualize that at which time which phase is going on in which State or Country with the help of actual data collected and used to plot the graph, when is outbreak phase when is the accelerated phase when is infection phase (the peak point in the graph, for those states or Countries which are still in their accelerated phase). And when these were able to control it to some extent that is the end or deaccelerated phase.

This is the only reason our Predictive monitoring is focusing on this kind of transitioning of the spread of the Coronavirus life cycle. When we try to forecast for a whole month instead of a daily basis then we can see the peak value that the infection life cycle will attain during that period which is the infection point of the curve. But predicting the ending point of the spread of coronavirus is not that easy or straightforward and it may vary because of different parameters of different states. As we can hypothetically think of as the end of coronavirus spread will be when there will be last infected person which can vary greatly due to some parameters that is if any vaccine is made that it will not take much time. Anyways as of knowledge of end date of this spread of coronavirus is unpredictable in nature. for ease of mind, one can simply explore the life cycle of the infection as we suggested above. the Pandemic curve need to be continuously re-estimated with help of past collected data, and these predictions that are made by these predictive models or the things that we are visualizing by these Pandemic curves need to be continuously monitored, and that's how when there will be the change in prediction we will be able to understand the changes in real-world over-time. When these kinds of changes are monitored and detected it provides meaning to the predictive monitoring.

In figure 6, we can see that all are the states of the same country even though the effect of coronavirus is different for different states this shows the dynamic nature of our world and how uncertain it is to predict the future cases. This is the reason we need predictive monitoring over traditional monitoring and will have to expect changes in prediction over time. To put it in another way, we can say that our basic expectation from predictive monitoring is that that future prediction of infected people cases will change, most importantly when the real-world scenarios are also changing very rapidly.

The variation we see in the predicted events, like the ending phase, can be assumed as the uncertainty that we face from the real-world scenario. This is the only reason why we see a deviation in the latest actual condition and predicted future infected cases. We can relate this kind of thing with finance when we try to estimate variability related to stock price or the stock market. if the difference between the predicted and actual value is small that definitely means extra parameters are also not changing and vice versa.

![Percent of Population infected per State](image)

Figure 6: The past few months, percentage of Population Infected per Population in each state.

VIII. SUMMARY

We have concluded that the advantage of using predictive monitoring is much greater when it comes to monitoring something which is affected by our uncertain real-world scenarios. We have planned to continuously monitor the life cycle curve of the pandemic, end date and get important insights from the changes in the monitored predictions, to understand fully the potential of predictive monitoring, and also introduce some guidelines for good predictive monitoring practice. Until then readers need to take prediction with caution no matter what the model is or from where the data is collected. If we try to over-optimize the data or model it may result dangerous because it might lose the basic disciplines and can result in the increased spread of virus due to wrong information.

The models that we are using are only hypothetically suited for few waves of the epidemic, and it can only be proved more meaningful when applied to data for every wave individually for the countries or states which have faced multiple waves (like India, US, Canada, etc.).
As we know future is always proven to be unpredictable and we need to keep this in our mind whenever we are reading or doing any prediction. As nobody knew that there will be any Coronavirus outbreak and it will be proven an epidemic. By keeping in mind the unpredictability of the current spread of infection and parameters which are influencing it, the thing which is going to be probably needed is the Flexibility and coordination between people, government, and other organizations which are here to help, that’s how we will be able to deal with such unpredictable epidemic.

**REFERENCE**

1. A. A., “Mental health of scientists in the time of covid...” Brain Behav Immun, 577, pp. 460–461, 2020.
2. J. C. Can Hou., “The effectiveness of quarantine of wuhan city against the corona virus disease...” 122, pp. 448–455, 2020.
3. C. Wang, Z. Wang, “Covid-19 in early 2021: current status and looking forward.” Journal of Aeronautics, Astronautics and Aviation, Series A, 114. 2020.
4. 4.D. Park, L. Pia1 Emma Risson, “Immunology of covid: Current state of the science.” https://doi.org/10.1016/j.immuni.2020.05.002, 69, 617–625, 2022.
5. A. Einstein, B. P. and Rosen, “Evolution of the covid-19 vaccine development landscape.” Journal of Computing in Civil Engineering, 47, 777–780, 2020.
6. A. A. Rahimi., “Tackling the covid pandemic.” Journal of Archives of Medical Research, 51, 468–470 . 2020.
7. J. Luo, “Predictive monitoring of covid.” Data-Driven Innovation Lab, 2020.
8. R. F. D. Marcus Renato Castro Ribeiro, “The role of spirituality in the covid pandemic: a spiritual hotline project.” Journal of Public Health, 42, 855–856, 2020.
9. R. D. M. M. P. Noonan, “The covid emergency department quality improvement project protocol.” Emergency Medicine Australasia., 1, 145–51, 2020.
10. T. Than Le, Z. Andreidakis, and S. Mayhew., “The covid vaccine development landscape.” From The analyst’s Couch, 19, 305–306, 2020.

**AUTHORS PROFILE**

**Abhay Pratap Mishra**, currently doing a Bachelor of Technology (B.tech) degree in CSE from SRM Institute of Science and Technology, Delhi NCR Campus Modinagar, Ghaziabad, Uttar Pradesh. He has done Certificate in Python for Data Science from NPTEL (National Programme on Technology Enhanced Learning), provided by the Indian Institute of Technology Madras, India. He has worked in Sceon Networks India (Pvt.) Ltd. as Software Intern. He is currently working in Cognizant as an Intern in ADM.net. His fields of interest are Data Science, Cyber Security, and ADM.net. Winner of Basketball C.B.S.E CLUSTER IV Tournament Organized by Little Flowers House Nagwa, Varanasi in less than 19 age group in year 2014. Participated in the C.B.S.E National Basketball Championship organized by Shree Vidyamkhetan International School Tirupati, Andhra Pradesh in under 19 age group in year 2014.

**Aman Singh**, currently doing a Bachelor of Technology (B.tech) degree in CSE from SRM Institute of Science and Technology, Delhi NCR Campus Modinagar, Ghaziabad, Uttar Pradesh. He has done the training at Shivakshi Tomifying Education LLP based on CORE JAVA of 2.5 months. He has done three months long Internship based on Mobile Application Development at Shivakshi Technifying Education LLP. He is currently working in Cognizant as an Intern in ADM Java FSE. His fields of interest are App Development, ADM Java, Data Science.

**Dr. Rakesh Kumar Yadav**, received Doctor of Philosophy (Ph.D) degree in CSE from SRM Institute of Science and Technology, Kattankulathur, Chennai, currently working as Assistant professor in Department of CSE at SRM IST, Delhi NCR Campus Modinagar, Ghaziabad, Uttar Pradesh. He received the Master of Engineering (M.E.) degree in CSE from NITTTR (National Institute of Technical Teacher’s Training and Research), Chandigarh, India, and the Bachelor’s degree in Computer Science and Engineering from Dr. Kedhar Nath Modi Institute of Engineering and Technology, Modinagar, Ghaziabad, India. He has more than 14 years teaching experience in the field of Computer Science and Engineering Department in various colleges. He has published more than 30 research papers in journals and conferences. His fields of interest are Wireless Sensor Networks, Mobile Adhoc Network and Internet of Things. He is a member of various professional societies are: IEEE, ISTE, CSI, IACSIT, and IAENG.