IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AUTOMATED DATA ANALYSIS AND PREDICTION OF COVID-19 PANDEMIC USING MACHINE LEARNING TECHNIQUES

[1]Mrs.Thangamma K C, [2]Bhavana N, [3]Gnaneshwari D S, [4]Syeda Sukaina Fatima, [5]Yashaswini S

[1]Assistant Professor, [2,3,4,5]Student

Department of Information Science and Engineering

[1,2,3,4,5]GSSS Institute of Engineering and Technology for Women, Karnataka, India.

Abstract: The coronavirus disease originated from China is spread across world wide. As per the recent count India has the second highest number of cases in the world. Machine learning tools and techniques are widely used to curb the disease. The paper firstly gives an overview of methodologies that can be adopted and the system proposed to design the framework. Our project aims in develop a robust automated model for exploratory data spread analysis and forecasting for COVID-19.

Index Terms **- Artificial intelligence (AI), Machine learning (ML), COVID-19, Coronavirus.**

I. INTRODUCTION

The Covid-19 pandemic has affected our lives through various factors causing deaths worldwide. In Wuhan of China had reported the first case since then the number of cases kept increasing and has now spread to most of the countries affecting lives of the people and thus reducing the population world-wide. This pandemic has led to a huge economic and social devastation. There are many people who are in danger of falling into extreme poverty. Large number of enterprises have been facing existence threat. Almost half of world's 3.3 billion workforce is at risk. There are about 219 countries and territories that have recorded COVID-19 cases. The necessity is to develop right and effective vaccine against COVID-19 virus. The AI, ML and many other technologies have their major role in countering to the novel coronavirus disease. The Artificial Intelligence and Machine learning techniques have contributed to rapid and vigorous research that has provided the capacity to understand virus and its transmission through various applications. Government's efforts in curving the pandemic such as lockdowns, curfews and policies are helping people to cope with the COVID-19 situations. This paper provides a knowledge of frameworks that can be developed using ML techniques to examine the increased spread analysis of infectious disease so that we can predict the future pattern of the disease. We discuss about the possible techniques and provide a survey of most appropriate ML technique that can be used for prediction of the infectious disease. Thus, it gives us important benefits to reduce the infectious virus. An overview of the paper is organized as follows. Section II describes Literature review of papers that provide essential knowledge of the ML techniques and Section III about the various machine learning algorithms. Section IV illustrates the proposed method of the project. Lastly, Section V concludes this paper.

II. LITERATURE REVIEW

The new born coronavirus disease has affected all the parts of the world and has become a pandemic that nobody is unaware of .Even after a long time it is being difficult to tackle with. The technology of today's world has various things to offer .Many developers around the globe have mentioned various ways to tackle with the pandemic technically. Research is going on to evaluate and contain the worldwide disaster of COVID-19 on the human race. Research studies include predictions about the cases of the future, and analysis of the variables responsible for spread of the coronavirus.

[1] The first paper is a base paper,which has it that the future dynamics of the pandemic can be predicted by using different algorithms. The project uses mainly three algorithms-(1) The ANN algorithm (2) Linear regression and (3) Support vector mission regression. By predicting the covid -19 it helps the government to take advance measures .Instead of using one algorithm, the three algorithms combined give a reduced error and improves the accuracy. We have a similar technique but we use (1) ARIMA time series (2) Support vector machine and (3) Linear regression. We will also be giving the government insights as to what might happen

in the future along with an analysis. We will also be recommending the best algorithm out of the three which has the least error using the root mean square method.[2]In this second paper, a survey is presented on the state-of-the-art solutions in the battle against the COVID-19 pandemic. The techniques used here are big data for spatial analysis methods and Geographic Information Systems (GIS) technology which would help collect the data from clinics, government hospitals etc. The technique used for classifying respiratory viruses is the reverse transcription polymerase chain reaction (RT-PCR) detection technique. We collect the W.H.O dataset and use that for the analysis.[3]The third paper is about directing the future by surveying the data. The survey paper mentions the deep learning algorithms for image processing. The images of radiology such as chest X-ray and CT scans are high-dimensional data that require processing capabilities of deep learning methods in which CNN-based models are common and most suitable. A model known as susceptible infected removed model is also used: $dS/dt = -\beta SI$, $dI/dt = \beta SI - \gamma I$, $dR/dt = -\gamma I$. Hence it helps the doctors differentiate between the coronavirus and other respiratory syndromes.

[4]The fourth paper analyzes the distribution of covid-19 in Italy using clustering methods. The paper mentions the use of K-means clustering and hierarchical clustering algorithms. This gives government insights on geographical areas where the corona affected patients are more and where they are less.[5]The fifth paper uses polynomial regression algorithm and support vector machine to analyze the situation of the pandemic. It gives a clear picture of the present state of the spread as well as predicts the future spread possibility of the covid-19 pandemic. It concludes to say that the polynomial regression algorithm gives 93% accuracy which is better than the SVM algorithm. The reason for doing state wise spread analysis is also substantiated along with the facts that describe why India’s mortality state is higher than other countries inspite of being the second most COVID-19 affected country in the world.[6]The sixth paper analyses the spread in turkey.It reviews the SIR model and it properties. A new model named the SPID model has been inspired by it and established. LS based parametric optimization approach is used to as certain the unknown parameters. It conclude that the SPID model can be used which gives the accurate number of covid cases.[7]The seventh paper predicts the COVID-19 future possible cases. It uses various models such as LASSO Regression ,Exponential smoothing(ES) ,Linear regression(LR) and Support vector machine(SVM) .It predicts that the number of death cases increases and the recovered case will decreases. The paper concludes that ES gives the the best accuracy amongst all the models. LASSO and LR also turn out to be reliable. SVM is considered to be the poor model out of all.The paper finally gives the governments a very good documentation and forecast for them to take the corrective measures.[8]In this eighth paper the popular SIR model is used and from this SIRD is developed .It connects the parameters such as recovery state denoted by Y and the mortality rate denoted by S. The government uses zone bifurcation to predict the patients that all most likely to be infected ,but from the SIR model it is inferred that the number of tests conducted to find out the number of infected people have to be increased. The survey also says that the height of danger that is the number of people most likely to lose their life in India is less fatality rate and the lesser value of R0 than other countries.

III. METHODOLOGY

Here we use the WHO Covid-19 datasets and extract India specific data.

Polynomial Regression :

Polynomial regression is also called as linear regression. Here we use this equation on the relationship between the target and independent values. In general, we can model the target value of y as an nth degree polynomial equation represented as:

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \cdots . + \theta_n x^n$$

where, θ_0 is the bias, $\theta_1, \theta_2, \theta_3 \theta_3, \dots, \theta_n$ are the weights in the equation of the polynomial regression, and n is the degree of the polynomial.

Support Vector Regression:

This is another version of support vector machine.This model we give importance on a subset of the training data, because the cost function for joining the model does not care about training points which is lining beyond the margin .

Thus we can write the training the original SVR means solving:

$$minimize\ \frac{1}{2} * \|w\|^2$$

ARIMA Model :

The aims to explain data by using time series data on its past values and uses linear regression to make predictions. It is represented by three parameters: (p, d, q). This model fitting is also called as the “Box-Jenkins” method. Auto regressive parameter (AR (p)) is a parameter which deals with using legacy values in the equation. Here p stands for the no. of lags considered in the model. If p=2, then it is represented as AR (2) or, ARIMA (2, 0, 0) and the equation is:

$$Y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + e_t \qquad (1)$$

where $\emptyset_1 and\ \emptyset_2$ are model parameters. And the parameter d signifies a degree to which the series differs.

$$Y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots .. + \theta_q e_{t-q} + e_t \qquad (2)$$

Combining all the three components - differencing, autoregressive, and moving average the ARIMA model can be represented as:

$$Y_t = c + \emptyset_1 y_{d\,t-1} + \emptyset_p y_{d\,t-p} + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t \quad (3)$$

where y_d is Y differenced d number of times with a constant value c.

IV. PROPOSED SYSTEM

The proposed system uses machine learning techniques, which are performed using the python library to predict the covid-19 outbreak in India. Dataset is collected by world health organization (WHO) .After collecting the data from WHO, formulative analysis is performed to get the data in date-time format. Then by performing the machine learning algorithms, prediction of covid-19 in India. Then we also recommend best algorithm to get accuracy in prediction by using root mean square method.

System Modules:

- **Exploratory data analysis:** Here first we carryout use full exploratory data analysis on the dataset, including a study of covid-19 growth rate. Under pre-processing here, we convert the data into date-time format and group the data under three categories: confirmed, recovered and deaths based on the date. we also add week column to perform weekly analysis further ahead.

- **Predictive analysis:** Here prediction of covid-19 is done using three algorithms .They are polynomial regression, Support vector machine regression and ARIMA time-series forecasting. Here first we predict confirmed cases for near future using the above machine learning algorithms and plot the graphs. We then use Root mean square error as an error metric to rate the prediction. Here we also recommend the best algorithm for the prediction.

- **Vaccination planning:** In vaccination planning, based on the severity of the spread analyzed using our model at the state and district level, proactive steps can be taken to ship the vaccination to state government vaccination hubs. Then state government can supply the vaccine shots to various cities and towns based on the current status and prediction data.

V. CONCLUSION AND FUTURE SCOPE

The COVID-19 pandemic has severely affected people’s life in one or the other way resulting in drastic loss. In this survey we investigated that among most of the published works the three algorithms i.e. Support-Vector algorithm, Linear regression and ARIMA model are mainly used algorithms. Hence the framework we would develop will use Support-Vector algorithm, Linear regression and ARIMA model algorithms for best results. The paper presents the contribution of Machine Learning techniques and how they can be successfully applied.

REFERENCES

[1] Roseline Oluwaseun Ogundkodun ,Joseph Bamidele awotunde , “ Machine Learning Prediction for Covid-19 Pandemic in India” ,2020 ,IEEE.

[2] Quoc-Viet Pham , Dinh C.Nguyen , Thien Huynh-The,Won-Joo Hwang and Pubudu N.Pathirana , “ Artificial Intelligence (AI) and Big Data for coronavirus(COVID-19) Pandemic : A syrvey on the State-of-the-arts” ,2020 IEEE.

[3] Thanh Thi Nguyen , “Artificial Intelligence in the Battle against Coronavirus(COVID-19):A Survey and Future Research Directions”,2020.

[4] Anastasiya Doroshenko , “ Analysis of the Distribution of COVID-19 in Italy Using Clustering Algorithms “, 2020,IEEE.

[5]Ekta Gambhir , Ritika Jain ,Alankrit Gupta , Uma Tomer , “ Regression Analysis of Covid-19 Using Machine Learning” ,2020,ICOSEC.

[6] Onder Tutsoy , Sule Colak , Adem Polat, and Kemal Balikci , “A Novel Parametric Model for the Prediction and Analysis of the COVID-19 Casualties”2020 ,IEEE.

[7] Furqan Rustam ,Aijz Ahmad Reshi , Arif Mehmood . Saleem Ullah , Byung-Won On , Waqar Aslam and Gyu Sang Choi , “COVID-19 Future Forecasting Using Supervised Machine Learning Models”,2020,IEEE.

[8]Siddharth Singh ,Piyush Raj ,Raman Kumar ,Rishu Chaujar , “ Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models”,2020 ,ICIRCA.

[9] H. S. Maghdid, K. Z. Ghafoor, A. S. Sadiq, K. Curran, and K. Ra-bie, “A novel AI-enabled framework to diagnose coronavirus COVID-19 using smartphone embedded sensors: Design study,” arXiv preprint arXiv:2003.07434, 2020.

[10] A. S. S. Rao and J. A. Vazquez, “Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone- based survey in the populations when cities/towns are under quarantine,” Infection Control & Hospital Epidemiology, p. 1–18, 2020.

[11] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang,Y. Li, X. Meng et al., “A deep learning algorithm using CT images to screen for corona virus disease (COVID-19),” medRxiv, 2020.

[12] D. Ivanov, “Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case,” Transportation Research Part E: Logis-tics and Transportation Review, vol. 136, p. 101922, 2020.

[13] V. Rees, “AI and cloud computing used to develop COVID-19 vaccine,” 2020. [Online]. Available: https://www.drugtargetreview.com/news/ 59650/ai-and-cloud-computing-used-to-develop-covid-19-vaccine/

[14] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, “A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact,” IEEE Access, vol. 8, pp.90 225–90 265, 2020.

[15]J. A. Lewnard and N. C. Lo, “Scientific and ethical basis for social- distancing interventions against COVID-19,” The Lancet Infectious dis- eases, vol. 20, no. 6, p. 631, 2020.