Data Analysis and Future Forecasting of COVID-19 pandemic

K. Sreesh Reddy
Computer Science and Engineering
PES University
Bangalore, India
reddy.sreesh224@gmail.com

Anup Nagareddy
Computer Science and Engineering
PES University
Bangalore, India
anupnagareddy@gmail.com

Prathvik Nayak
Computer Science and Engineering
PES University
Bangalore, India
prtvk13@gmail.com

***Abstract*— The outbreak of COVID-19 has affected over 185 countries around the world. The number of infected and deceased patients have been increasing at an alarming rate. It has therefore become necessary for the inculcation of forecasting techniques to develop better strategies and taking preventive decisions. In this study, we focus on in-depth analysis and forecasting techniques to predict the number of cases in the foreseeable future. These predictions might help to prepare against possible threats and consequences.**

Keywords: Analysis and forecasting, COVID-19

INTRODUCTION

It is no new news by now that the COVID-19 (also known as the Coronavirus) has taken the world by storm this year. It originated in Wuhan, China. Barring the lifelong health issues that come along with it, the lockdowns in the midst of the pandemic implemented by several countries around the globe have disrupted the economic and social sectors of countless lives. The virus caused work life to come to a standstill. Symptoms of the virus vary from having no symptoms at all (asymptomatic) to having fatigue, cough, fever, general weakness, sore throat, muscular pain and in the extreme cases, sepsis, severe pneumonia, acute respiratory distress syndrome, and septic shock, all potentially leading to death. Reports show that clinical deterioration occurs quickly, most probably in the 14 days since getting it. Hence, the quarantine rule was kept in place. Many clinical trials have been ongoing to assess the effectiveness of various vaccines but as of now, no solid result is presentable. Due to an insufficient number of test kits, ventilators, oxygen tanks, hospital beds, and the current unavailability of treatment or vaccine, it is essential to analyze the growth rates of the positive cases, number of recoveries, and several other factors affecting the growth of the disease. For example, if the government had an idea of the number of forecasted cases which might occur the next day, they can make preparations accordingly for the necessary medical equipment. In this literature survey, we analyze and visualize the COVID data since visualizations are easily understandable and to forecast the future cases using the present data. Machine learning and Artificial Intelligence methods have recently made their way into the healthcare field and have had a huge impact and thus, helping medical staff in the long run.

LITERATURE SURVEY

A. *Modelling the Spread of COVID-19 infection using Multilayer Perceptron [1]*

The aim of this paper was to achieve a better global model for the spread of COVID-19 virus by using AI algorithms. Hyperparameter tuning is one of the main steps to improve the accuracy of the model. It views all the possible values for the model and chooses the one with the best parameters. Cross-validation is another technique to obtain trained datasets.

The authors analyzed the number of COVID cases based on the longitude and latitude which gives a fair idea about where ground zero was. The model they used was a Multilayer perceptron. The cross-validation method used here is the K-fold algorithm, where the data set is split into k sets and each of them is used as a test set and the remaining (k – 1) sets are used as training sets. They trained the model using a high-performance computer (HPC) which is the Bura Supercomputer. A total of 768 logical CPUs were used. They also tuned the following hyperparameters: solver, initial kernel rate α, adjustment of learning rate. A number of hidden layers, activation function. Regularization parameter L2. They used the coefficient of determination (R2) to determine the quality of the model.

The data has been considered only until March 2020 when the cases in China were at their peak. Model fitting to largely the Chinese patient population shows that using the number of patients per country is not necessarily a good metric to use as a training goal. Also, the MLP method used is complex and less transparent and better techniques such as recurrent neural networks could be applied for the analyses of infection models using time-series data.

B. *COVID-19 pandemic Data Analysis and Forecasting using Machine Learning. [2]*

The research paper and their method are based on the assumption that the predicted number of cases in the month of April would be 5000 but the actual scenario was actually much worse.

The authors have used 4 models here for prediction of future COVID cases.

i) Arima (Auto-regressive integrated moving average) which is dynamic and can be run again and again over a course of the time period. It assumes that there is a linear relationship between time and variables.

ii) Prophet by Facebook, which is mainly designed to manage business problems. It has two techniques and the saturated method fits the best for this problem.

iii) The Linear Regression model trains itself based on historical data and predicts the future considering a linear relationship between the two (in this case, no. of cases and time).

iv) Support Vector Regression, which traditionally has huge forecasting ability is also used here.

The claim they make here is that Classical methods like the Arima model, Facebook Prophet model outperform the machine learning and deep learning models like LSTM. They also claim at the end that the Sigmoid Model is the most accurate since it has the least RMS value.

A few limitations encountered in this paper are that the forecasting techniques used here may not have a proper fit when a global case is considered. This is mainly because they have considered only India and India is considered to have a very poor healthcare system and it might vary around the world. Also, most techniques they use fit well only when there is a linear relationship between time and the variables.

C. COVID-19 Future Forecasting using Supervised Machine Learning Models. [3]

The study demonstrates the usage of Linear regression, LASSO, Support vector machine, and exponential smoothing to forecast key COVID-19 trends. The main predictions of this paper are the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days.

i) Linear Regression: It is used to find out the relationship between independent and dependent variables. And it is most suitable for forecasting and predictive analysis.

ii) LASSO: this method removes the extreme values hence making the model more stable and accurate. It is considered suitable for multicollinearity scenarios.

iii) Support Vector Machines: It is a type of supervised ML algorithm used for both regression and classification.

iv) Exponential Smoothing: In exponential smoothing, forecasting is done based on previous period's data. The past data observations influence decreases exponentially as they become older.

SVM produces poor results in all scenarios because of the ups and downs in the dataset values. For predicting the death rate forecasting and newly infected cases forecasting ES performs the best and SVM the worst in both the situations. The ES model is the best fit for predicting the recovery rate mainly due to the nature of the time-series data. LR and LASSO also perform well for forecasting to some extent to predict death rate and confirm cases.

The authors have not considered a real time scenario and they predict only a few COVID-19 statistics.

PROBLEM STATEMENT

Our main problem statement is to analyze the COVID-19 data from various sources such as the John Hopkins dataset (which gets constantly updated) and have visualizations which are simple to understand for a layman. For example, plotting a bar chart with the number of cases will show that the bar of US and India's is much higher than the rest of the world, showing the seriousness and the impact in those countries is much deeper than others. We then try to forecast the cases by using appropriate models to predict the future cases.

The specific issue we want to address and help with is the unsureness of the cases in the future and helping everyone have an idea of how the graph might look after a while. Machine learning algorithms play an important role in epidemic analysis and forecasting. In the presence of massive epidemic data, machine learning techniques help to find the epidemic patterns so that early action can be planned to stop the spread of the virus.

We also want to address the big problem of the asymmetry of risks and the irrational fear of a pandemic with its possible catastrophic consequences, as happened with the 1918 Spanish flu that killed an estimated 50 million worldwide. Our goal is to not present the model which obtains 100% accuracy but rather help understand the method and the possibility of the existence of prediction and forecasting the cases and accordingly make sense of it.

A Different Approach

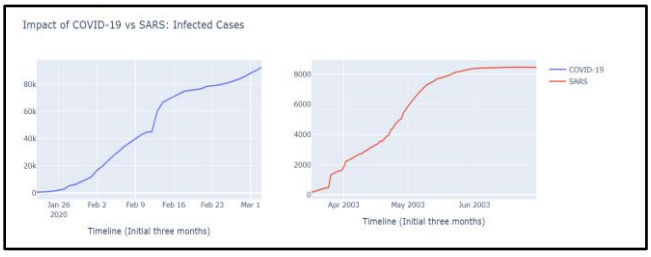
Although the type of problem we have stated here has been seen before, many people have tried to use the preexisting models such as SVR and LSTM. But in the year 1927, Kermack and McKendrick developed a fundamental epidemic model for human-to-human transmission to describe the dynamics of populations through three mutually exclusive phages of infection, namely susceptible (S), infected (I) and removed (R) classes. Mathematical modeling of infectious diseases is now ubiquitous and many of them can precisely depict the dynamic spread of particular epidemics. Several mathematical models have been developed to study the transmission dynamics of COVID-19 pandemic.

Since this model is specifically for a situation where it is used for computing the theoretical number of people infected with a contagious illness in a closed population over time, we will use this model.

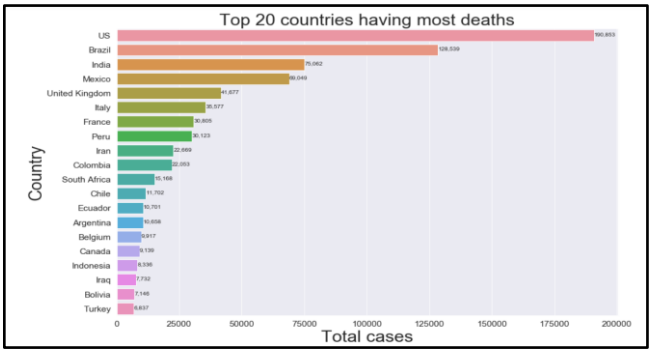
Another challenge or approach we have seen in the papers is that the data was cleaned so much that the data lost its essence and integrity by removing outliers at places where they would be considered necessary. We will try to have data which represents accurate and neat information regarding COVID.

Observations

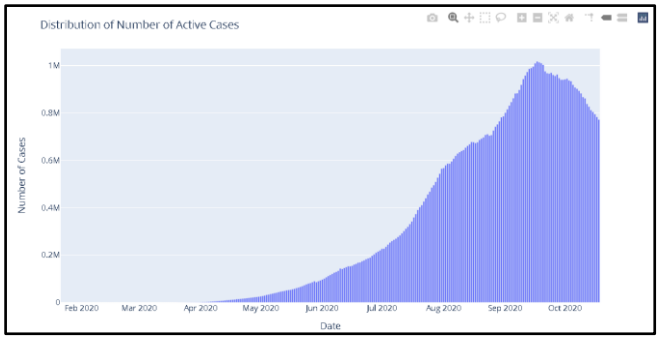
1) Comparison of Growth of COVID-19 and SARS (1918 Plague):



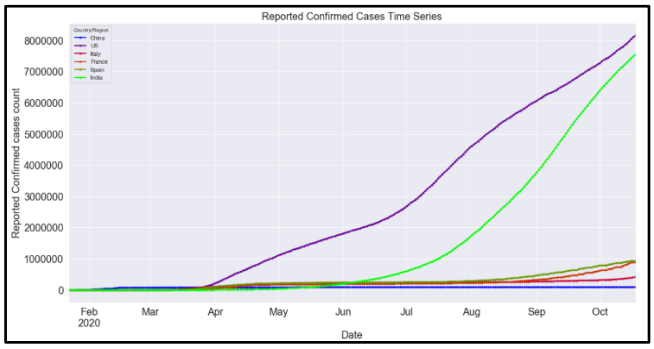
2) Top 20 countries with most deaths:



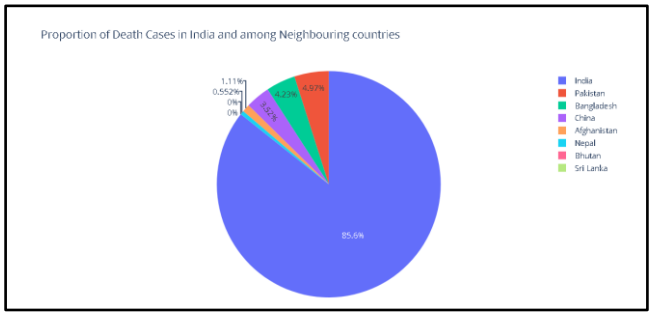
3) Distribution of number of active cases of India:



4) Time Series of Confirmed Cases:



5) Proportion of death in India as compared to its neighboring countries:



REFERENCES

[1] Zlatan Car, Sandi Baressi Segota, Nikola Andelic, Ivan Lorencin, and Vedran Mrzljak: *Modelling the spread of COVID-19 Infection Using a Multilayer Perceptron*
http://downloads.hindawi.com/journals/cmmm/2020/5714714.pdf

[2] Sohini Sengupta, Sareeta Mugde, Dr. Garima Sharma: *COVID-19 Pandemic Data Analysis and Forecasting using Machine Learning Algorithms*
https://www.medrxiv.org/content/10.1101/2020.06.25.20140004v2.full.pdf

[3] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Wawar Aslam, Gyu Sang Choi: *COVID-19 Future Forecasting Using Supervised Machine Learning Models*
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9099302