Quest Journals
Journal of Medical and Dental Science Research
Volume 8~ Issue 11 (2021) pp: 01-06
ISSN(Online) : 2394-076X ISSN (Print):2394-0751
www.questjournals.org

Research Paper

A Systematic Review of Artificial Intelligence Based Approaches to Map and Forecast the Spread of COVID-19

Shiv Batra, Pragya Verma
[1](Student, DPS International Saket, New Delhi))
[2](Research Mentor, DPS International, Saket, New Delhi)

ABSTRACT: *COVID-19 is the most infectious and deadly pandemic of recent times. Till date there has been no effective medicines released. As modern medicine is not able to cope with the unpredictable nature of COVID-19, new technologies such as machine learning need to be used. Machine learning can be used to understand the nature of this virus and predict the upcoming issues. This paper is a review of machine learning techniques used in forecasting the number of COVID-19 cases and also which is the most significant factor in the spread of COVID-19. We identify the best algorithm's currently being used. Furthermore we discuss the shortcomings of each model and approach and we discuss what has to be done in the future so that machine learning algorithms can be used widely by healthcare authorities.*

KEYWORDS: *COVID-19, Infectious, Algortihm, pandemic, machine learning, Artificial Intelligence*

Received 17 November, 2021; Revised: 29 November, 2021; Accepted 01 December, 2021 © The author(s) 2021. Published with open access at www.questjournals.org

I. INTRODUCTION

COVID-19's pathogen is severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [11]. It is more virulent than previously known viruses such as influenza, severe acute respiratory syndrome (SARS), Middle East Respiratory Syndrome (MERS), and Ebola [12]. COVID-19 has resulted in the death of over 4,800,000 people and losses amounting to trillions of dollars [5]. COVID-19 often spreads through sneezing or coughing, like influenza [19,20]. The first case of COVID-19 was reported from Wuhan, China on 31st December 2019. That was the day which marked the start to one of the most devastating global crises in the recent history of mankind. An International Labor Organization (ILO) report dated 25th January 2021 concluded that the working hour losses in 2020 were 255 million FTE (FTE: Full‐ time equivalent jobs (assuming a 48‐ hour working week) leading to global labor income loss of 3.7 trillion US dollars [6]. According to the UN World Tourism Organization (UNWTO) international tourism and its closely linked sectors suffered an estimated loss of 2.4 trillion US dollars in 2020 due to a steep drop in international tourist activities [7]. As per a WHO report over two decades of progress in the reduction of extreme poverty is now being reversed as an additional 150 million people are being pushed into extreme poverty due to covid-19[8].

All of these figures and statistics are still unable to fully describe how grim and devastating the current scenario is. COVID-19 has brought about an unprecedented situation wherein countries are forced to close their borders, governments are fighting over basic resources and lifesaving vaccines, and people are dying on the footsteps of hospitals due to healthcare facilities and supplies. COVID-19 has always been a few steps ahead of the combined scientific power of the world, so this begs the question how will humanity respond?

In order to respond to this crisis, we need to leverage one of the most cutting-edge modern-day technologies, Artificial Intelligence. In the broadest sense, Artificial intelligence leverages computers and machines to mimic the problem solving and decision-making capabilities of the human mind. Its roots can be traced back to Alan Turing's groundbreaking work; "Computing Machinery and Intelligence" where he posed the question: Can Machines Think?

Artificial intelligence can be used for computational epidemiology, for detection and diagnosis and disease progression. COVID 19 disease trajectory, molecular analysis and drug discovery, and facilitating covid-19 responses through building tools such as an image repository are all under the purview of computational epidemiology.

The scope of this review is limited to research papers that used Machine Learning('ML') for forecasting the number of COVID-19 cases, finding factors which affect the spread, and the adoption of ML for the detection and diagnosis of COVID-19.

1.1 Theme 1: Spread of COVID-19 and its factors

ML as a technology has been proven effective in extracting information from data sets to build a predictive model. This technology can be applied to predict the spread of COVID-19. Machine learning also has the potential to show the relationship between a dependent and an independent variable. This means that the machine learning technology can be used to find the most important factors which determine the spread of COVID-19 and their relative significance.

In this theme we review different papers to see the current state of application of ML models for the short term (daily or weekly) prediction of COVID-19 cases and also to find out which factors impact the spread of COVID-19. The factors which were evaluated included PM 2.5 (it is a measure of pollution as PM 2.5 are particulate matter smaller than 2.5 micrometers which are harmful), population density, per capita GDP, and temperature. Each paper presented different models and approaches. We analyze each approach for their ability to predict the spread of COVID-19 and their limitations. We also discuss how each paper should be
further developed to make a more accurate and reliable model so that it can be used extensive by various stakeholders such as the health care authorities, hospitals etc. We also give a broad summary of the approach each paper took in terms of what type of dataset they had and how they trained their models.

1.2 Techniques for Short Term Prediction of cases

Proposed Models

Different papers have presented various approaches for the short-term prediction of COVID-19 cases. Rustam et al. & Satu et al. used the SIR (Suspected Individuals, infection individuals and removed individuals which includes recovered and deceased) [2,4]. SIR is a common model used in epidemiology to model the spread of a virus. It is a compartmental model as it places everybody into one of the three groups. People may progress with time between these compartments.

Rustam's and Satu's group built many models including linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), exponential smoothing (ES), and prophet. These models were chosen because they are easy to use, and are proven to give accurate forecasting results. For further information on each model refer to appendix 1.

Appendix 1

Regression models are used in machine learning models to find the relationship between a dependent variable and one or more independent variables. [2,4] used different regression models; which are described below.

1.Linear Regression:

Linear regression is a type of regression modeling which depends on 2 values; the dependent variable and the independent variable. Linear regression forms a linear relationship between the dependent and independent variable.

The two variables (x,y) are put in the form: $y = \beta 0 + \beta 1x + \varepsilon$

$\beta 0$: Y intercept $\beta 1$: Slope/Gradient ε: error term of linear regression

The machine learning algorithm is used to find the best values of the Y intercept and the gradient to get the best fit regression line.

2. LASSO

LASSO is a linear regression technique which uses shrinkage. Shrinkage is the process of shrinking extreme values to central values. In order to increase the stability of the model LASSO tries to minimize the sum of the coefficient of variables so that the model is not oversensitive to inputs. It does this through an L1 penalty. An L1 penalty minimizes the size of all coefficients and some to zero effectively removing input features from the model. It decreases the coefficients by checking each variable separately and whether it improves the fit of the line.

3.Exponential Smoothing:

Exponential smoothing is a forecasting method for time series data. It uses weighted averages of past observations to forecast new values. It gives greater weightage to recent values and less to older values.

4.Support Vector Machine:

Support Vector Machine is a supervised machine learning algorithm that can be used both for classification and regression models. In an SVM algorithm each data is plotted in an n- dimensional space where

n is the number of features that you have. It then finds a hyperplane which has the maximum number of points and treats it as the best fine line.

5.Prophet:

It is an additive model for time series data which means it detects the trends and seasonality from the data and then combines them together to get the forecasted value. It is a nonlinear regression model which transforms the data to its natural logarithm form and then estimates a regression model. It is an open-source software released by Facebooks's Core Data Science team.

The dataset which [2] used was obtained from GitHub repository provided by Center for Systems Science and Engineering, Johns Hopkins University. It contained daily time series summary tables including the number of confirmed cases, deaths, and recoveries. [2] had a 56-day training set and a 10-day testing set. In the preprocessing step [2] found the global statistics of the daily number of deaths, confirmed cases and recoveries.

[4] used the same source for its data but only extracted confirmed infection and fatality instances in Bangladesh from 8 March 2020 to 28 November 2020. [4] took an interesting approach where it splits its data into 32 windows of 32 days each. 85% of the last 25 days of the window were used as the training set and 15% was used as a test set. This form of data processing is known as walk forward validation and is the gold standard for time series data as it respects the order of data and tests it in that manner. Time series data is a sequence of data points indexed in time order such as the number of COVID-19 cases each day. The model then gave the forecast for the next 7 days in Bangladesh.

The effectiveness of the models was measured by Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 (Table 1.1 defines each metric and what it represents). These metrics are used to describe how accurately the model has been trained and how accurate its forecasts using the testing set was. Only if a model is proven to be accurate through these metrics can its future predictions be relied on. Rustam provided a more holistic approach in the application of the ML models as they had been used to predict Death Rate, New Infections Confirmed Cases, and Recovery Rate. The Exponential Smoothing (ES) model had consistently the highest R^2 score and the lowest MSE, MAE, and RMSE. Satu concluded that the Prophet model was the most effective as it consistently had the highest R^2 score and the lowest MSE, MAE, and RMSE.

Metric	Meaning
Root Mean Squared Error (RMSE) $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, where Yi and yˆi specifies the data points and predicted data points respectively	It is the standard deviation of prediction errors that shows how close the predicted values are to the real values. Lower it is the better.
Mean Squared Error (MSE) $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ where n, yi and yˆi specifies the number of data points, data points and predicted data points, respectively	It is the average of the square of the difference between actual and estimated values. A lower value shows that the estimated values are close to actual values
Mean Absolute Error (MAE) $MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$, where n, yi and yˆi specifies the number of data points, data points and predicted data points, respectively	It is the magnitude of difference between the predicted value and the true value. The lower it is the closer are the estimated and true values
R^2 $R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$. SSRES and SSTOT indicate the sum of regression and total regression error. yi, yˆi and $\bar{y}$ denotes as data points, predicted data points and mean values respectively.	It is a statistical measure that represents the goodness of a fit of a regression model. If it is 1 then the variables are perfectly correlated.

Table 1.1: This table explains what each metric means and how it is calculated

Apart from the common SIR model Zou et al. presented a new epidemic model known as SuEIR [3]. This acknowledged an overseen fact of the COVID-19 pandemic which is that there are many untested/unreported cases of COVID-19. This feature is not characterized by the classical epidemic model SIR. Zou did not use any normal evaluation metrics but presented data of predicted cases versus the actual cases on those days from the JHU CSSE (John Hopkins University Center for Systems Science and Engineering) and The New York Times showing that the results were accurate as seen in Table 1.2. For further comparison I calculated the RMSE, MSE, and MAE myself as seen in using the equations in Table 1.1. The values are:

RMSE (x10^3): 0.40795658
MSE (x10^3): 0.166428571
MAE (x10^3): 0.334285714

Date (MM/DD)	Prediction of total deaths in the US (×10^3)	Reported number of deaths from the JHU CSSE (×10^3)
05/12	82.60	82.38
05/13	84.10	84.12
05/14	85.57	85.90
05/15	87.01	87.53
05/16	88.42	88.75
05/17	89.79	89.56
05/18	91.13	90.35

Table 1.2: The table provides both the predicted number of cases and the reported number of cases.

Limitations

Satu only used Bangladesh as its sample size so prophet being the best model for forecasting COVID - 19 cases and deaths on a global scale cannot be verified. Rustam and Satu have not shown whether the models consider the mutation of the covid 19 virus, and the usage of the vaccine. New mutations are more virulent and more infectious. The Delta- B.1.617.2 variant was the new variant first detected in India and spreads much faster than other variants and can cause more severe cases than the other variants. As most models were regression based i.e., finding the line of best fit they cannot be used for the new variant. The new variant would have a greater gradient/slope for both number of cases and number of deaths. Furthermore, another problem would be that at the same time both variants can be present in the country. This means that a regression model may not work unless the model treats the type of variant and which variant is predominant in the specific country as a factor.

Vaccines have also been distributed in countries. As result of the vaccine the transmission rate would decrease. The magnitude of decrease depends on the proportion of population vaccinated and the efficacy of the specific vaccine given to the population. In order to give accurate predictions of cases these factors need to be taken in account but have not been.

The training data of Zou only included the USA where extensive testing takes place. As a result, there isn't a significant number of untested/unreported cases in the USA. Until the model in Zou's paper is tested in other countries where there is a large population which is untested/unreported the efficacy of the model cannot be confirmed. .

Future

Machine Learning models need to inculcate other variables such as proportion of population which has been vaccinated, county wise restrictions and curbs on movement and gatherings, the variation in virulence of COVID -19 through feature engineering in order to give real time COVID-19 forecasts. Satu touched upon the possibility of using cloud computing for real time data and this needs to be developed further. Cloud computing would provide safe computation and quick data analysis as the models will be deployed on cloud datacenters. This would allow the model to use real time data on all factors through continuous data input from hospitals and health care centers for more accurate and precise real time forecasting. The SuEIR model presented by Zou is accurate as seen by the RMSE, MAE and MSE score. This means that it should be tested and trained using a

larger sample size including South Asian and African countries wherein lack of testing is a more significant problem.

1.3 Evaluating the factors which determine the spread of COVID-19

Proposed Model

Gupta et al used five different models: Linear Regression, Support Vector Machine (SVM) Linear Kernel, SVM Radial Kernel, SVM Polynomial Kernel, and Decision Tree [1]. Estimated reproduction number through three different methods; time-dependent, exponential growth and maximum likelihood. The estimated reproduction number, PM 2.5, population density, per capita GDP, and temperature, were inputs the machine learning model used to predict the spread of COVID-19. Table 1.3 shows the importance of each factor in the spread of COVID-19.

The metrics used to compare each model were RMSE, R-squared and MAE. The SVM with radial Kernel indicated the best performance with lowest RMSE, and MAE, and highest R-squared. Though it is not mentioned in the paper, I believe that SVM (RK) performed the best because it has the ability to non-linearly map data into a higher dimensional space so unlike linear SVM (LK) it can handle the case when relationship between independent and dependent variable is not linear. SVM (RK) was also more suitable then SVM (PK) as the dataset was not very large.

Using the SVM (RK) the relative importance of predictors was found showing that population density is the most important factor, followed by PM 2.5, daily temperature and then per capita GDP. The relative importance is found by accessing the classifier coefficient on the trained model. Finding the relative weightage is known as feature importance and this can help in getting rid of the factors which are less important which can speed up training, avoid over fitting and ultimately lead to more accurate results thanks to the reduced noise in the data.

Factor	**Effect on Spread**
Reproduction Number	It represents the number of people who will contract a contagious disease from one person with that disease, therefore it is an indicator of how contagious an infectious disease such as COVID-19 is.
PM 2.5	PM 2.5 is a measure of pollution. Pollution can predispose people who have lived in polluted air for decades and can also act as a vehicle for viral transmission therefore it can affect the spread of COVID-19
Population Density	COVID-19 is an airborne disease. This means that people can contract the disease if they are in close quarters with another person through droplets from the infected one's mouth or nose. Higher population density usually means higher transmission of an infectious disease
Temperature	It is a well-known fact that transmission of diseases such as covid-19 can vary due to temperature due to the response of a virus to different temperatures. The virus can prefer either higher or lower temperatures.
GDP per capita	GDP per capita can be a factor in the spread as it can be an indicator of the level of education in a country and also the healthcare facilities available in a country. If a country has a higher GDP per capita usually the spread is less.

Table 1.2: The table describes why each factor has been evaluated and how each factor affects the spread of covid-19

1.4 Limitations

The sample size was only the US and the model compared the reproduction number of different states in the USA. It cannot be said conclusively that population density is the most important factor if this model does not also use data from other countries and compare the reproduction number

Future

More research has to be conducted to accurately forecast the COVID-19 cases using the features importance of the factors. As we saw in the previous theme the factors of the spread were not given priority

when developing the model. If we can use the result from this theme and integrate it into other ML models for the prediction of the spread of COVID-19 then we could build more accurate and precise models.

Lastly other factors such as race, literacy rate, median age, and humidity should also be included in the experiment. Median age would be better than mean age as it would not be skewed depending on whether a country has a top heavy or bottom-heavy population density. All of these factors need to be evaluated as in previous studies by both WHO and CDC these factors can be responsible for the way an infectious disease such as COVID-19 spreads [9,10]. These factors have also proven to be consequential in pandemics as shown by previous research done on pandemics [13,14,15,16,17,18]

REFERENCES

[1]. Gupta, A. and Gharehgozli, A. Developing a Machine Learning Framework to Determine the Spread of COVID-19. SSRN Electron. J. 2020. 1, 1–19.
[2]. Rustam, F., et al, A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS ONE, 2021.16(2): e0245909
[3]. Difan, Z., et al, Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States. MedRxiv, 2020.
[4]. Satu, M.S., et al, Short-term prediction of COVID-19 cases using machine learning models Appl. Sci. 2021
[5]. WHO Coronavirus (Covid-19) Dashboard.
[6]. ILO Monitor: COVID-19 and the world of work. Seventh edition Updated estimates and analysis 2021 January.
[7]. 2021, United Nations Conference on Trade and Development. Covid-19 and tourism an update
[8]. COVID-19 to Add as Many as 150 Million Extreme Poor by 2021. The World Bank 2020
[9]. Morse, S. S. Factors in the emergence of infectious diseases. Emerging Infectious Diseases, 1995. 1, 7-15
[10]. Environment, climate change, social factors and the implications for controlling infectious diseases of poverty. Global report for Research on Infectious Diseases of Poverty. Chapter 2
[11]. Gorbalenya, A., et al, The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2," Nature Microbiology, 2020
[12]. Callaway, E., et al, The coronavirus pandemic in five powerful charts. Nature. 2020;79:482–483.
[13]. Davis, R.E., et al, The impact of weather on influenza and pneumonia mortality in new york city, 1975–2002: a retrospective study," PloS one, 2012 vol. 7, no. 3,.
[14]. Yang, W. and Marr, L.C. Dynamics of airborne influenza a viruses indoors and dependence on humidity, PloS one, 2011 vol. 6, no. 6.
[15]. Jaspers, I. et al, Diesel exhaust enhances influenza virus infections in respiratory epithelial cells. *Toxicol Sci* 2005, 85: 990–1002.
[16]. Kesic, M.J., et al, Exposure to ozone modulates human airway protease/antiprotease balance contributing to increased influenza a infection. PLoS ONE. 2012, 7 (4)
[17]. Lowen, A.C., et al, Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathog. 2007* 3**:**e151.
[18]. Shaman, J. and Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality, Proceedings of the National Academy of Sciences, 2009, vol. 106, no. 9, pp. 3243–3248,
[19]. Yamin, D. and Gavious, A. Incentives' effect in influenza vaccination policy, Management Science, 2013, vol. 59, no. 12, pp. 2667–2686,
[20]. Mamani, H., et al, A game-theoretic model of international influenza vaccination coordination, Management Science, 2013, vol. 59, no. 7, pp. 1650–1670.