International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.SP185

Scopus®

Research Article

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Predicting COVID-19 Cases in Indian States using Random Forest Regression

Aravind M, Srinath KR, Maheswari N, Sivagami M

School of Computer Science and Engineering, Vellore Institute of Technology, Vandalur, Kelambakkam Road, Chennai, Tamil Nadu-600127, India.

ABSTRACT

Introduction: Coronaviruses are single-stranded RNA viruses that affect human and non-human mammals and birds. Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. The World Health Organization (WHO) declared this as a pandemic on the 11th of March 2020. Since then, governments have been on war footing imposing lockdowns with various restrictions, ramping up medical infrastructure and creating awareness among the people asking them to wear masks, to follow physical distancing, to wash hands regularly and various other safety measures.

Objective: To forecast the number of cases in the states of Tamil Nadu and Maharashtra for 20 days and visualize the numbers for each state.

Methods: Different approaches using the number of deaths, the number of recovered have experimented, but the results using the number of tests done turned out to be more accurate. The authors have presented a methodology that could predict the number of infections based on the number of tests done using the Random Forest Regressor method. The forecast helps in estimating the spread of the disease and act as a tool for the government to take appropriate actions.

Results and Conclusion: It was concluded that the accuracy of the predictions was heavily dependent on the consistency of the number of tests taken. More consistency, greater accuracy.

Key Words: COVID-19, Prediction, Random Forest Regression, Test Positivity Rate, Tamil Nadu, Maharashtra

INTRODUCTION

The COVID-19 pandemic is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).[1] The first case in India was reported on 30 January 2020 in the state of Kerala. India has the largest number of confirmed cases in Asia and stands only behind the United States of America worldwide.

The virus spreads through small droplets released when people talk, sneeze or cough. The droplets don't travel a long distance. Instead, droplets of the virus fall on the ground, stay suspended in the air or cling on to a surface. A person testing positive can be symptomatic or asymptomatic. In case of asymptomatic case, common symptoms include fever, cough, cold, irritation in the throat, breathing difficulty, loss of taste sense and fatigue. Personal safety measures like Social Distancing, wearing masks and washing hands with soaps or sanitisers (commonly abbreviated as SMS of Corona) are being practised to contain the spreading of the virus. And governments worldwide are imposing lockdowns with various degrees of restrictions.[2] To date, there isn't any vaccine approved by the World Health Organization. Various scientists are working at neck-breaking speeds to come out with a vaccine as sooner as possible.

As of September 28, 2020, there are over 6 million total confirmed cases in India.[3] And around 1 lakh people have succumbed to the virus. Gradually the total daily numbers are increasing and it's an alarming sign. The government has gradually started to 'unlock' in phases. People have started to get back to work and the economy is slowly getting back on track. The pandemic has affected almost all people in some form or the other. Successive lockdowns were imposed that disturbed the nation's functioning on the whole. People were forced to stay at homes to contain the spread.[4] Initially, the people who were returning from abroad tested positive

Corresponding Author:

Maheswari N, Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vandalur, Kelambakkam Road, Chennai, Tamil Nadu-600127, India; Email: maheswari.n@vit.ac.in

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 11.12.2020 **Revised:** 03.01.2021 **Accepted:** 02.02.2021 **Published:** 30.03.2021

for the virus. Later it started spreading to their contacts and eventually it started spreading locally. Forecasting the cases for the next few days helps governments to be better prepared to tackle this pandemic.

In this paper, We have compared two of the most affected states - Maharashtra and Tamil Nadu with 13.4 lakh and 5.8 lakh cases as of September 30, 2020. The Test Positivity Rate (TPR), which is calculated by diving the number of confirmed cases by the number of tests done, daily and cumulative numbers of testing and the confirmed cases between the two states are visualized and interpreted by the authors in this paper.[5] For forecasting of the confirmed cases each day, Random Forest Regression was used. The model was trained with 4 months' data and the prediction was done for 20 days and the results were observed.[6]

Related Work

Arora et al.[2] used various deep learning models such as Bi-Directional LSTM, Convolutional LSTM and Deep LSTM to predict cases in all of India's States and Union Territories. According to the authors, based on the metric 'prediction error', bi-directional LSTM and convolutional LSTM gave the best and worst results respectively. Also, based on daily and weekly predictions, bi-directional LSTM gave better results for short-term prediction.

Farooq and Bazaz et al.[3] proposed an Artificial Neural Network (ANN) based online incremental learning technique to forecast the pandemic in India. The unique feature of this model is that it adapts itself to new datasets by updating its parameters intelligently, continuously. After validating the model with the historical data, a prediction of no. of cases for 30 days was given for the five badly affected Indian states at that time.

Gurucharan et al.[4] worked to predict kidney disease for the clinical data using classification techniques. The causes and effect of disease after the period, from old data, will be predicted using efficient classification techniques.

Ivorra et al.[5] proposed a new statistical model to understand the spread of this deadly disease. It is called the θ-SEIHRD model which used the known special characteristics of the disease - the presence of infectious cases which go/went undetected and the different health conditions of hospitalized people. Particularly, this model considers the fraction θ of detected cases over the total actual infected cases. In this way, the significance, this ratio has on COVID-19, is studied.

Kavadi et al.[6] proposed a Partial Derivative Regression (PDR) and Nonlinear Machine Learning (NML) method for prediction. To identify the best parameters in the dataset in an efficient manner computationally, PDR was used. Later, to make accurate predictions, NML was applied to the normalized features. According to the authors, the proposed methods performed way better than the traditional ML methods.

Poonia and Azad[7] used two models - Holt's second-order exponential smoothing and Auto-Regressive Integrated Moving Average (ARIMA) to do 10-day forecasts of the pandemic in India. Based on the predictions, the authors have also made a few recommendations on the extension of lockdown in different states and is based on the zone in which they fall – red (extremely affected), blue (moderately affected) and green (least affected).

Rath et al.[8] used various regression models such as Linear Regression and Multiple Linear Regression to visualize and to predict the trend of cases for the next few days for the state of Odisha and India. According to the authors, both models showed a very high correlation score. The relationship between the dependent (active) and the independent (positive, recovered and deceased) variables is determined by a strong correlation score.

Rustam et al.[9] used various Machine Learning models such as Linear Regression, Lasso Regression, SVM and Exponential Smoothening to predict the number of COVID-19 cases. Each model predicts the number of new cases, deaths, and recoveries for 10 days. Using various parameters like R-squared score, adjusted R-squared score, MAE, MSE and RMSE, it was found that Exponential Smoothening performed well, followed by Linear Regression, Lasso and SVM.

Sardar et al.[10] developed a model which takes into account the interaction of the following mutually exclusive subclasses: Susceptible, Lockdown, Exposed, Asymptomatic, Symptomatic, Hospitalized, and Recovered. Models with and without lockdown were used to emphasize that in places where there is a higher percentage of symptomatic infections, lockdown would be effective.

Sarkar et al.[11] proposed a new dynamic model to predict the dynamics of the pandemic in India. The model takes into account six types of individuals, namely Susceptible, Asymptomatic, Recovered, Infected, Isolated Infected (I_q) and Quarantined Susceptible (S_q), and is known as SARIIqSq. The results suggested that quarantining the susceptible individuals will, firstly, reduce the contact between infected and uninfected ones and, secondly, effectively reduce the basic reproduction number.

Sujatha et al.[12] proposed a method that compared 3 models - Linear Regression (LR), Multilayer perceptron (MLP) and Vector Auto Regression (VAR) on data taken from Kaggle and implemented using WEKA and Orange. For each model, the predicted number of cases was compared with John Hopkins University (JHU) data. Based on this, the authors concluded that the MLP model performed better than the LR and VAR models.

Tomar and Gupta et al.[13] used methods such as Curve fitting and LSTM to predict the number of COVID-19 cases in India for 30 days. The authors also analyzed various measures like lockdown and social isolation and said that by following these measures, the transmission of the disease can be reduced significantly.

Yang et al.[14] made use of a modified SEIR model to model the spread of COVID-19 in China. The LSTM time series model, which was trained on the 2003 SARS (Severe Acute Respiratory Syndrome) outbreak data, was used to study the transmission trend and to predict the transmission of COVID-19. They found the model helped predict the sizes and peaks of the pandemic.

MATERIALS AND METHODS

Data

Data was taken from https://www.covid19india.org/,[15] a crowd sourced initiative to collect, prepare and maintain the dashboard for all the States and Union Territories of India. It contains district-level data for confirmed cases, recovered cases, active cases and deceased cases and testing data on a state level. The data is updated multiple times a day to make sure that the dashboard is up-to-date.

The data for the two states was taken from the 10th of April 2020 till the 9th of September 2020 using their API. It was then cleaned, preprocessed and converted to a format for convenient use.

Procedure

After experimenting with various algorithms such as Linear Regression, Support Vector Machine Regression and Random Forest Regression, it was found that Random Forest Regression produced better results when compared to the other algorithms. Random Forests[16] is an ensemble machine learning technique. In ensemble methods, a sample of decision trees is taken and features to be used are calculated at each split and those results are aggregated to arrive at a prediction. The Random Forest Regressor model was used for the prediction. It is a supervised learning algorithm that uses an ensemble learning method called bagging for regression.[17]

Consider a training set X containing n samples x1, x2, ..., xn with n responses Y=y1,y2,...,yn. Bagging happens B times and each time a random sample is selected with replacement from the training set and fit into these samples. Predictions for new (unseen) samples x' can be made either by taking the average of all regression trees generated individually on x' or by doing a majority vote in classification problems shown in equation 1.[16]

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b\left(x'\right) \tag{1}$$

In the bagging technique, many different decision trees are made. Each tree is made on different subsets of training data. Classification and Regression Trees (CART) are used for this purpose. They are unpruned and hence there would a slight overfit of the training data. Each tree is different and predictions are less correlated, which means the prediction errors would below.[18]

RESULTS

Forecasting cases state-wise

The model was trained on the data from the 10th of April 2020 till the 15th of August 2020. It was tested on the data from the 16th of August 2020 till the 9th of September 2020.

The results for Tamil Nadu are as shown in Table 1.

Table 1: Prediction for Tamil Nadu

Date	No. of tests done	Actual no. of cases	Predicted no. of cases
2020-08-16	70450	5950	5961
2020-08-17	67532	5890	5883
2020-08-18	67025	5709	5785
2020-08-19	67720	5795	5883
2020-08-20	75076	5986	5902
2020-08-21	74344	5995	5902
2020-08-22	71679	5986	5902
2020-08-23	73547	5980	5902
2020-08-24	70127	5975	5909
2020-08-25	70023	5967	5909
2020-08-26	70221	5951	5961
2020-08-27	75500	5958	5902
2020-08-28	76345	5981	5902
2020-08-29	75103	5996	5902
2020-08-30	80988	6352	5902
2020-08-31	83250	6495	5902
2020-09-01	75100	5956	5902
2020-09-02	75165	5928	5902
2020-09-03	75829	5990	5902
2020-09-04	82901	5892	5902
2020-09-05	83699	5976	5902
2020-09-06	81793	5870	5902
2020-09-07	85974	5783	5902
2020-09-08	80503	5776	5902
2020-09-09	83266	5684	5902

The results from Maharashtra are as shown in Table 2.

Table 2: Prediction for Maharashtra

Date	No. of tests done	Actual no. of cases	Predicted no. of cases
2020-08-16	49384	11111	10130
2020-08-17	43185	8493	8369
2020-08-18	58890	11119	9526
2020-08-19	75427	13165	12618
2020-08-20	76591	14647	12613
2020-08-21	78523	14161	12843
2020-08-22	77400	14492	12817
2020-08-23	45755	10441	9293
2020-08-24	46616	11015	9719
2020-08-25	43186	10125	8369
2020-08-26	87183	14888	11897
2020-08-27	70449	14857	13064
2020-08-28	71006	14427	13087
2020-08-29	77414	16286	12817
2020-08-30	75445	16408	12618
2020-08-31	52503	11852	9517
2020-09-01	52503	15765	9517
2020-09-02	76851	17433	12613
2020-09-03	93986	18105	11897
2020-09-04	91986	19218	11897
2020-09-05	90397	20800	11897
2020-09-06	90177	23350	11897
2020-09-07	56472	16429	9784
2020-09-08	80334	20131	12693
2020-09-09	97735	23577	11897

Visualizing the data

The Test Positivity Rate (TPR) was evaluated for both states and the results are as shown in Figure 1.

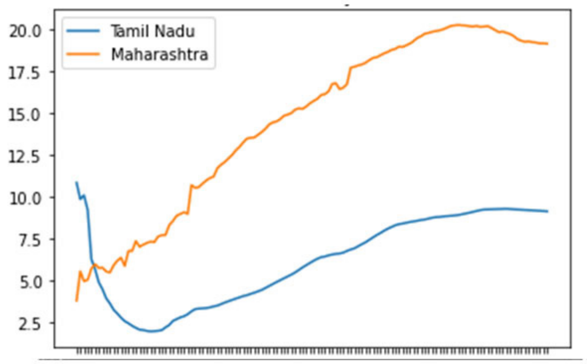

Figure 1: Test Positivity Rate - a comparison.

Initially, the rate was low for Maharashtra and it continued increasing to as high as 20%. Then there are signs of it flattening. Whereas for Tamil Nadu, it was high in the beginning and it slowly started to drop. Then it increased and currently saturated at around 10% for the past month or so. Figure 2 shows a plot of the cumulative cases was made to observe the increasing trend of the curve.

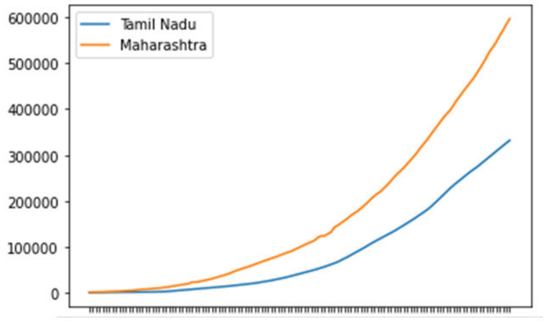

Figure 2: Cumulative daily case numbers in Tamil Nadu and Maharashtra.

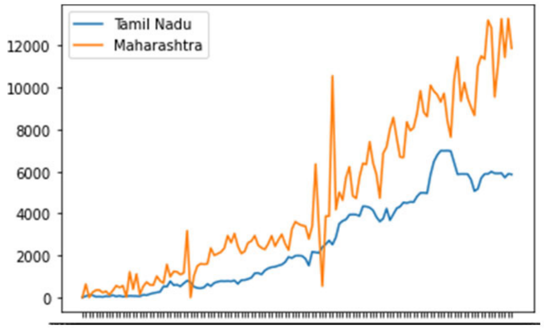

Figure 3: Daily Infected - Tamil Nadu vs Maharashtra.

As one can observe, the exponential growth continues for Maharashtra while towards the end, the graph is almost linear in the case of Tamil Nadu. A graph of the daily cases is shown in Figure 3. In the case of Maharashtra, the general trend seems to be an increasing number of cases daily with the number dipping occasionally due to low testing. Figure 4 shows the comparison of daily testing data of both states.

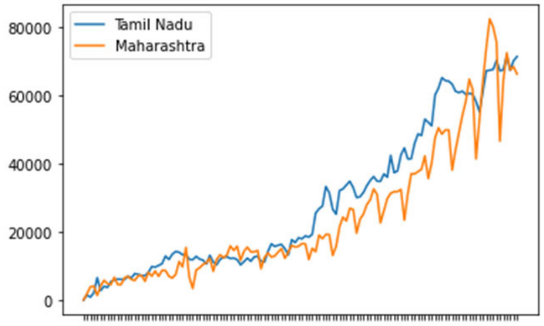

Figure 4: Daily Testing - a comparison.

The low testing number on certain days corresponds to the dips in the daily cases numbers for Maharashtra. Tamil Nadu does more tests than Maharashtra on most of the days but still reports fewer cases. This shows the extent of spread in more in Maharashtra. But overall, both states show an increasing trend in daily testing numbers which is a positive trend. Figure 5 shows the cumulative testing numbers of both states. Figure 6 and 7 show the daily trends - Confirmed, Recovered and Deceased cases in both the states.

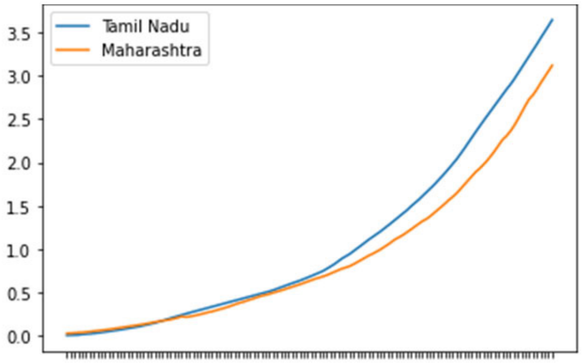

Figure 5: Cumulative Testing and a comparison.

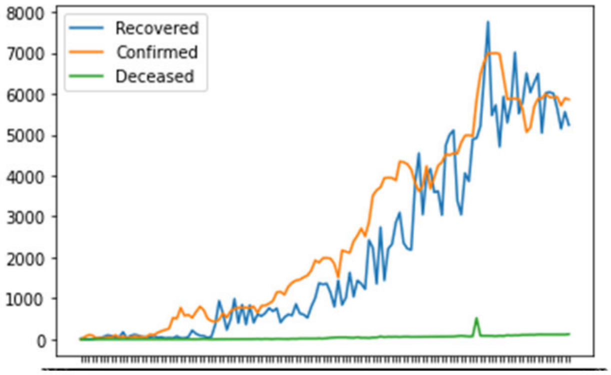

Figure 6: Daily trends in Tamil Nadu.

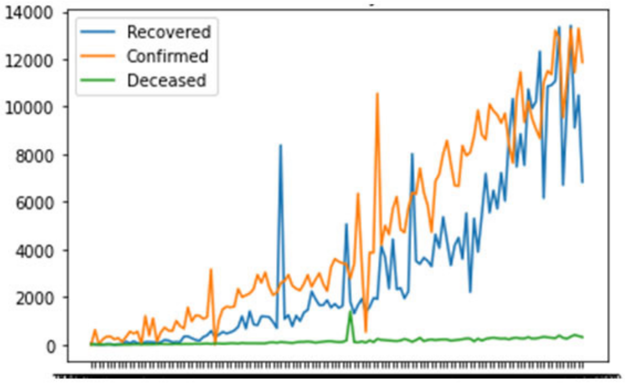

Figure 7: Daily trends in Maharashtra.

DISCUSSION

The Absolute Mean Error observed for Tamil Nadu was found to be 1.87% and that of Maharashtra to be 22.7%. It can be interpreted that the case growth rate was steady overall in Tamil Nadu while Maharashtra showed drastic variations in their testing numbers as well as in the positive cases.[15,16]

As seen in Figure 1, we can see the Test Positivity Rate in Tamil Nadu initially increased and saturated and has remained almost constant throughout which the forecasts also show. For Maharashtra, the peak kept moving up until rates hit almost 20% before it started reducing. The sudden fluctuations in the rates in Maharashtra many times is the reason for a large error and deviation in the forecast.

CONCLUSION

Forecasting helps to a certain extent in determining the probable number of cases provided the growth rate or slowdown rate doesn't change rapidly. Better results can be obtained using Machine Learning techniques if there is district wise testing data. It helps to identify the "hotspot" districts by monitoring the test positivity rate continuously. As of now, we need to be extremely vigil and safe in this crucial situation by social distancing, wearing masks and follow personal hygiene practices to prevent transmission. Currently, the comparison was made only for two states. Extending the same to other states would help to distinguish and analyze how each state has fared. Also, the forecasting is being done with the help of daily testing data state-wise. The accuracy of the forecasts could be improved if further details like district wise testing data and the number of ICU beds occupied, vacant beds, etc. are released by the governments.

Conflict of Interest: Nil

Source of Funding: Nil

Authors Contribution: All the authors have contributed to the research work.

ACKNOWLEDGEMENT

We are thankful to the dataset provider and also like to express our gratitude to Vellore Institute of Technology for providing a working atmosphere and inspiration during the work.

REFERENCES

1. Wikipedia contributors. COVID-19 pandemic in India. 2021 [cited 2020 Sep 20]. https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_India&oldid=1003049780

2. Arora P, Kumar H, Panigrahi BK. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. Chaos Solitons Fractals 2020;139(110017):110017.
3. Farooq J, Bazaz MA. A deep learning algorithm for modelling and forecasting of COVID-19 in five worst-affected states of India. Alex Eng J 2021;60(1):587–696.
4. Singh S, Maheswari N, Comparative study to predict the kidney disease for the clinical data using classification techniques. Ind J Public Health Res Dev 2019;10(8):372.
5. Ivorra B, Ferrández MR, Vela-Pérez M, Ramos AM. Mathematical modelling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. Commun Nonlinear Sci Numer Simul 2020;88(105303):105303.
6. Kavadi DP, Patan R, Ramachandran M, Gandomi AH. Partial derivative Nonlinear Global Pandemic Machine Learning prediction of COVID 19. Chaos Solitons Fractals 2020;139(110056):110056.
7. Azad S, Poonia N. Short-term forecasts of COVID-19 spread across Indian states until 1 May 2020. Preprints. 2020. DoI: http://arxiv.org/abs/2004.13538
8. Rath S, Tripathy A, Tripathy AR. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using the multiple linear regression model. Diabetes Metab Syndr 2020;14(5):1467–1474.
9. Rustam F, Reshi AA, Mehmood A, Ullah S, On B-W, Aslam W, et al. COVID-19 future forecasting using supervised machine learning models. IEEE 2020;8:101489–99.
10. Tridip S, Shahid N, Sourav R, Joydev C. Assessment of lockdown effect in some states and overall India: A predictive mathematical study on COVID-19 outbreak. Chaos Solit Fract 2020;139(110078):110078.
11. Sarkar K, Khajanchi S, Nieto JJ. Modelling and forecasting the COVID-19 pandemic in India. Chaos Solit Fract 2020;139(110049):110049.
12. Sujath R, Chatterjee JM, Hassanien AE. A machine learning forecasting model for COVID-19 pandemic in India. Stoch Envt Res Risk Asses 2020;34(7):1–14.
13. Tomar A, Gupta N. Prediction for the spread of COVID-19 in India and the effectiveness of preventive measures. Sci Tot Envt 2020;728(138762):138762.
14. Yang Z, Zeng Z, Wang K, Wong S-S, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the trend of the epidemic of COVID-19 in China under public health interventions. J Thor Dis 2020;12(3):165–174.
15. COVID19-India API [cited 2020 Sep 20]. https://api.covid19india.org
16. Wikipedia.org. [cited 2020 Sep 20]. https://en.wikipedia.org/wiki/Random_forest.
17. Who.int. [cited 2020 Sep 20]. https://www.who.int/docs/default-source/coronaviruse/key-messages-and-actions-for-covid-19-prevention-and-control-in-schools-march-2020.pdf?sfvrsn=baf81d52_4
18. CDC Labs [cited 2020 Sep 20]. https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/calculating-percent-positivity.html