Covid-19 outbreak prediction using Machine Learning and Deep Learning

Sumana H V [1], Arati Ashok Neelammanavar[2], Sudeep K R[3], Veena R[4] and

Dr Rajashree Shetter[5],

[1-5]Department of Computer Science & Engineering, RV College of Engineering®, Karnataka, India

Abstract- Coronavirus Disease 19 or COVID-19 is a highly infectious pandemic caused by novel Coronavirus (SARS Cov 2). The peculiar characteristic of the virus to swiftly proliferate into human existence coupled with scanty data and understanding has resulted in unprecedented mortalities across the globe. With millions of infections and thousands of deaths being reported on daily basis, recording, analyzing and making sense of data in real time significantly aids in combating the virus.

This study aims to offer an amicable solution to this problem by developing a prediction model to track the infection and mitigate the risk. A comparative study of legacy Machine Learning algorithms is presented, keeping speed and accuracy in focus. Open-source live data on Covid-19 infections is obtained for prediction and results are presented both worldwide and country-wise.

Key Words: COVID-19, Prediction Model, Machine Learning, SVM, PR, Deep Learning, LSTM.

1. INTRODUCTION

The twenty-first century is considered one of the most eventful and path breaking centuries humanity has witnessed yet and going to remain so for eons to come. However, one such event that has had a humongous impact on human livelihood perhaps is the COVID-19 pandemic. A new variant of the then familiar Coronavirus, now called the novel coronavirus or SARS Cov 2 which originated in China was sufficiently quick and deadly to wreak havoc around the world. COVID-19 was declared a worldwide pandemic by the World Health Organization (WHO) on March 11, 2020, with greater rates of infection and mortality than its predecessors, SARS and MERS. As of 10 Jun 21, the virus has infected 17.7 Cr people and caused 38.3L deaths across world. The rate and pattern of infection and death however, has been different in different countries depending on various factors such as location, demographics, healthcare infrastructure and governmental policies.

Although pandemics are not new to the world, technological advancements and availability of modern computing capability have now made it possible to device an accurate prediction of the infection. This enables seamless tracking and elevated preparedness.

2. RELATED WORK

In [1], the research shows that machine learning algorithms can forecast the number of future patients who will be impacted by COVID-19, which is now regarded as a possible threat to humankind. To forecast COVID-19 risks, this study employed predictive models, namely Support Vector Machine (SVM), and Exponential Smoothing (ES). The results of this study prove that ES performs best in the current forecasting domain given the nature and size of the dataset. LR and LASSO are also having equal hands by forecasting to some extent of predicting the death rate and confirm cases. According to the results of these two models, the death rates will increase in upcoming days, and recoveries rate will be slowed down. SVM produces poor results in all scenarios due to ups and downs of the dataset values.

The paper [2] presented a comprehensive study of the spread of the virus outbreak situation in India and also considered world records which will further help in taking necessary steps to manage the huge population of India. For the same, two Machine Learning models were used such as Support Vector Machine (SVM) and polynomial regression models but in future Deep Learning models or hybrid two or models can be

used to forecast the further spread of the virus, by researching the paper the result depicted that the performance of polynomial regression model is much higher than the regression model by 93% approx. of predicting the future cases by one month.

Based on the paper [3], the study was conducted to study how the retrospective model would use the Covid-19 prediction based on time series data. In this study, two machine learning models SEIR and Regression were used to analyze and predict the change in spread of COVID-19 disease. With the help of the SEIR model, the value of R_0 was computed to be 2.84 and also predicted the number of confirmed cases of COVID-19 for the next 21 days starting from 11th May, 2020–31st May, 2020, which was very close to the actual number of cases that happened in India.

In [4], findings from research, comparative analysis of machine learning and soft computer models predict Covid-19 outbreaks. The researcher used an alternative to SIR and SEIR models. The Results depicted in this paper signifies that the SIR model, which has better abilities for long term forecasts. The statistical parameters for this regression model are also satisfied, with a coefficient of determination of R^2 = 0.992, and p-value very close to zero indicating high statistical significance.

One of the most important areas of ML prediction the paper [5], depicted the forecasting models comprising autoregressive integrated moving average (ARIMA), support vector regression (SVR), long shot term memory (LSTM), bidirectional long short-term memory (Bi-LSTM) are assessed for time series prediction of confirmed cases, deaths and recoveries in ten major countries affected due to COVID-19. On the basis of the result, the prediction rate of Bi-LSTM can be exploited for pandemic prediction for better planning and management. The various types of retardation and neural networks have extensive use in predicting future conditions of patients with a specific disease [6].

Essentially, the study focused on the 9 different machine learning (ML) algorithms namely Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Regressor (SVR), Linear Regressor polynomial (LRP), Bayesian Ridge Regression (BRR), Linear Regression (LR), Random Forest Regressor (RFR), Holt-Winter Exponential Smoothing (HW), and Extreme Gradient Boost Regressor (XGB). Among all the models used for the various countries, the highest accuracy achieved was of 99.93% for Ethiopia by using the ARMA model. ARIMA gave an accuracy of more than 85% most of the time for almost all countries. Almost all the models gave an accuracy of more than 80% at least for one of the 10 countries, except in the case of the Philippines.

In [7] the paper presents a comparative study of machine learning methods for COVID-19 transmission forecasting. They have investigated the performances of deep learning methods, including the hybrid convolutional neural networks-Long short-term memory (LSTM-CNN), the hybrid gated recurrent unit-convolutional neural networks (GAN-GRU), GAN, CNN, LSTM, and as well as baseline machine learning methods, namely logistic regression (LR) and support vector regression (SVR). The results depicted that hybrid deep learning models can efficiently forecast COVID-19 cases. Also, results confirmed the superior performance of deep learning models compared to the baseline machine learning models. Furthermore, results showed that LSTM-CNN achieved improved performances with an averaged mean absolute percentage error of 3.718%, among others.

These forecasting systems has become very helpful in making decisions to address the current situation to direct early intervention to reduce the epidemic with great success. Therefore, this paper focuses on a state-of-the-art monitoring of ML techniques and in-depth learning models such as SVM (Support Vector Machine), polynomial regression model and LSTM (Short-term Memory) Machine learning models using COVID-19 patient database provided by Johns Hopkins.

3. Proposed System

In this paper, the COVID-19 data analysis is done for detection and tracking of confirmed, acquired and death cases from live data sets from John Hopkins University [15]. The application includes a pipeline for creating Event-Centric Knowledge Graphs using COVID-19 data, as well as graph statistics for obtaining the best accurate forecasts based on epidemic dynamic model simulation. The efficiency of the training was verified by a survey of 128 nations or regions based on data given by John Hopkins University on COVID-19.

3.1 Input dataset

The initial steps involves of loading the preformatted data to the model for the best prediction but the dataset which has considered is a non-linear dataset of live world record of Covid-19 which includes confirmed cases, recovery cases and death cases with respect to time specific series. The dataset is taken from John Hopkins University.

3.2 Pre-Processing of the input data

This step includes the preprocessing of the input dataset. Preprocessing is being one of the primary steps in model building wherein the imported data has to be filtered through data cleaning, duplicate data removal and data formatting. Data is further divided into two sets as training set and testing set with the ratio of 80:20. Therefore, the pre-processing step is very crucial phase in model building cycle. Figure 1 shows the graph of number of confirmed cases with respect to various countries and figure 2 shows the graph of number of confirmed cases with respect to India

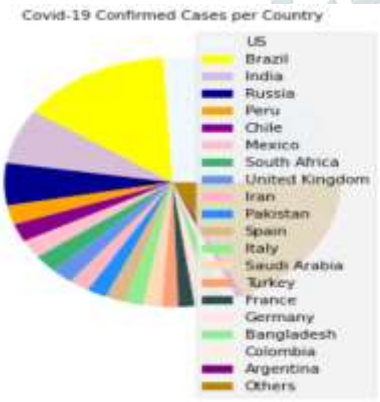

Figure 1: Depicting number of confirmed cases with respect to countries in Pi graph representation.

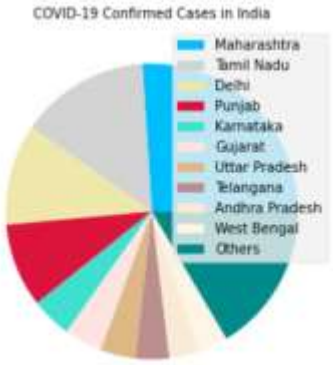

Fig 2: Showing the confirmed cases with respect to Indian States represented in Pi graph.

3.3 Model formation using Machine Learning and Deep Learning

In this phase the intention is to perform the comparison analysis of more than one models to predict the best result from one of the models. Basically, the proposed model comprises the most fitting models as SVM (Support Vector Machine), polynomial regression using 3rd-degree regression and an LSTM-based Deep Learning Model. The best Machine Learning model to adapt the gathered information is the Regression Model since the aim of regression is to create a function that approximates the mapping from the input domain to the actual numbers on the basis of training samples.

ML has recently gained focus and is growing rapidly in solving various problems such as speech processing and identifying an object, image separation, etc. In contrast to other areas of concern, the COVID-19 study with ML has increased dramatically in just two months. This emphasises the need of comprehending the disease's consequences as well as the requirement for enhanced research into intelligent computer approaches.

The prediction of COVID-19 by DL plays a very important role and has been the latest interest in the discovery and prediction of this epidemic. Major factors such as better performance, lack of human involvement in quality releases and recognition make DL techniques a more effective and popular method than ML COVID-19 predictive strategies. Programs for group learning. COVID-19 prediction by DL is absolutely vital, and it has sparked a lot of interest in the discovery and forecast for such pandemic. DL approaches are more successful and popular than ML COVID-19 prediction algorithms because to variables such as superior performance, absence of human participation in quality assurance, and recognition.

With incredibly complicated and deeper information, ML has progressed through time regardless of its revolutionary capacity. As a result of Covid-19 uncertainty, DL may be applied to optimize numerous layers of non-linear data and make direct strategic decisions on a difficult topic.

3.4 Model Functioning

The System architecture shows the flow design of model cycle where the first step is by extracting live Covid-19 dataset directly from the API of John Hopkinson University which consists of world related data. The data set will undergo pre-processing which gets filtered and the required parameters have been extracted from the imported data. The pre-processed data will be directly imported to three training models namely, SVM, 3rd-degree Polynomial Regression and LSTM. The training is done individually on each of the models and performance comparison of each of the model is done by comparing the results. The model which gives the best results is considered further analysis. Figure 3 shows the system architecture of the various models used for analysis of the data in this paper.

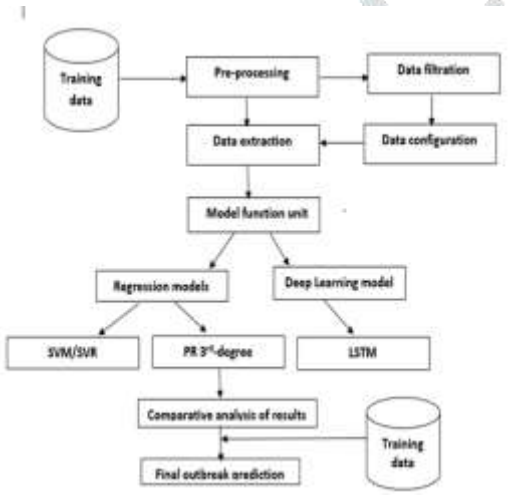

Fig 3: System architecture

Polynomial regression model

Polynomial Regression, also known as nth degree polynomial regression, is a regression procedure that depicts the relationship between dependent (y) and independent (x) variation as shown in the equation below:

$$y = b_0 + b_1x^1 + b_2x_1^2 + b_3x_1^3 + \ldots\ldots b_nx_1^n$$

In the ML, it's also known as a particular instance of Multiple Linear Regression. Because certain polynomial terms must be integrated into the Multiple Linear regression equation in order to transform it to Polynomial Regression. Figure 4 shows the graphs of Simple linear model and Polynomial model

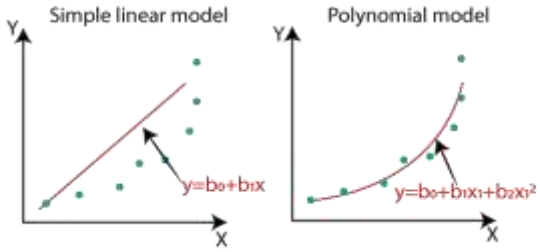

Fig 4: Depicting graphs of Simple linear model and Polynomial model

SVM/SVR Model

The most widely used machine learning algorithm in classification and regression is the Support Vector Machine (SVM). Support Vector Regression is analogous to Linear Regression.

In the line equation below, this straight line is known as a hyper plane in SVR.

$$y = wx + b,$$

Data points on either side of and closest to the hyper plane are the Support Vectors, which are used to draw the boundary line.

In contrast to traditional regression models, SVR tries to equate a better line within the limit value (the distance between the hyper plane and the boundary line), rather than reducing the error between the actual and projected value. Figure 5 shows the SVM graph

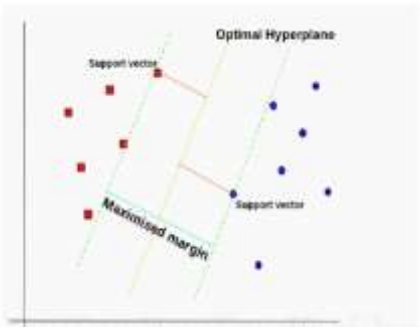

Figure 5: Displaying SVM graph

The kernel function in non-linear regression turns data into greater magnification and generates line separation. Finding the best fit line is the most basic need for Support Vector Regression. The best fit line in SVR is the hyper plane with the greatest number of points.

Deep learning-based LSTM model:

Long Short-Term Memory (LSTM) networks are a modified version of duplicate neural networks that make it simpler to recall information from the past. Below is a solution to the RNN disappearing problem. Given an uncertain time lag, the LSTM is well-suited to categories, analyze, and forecast a succession of events. Back distribution is used to train the model. There are three gates in the LSTM network in figure 6:

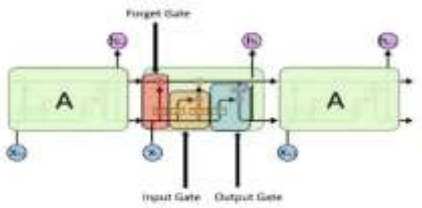

Fig 6 Block diagram of LSTM

The Sigmoid function specifies which values must be allowed to exceed 0.1, and the tanh performance assigns weight to the transferred values, determining their value level from 1 to 1.

The number of the input gate is displayed below.

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i)$$

$$\tilde{c}_t = \tanh (W_c.[h_{t-1},x_t] + b_c)$$

Figure out what data will be discarded in the block. The sigmoid function is used to determine this. Subtracts a number between 0 (leave this) and 1 (keep this) for each number in style Ct-1 from the preceding state (ht-1) and content insert (Xt). Below is a diagram for forgetting the gate.

$$Ft = \sigma(W_i . [h_{t-1}, x_t] + b_f)$$

Input and back memory are utilized to determine the outcome of the output gate. The Sigmoid function decides which values must be permitted to exceed 0, 1 and the Sigmoid output is multiplied by the tanh function, which adds weight to the transferred values that decide their value level from 1 to 1. The output gateway number is shown below.

$$O_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) , \quad h_t = O_t * \tanh (C_t)$$

4. Results and Discussion

The open source data on Covid-19 is obtained through the API from John Hopkins University. This dataset includes the details of the state, country, longitude, latitude, date and number of confirmed cases etc. The processed data is used to train and test the models to obtain the prediction. The LSTM model predicts the feature results more accurately than Polynomial regression and SVM.

Below graphs shown in figure 7 and 8 shows the predicted results for 10 days by using polynomial regression, SVM and LSTM. Table 1 shows the performance of the models used.

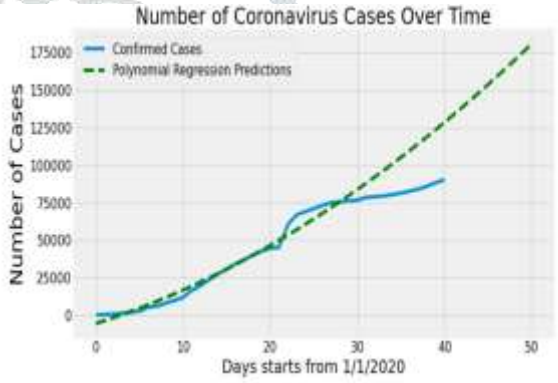

Fig 7: Result prediction from polynomial Regression

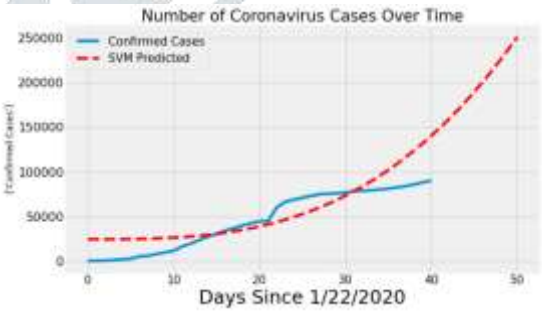

Fig 8: Result prediction from SVM

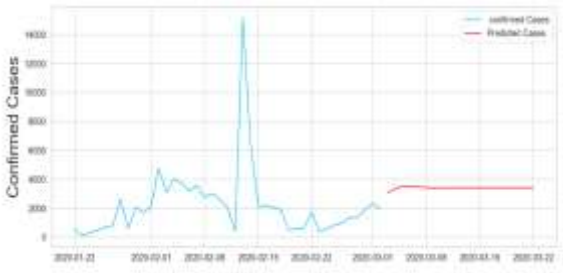

Fig 9: Result prediction from LSTM

Table 1 Performance of models

Model	Metrics	Value
Polynomial Regression	MAE	26463.63
SVM	MAE	26857.69
LSTM	Loss function	47.15

5. Conclusion and Future work

In many nations, the global pandemic of acute respiratory infections Coronavirus 2 (SARS-CoV-2) has turned into a serious humanitarian threat. The creation of reliable emergence prediction models is critical for gaining insight into the disease's prevalence and implications. Standard epidemiological models have demonstrated inadequate accuracy for long-term projections due to a high amount of uncertainty and a lack of critical data. To anticipate COVID-19 outbreaks, research study employs ML comparative analysis with soft computer models. The findings of machine learning models (SVM, PR, and LSTM) showed that they have a lot of potential for long-term performance prediction.

Although estimating the large number of infected individuals is the most challenging forecast, the individual mortality rate is equally significant. The accurate estimation of the number of patients and beds required in highly ill hospitals relies heavily on death measurement. Modeling of death rates may be the most essential aspect in the importance of nations designing new structures in future study. It is aimed to develop existing epidemiological models in terms of accuracy and long lead time by incorporating future research into machine learning and SIR / SEIR models.

References

[1] COVID-19 Future Forecasting Using Supervised Machine Learning Models FURQAN RUSTAM 1, AIJAZ AHMAD RESHI 2, (Member, IEEE), ARIF MEHMOOD 3, SALEEM ULLAH 1, BYUNG-WON ON, 4 WAQAR ASLAM 3, (Member, IEEE), AND GYU SANG CHOI 5 - https://ieeexplore.ieee.org/abstract/document/9099302/authors#authors

[2] Ekta Gambhir, Ritika Jain, Alankrit Gupta, Uma Tomar, "Regression Analysis of COVID-19 using Machine Learning Algorithms" - https://ieeexplore.ieee.org/document/9215356/references#references

[3] N. S Punn, S. K. Sonbhadra and S. Agarwal, "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms", medRxiv, June 2020, [online] Available: https://doi.org/10.1101/2020.04.08.20057679

[4] Siddharth Singh, Piyush Raj*, Raman Kumar, Rishu Chaujar ," Predictions for COVID-19 Outbreak in India using epidemiological models" Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2,[online] Available: doi: https://doi.org/10.1101/2020.04.02.20051466

[5] FarahShahid, AneelaZameer,MuhammadMuneeb
Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM
https://doi.org/10.1016/j.chaos.2020.110212

[6] "Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning"
Aman Khakharia, Vruddhi Shah, Sankalp Jain, Jash Shah, Amanshu Tiwari, Prathamesh Daphal, Mahesh Warang & Ninad Mehendale
Published: 16 October 2020

[7] "Comparative study of machine learning methods for COVID-19 transmission forecasting". Abdelkader Dairia, Fouzi Harroub, Abdelhafid Zeroualcd, Mohamad Mazen Hittaweb ,YingSun
Published: Journal of Biomedical Informatics
Volume 118, June 2021, 103791

[8] P. Ghosh, R. Ghosh and B. Chakraborty, "COVID-19 in India: State-wise Analysis and Prediction", medRxiv, May 2020, [online] Available: https://doi.org/ 10.1101/2020.04.24.20077792

[9] F. Petropoulos and S. Makridakis, ''Forecasting the novel coronavirus COVID-19,'' PLoS ONE, vol. 15, no. 3, Mar. 2020, Art. no. e0231236.

[10] G. Grasselli, A. Pesenti, and M. Cecconi, ''Critical care utilization for the COVID-19 outbreak in lombardy, italy: Early experience

and forecast during an emergency response,'' JAMA, vol. 323, no. 16, p. 1545, Apr. 2020.

[11] WHO. Naming the Coronavirus Disease (Covid-19) and the Virus That Causes it. Accessed: Apr. 1, 2020. [Online]. Available: https:// www.who.int/emergencies/diseases/novel-coronavirus-2019/technic% alguidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virusthat-cau%ses-it

[12] Siddharth Singh, Raman Kumar, Piyush Raj, Rishu Chaujar, 'Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models', International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE Xplore, 93-97, 2020.

[13] Amit Bhati, Anurag Jagetiya, 'Prediction of COVID-19 Outbreak in India adopting Bhilwara Model of Containment', International Conference on Communication and Electronics System(ICCES), 951-956, 2020.

[14] Xiaoyi Fu, Xu Jiang, Yunfei Qi, Meng Xu, Yuhang Song, Jie Zhang, and Xindong Wu, 'An Event-Centric Prediction System for COVID-19', 2020 IEEE International Conference on Knowledge Graph (ICKG), 195-202, 2020.

[15] Farah Shahid, Aneela Zameer, Muhammad Muneeb, 'Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM', Pakistan Institute of Engineering and Applied Science (PIEAS), 1-9, 2020.

[16] COVID-19 Outbreak Prediction with Machine Learning Sina F. Ardabili 1, Amir Mosavi 2,3,*, Pedram Ghamisi 4, Filip Ferdinand 2, Annamaria R. Varkonyi-Koczy 2, Uwe Reuter 3, Timon Rabczuk 5, Peter M. Atkinson 6

[17] https://raw.githubusercontent.com/CSSEGISandData/COVID19/master/csse_covid_19_data/csse_covid_19_time_series/