A Robust Machine Learning Model for Prediction of COVID-19 Pandemic with Climate & Air Quality Parameters

Satya Prakash, Pooja Pathak, Anand Singh Jalal

Dept. of Computer Engineering and Appications
GLA University Mathura - INDIA
write2satyap@gmail.com

Dept. of Computer Engineering and Applications
GLA University
Mathura- INDIA
pooja.pathak@gla.ac.in

Dept. of Computer Engineering and Applications
GLA University
Mathura- INDIA
anandsinghjalal@gmail.com

Abstract—

Most of the studies tries to predict the COVID-19 parameters in isolation. The current study aims to predict the COVID-19 spread in different regions of India in correlation with weather conditions and air quality index. A data-driven, machine learning based approach is employed for accurate long-term prediction. The accuracy of the model is measured on test data by capturing the regression metrics like R2-Score, RMSE and MAE. Air Quality Index seems to be of least effect on COVID-19 fatalities and COVID-19 cases count is showing an increased trend for the moderate temperature range of 25-40 °C for India. The proposed model is able to predict a long-term prediction (180 days) for COVID-19 cases spread and fatalities. It is also able to correlate the impact of weather Conditions and Air Quality Index on COVID-19 spread.

Keywords— *Machine learning, COVID-19, Random Forest, kNN Regressor, Multiple Linear Regressor.*

DOI: 10.48047/ecb/2023.12.Si8.680

Introduction

The world population is being rapidly infected by the SARS-COV-2 virus pandemic, also known as Covid-19 [1]. The novel Coronavirus disease (COVID-19) has been reported to infect more than 113 million people, with more than 2.5 million confirmed deaths worldwide. The first COVID-19 case was discovered in China in December 2019. Post that it has spread in 18 countries by January 2020 and World Health Organization (WHO) declared it as Pandemic on 11th March 2020 [2]. In India, the first COVID-19 case was reported on 30th January 2020 in the state of Kerala [3]. Till now around 10.3 million people are found COVID-19 positive and almost 154000 people have been deceased by this disease [3].

The recent global COVID-19 pandemic has exhibited a nonlinear and complex nature [4]. In addition, the outbreak has differences with other recent outbreaks, which brings into question the ability of standard models to deliver accurate results [5]. There are many studies performed for predicting the spread of disease using various methodologies. Ardabili et al. [6] provided a COVID 19 prediction spread over longer term using MLP and ANFSI methods. It demonstrated that Machine Learning based algorithms are better than mathematical models of SIR and SEIR. Similar work was done by Pinter et al., [7] where they used ML algorithms such as Adaptive network based Fuzzy Inference system (ANFIS) and multi layered perceptron-imperialist competitive algorithm (MLP-ICA) for predicting COVID-19 for Hungary. Chaurasia et al., [8] focused on predicting the number of deaths worldwide due to COVID-19 using timeseries methods like naïve method, simple average, moving average, single exponential smoothing, Holt linear trend method, Holt Winter method and ARIMA. Sujath et al., [9] used dataset from Kaggle and Liner regression, MLP and Vector Autoregression (VAR) methodologies tried to predict the COVID-19 parameters, but the forecasted data in their case does not match with actuals. Dhanwant et al., [10] used SIR model for predicting the COVID cases in India using beta, gamma and alpha as three factors utilizing a custom loss function. Malavika et al., [11] have presented a study comprising of predicting new cases incidences for short term (using Logistics growth curve model), forecast the maximum number of active cases and peak time using SIR model and evaluate the impact of three weeks lock down period using time interrupted regression model for India. Unfortunately, the prediction of peak in India does not match with actual and in fact the country has seen multiple waves and corresponding peaks of COVID cases. Pereira et al., [12] presented a study which uses neural networks for predicting COVID 19 in Brazil. First it tries to use LSTM and LSTM-SAE. But each model fails to give good result. So, it moves to clustering and MAE (Modified Auto-Encoder networks) which seems to have given better result. Wang et al., [13] used LSTM with rolling update for predicting COVID 19 for three countries –

Russia, Peru and Iran. Yadav et al., [14] used support vector regression for predicting the number of COVID cases. It also tries to figure out correlation between different weather parameters and the total number of cases using Pearson's method.

There are many studies on climate conditions and their relationship with the spread of virus. Chan et al., [15] has studied the stability of the virus at different temperatures and relative humidity on smooth surfaces. Wu et al., [16] provided a study to explore the effects of temperature and humidity on the daily new cases and deaths of COVID-19 worldwide using log-linear GAM to analyze the effects. As per their findings Temperature and relative humidity were both negatively related to the daily new cases and daily new deaths of COVID-19. Behnood et al., [17] in their study used a combination of the virus optimization algorithm (VOA) and adaptive network-based fuzzy inference system (ANFIS) to investigate the effects of various climate-related factors and population density on the spread of the COVID-19. Naqvi et al., [18] in their study tries to figure out the impact of reduced AQI levels during lockdown in India on COVID-19 mortality rate. As per the study NO2 and AQI pollutants are positively correlated with SARS-CoV-2 cases and mortalities. Xu et al., [19] in their study on possible impact of AQI on COVID-19 spread in China has confirmed that cases association is statistically significant. Heidari et al. [20] In their study, various medical imaging techniques, including computed tomography (CT) and X-ray, are put to use to give machine learning (ML) a fantastic platform for fighting the Covid-19 pandemic. This necessity has led to a sizable amount of research being done. According to the data, the majority of papers are primarily assessed on qualities like flexibility and accuracy, while other aspects like safety are disregarded. Vega et al. [21] uses the SIMLR model, which combines machine learning (ML) with the epidemiological SIR model, to handle the COVID-19 forecasting difficulty. In order to forecast the number of new infections one to four weeks in advance, the SIR model's time-varying parameters are estimated by SIMLR for each region by monitoring changes in the government-level policies that are in place. Martínez et al. [22] their study evaluates the effectiveness of Verhulst's, Gompertz's, and SIR models in describing COVID-19 behavior in Spain. By first solving the associated inverse problems to identify the model parameters in each wave independently and using the daily instances from the past as observed data, these mathematical models are utilized to forecast the course of the pandemic.

Most of the above studies, tries to predict the COVID-19 parameters in isolation. At one end there are many studies on predicting COVID-19 parameters using Machine learning and Mathematical models. At the other end, there are many studies done to understand statistically the impact of Climate parameters on COVID-19 spread. However, none of the studies have tried to predict COVID parameters in consideration with Climate and Air Quality conditions using Machine Learning. There is no modelling performed with extensive dataset. The present study is a trial to bridge this gap. A unique and extensive dataset is created by merging COVID-19 parameters along with Climate and Pollution parameters. Trusted regression algorithms of machine learning are applied on that dataset for modeling. Prediction is done using that trained model which consists of COVID-19, climate and Air pollution parameters to have accurate results.

I. MATERIAL AND METHOD

A. Dataset Description

For this study, a dataset is created containing COVID – 19 data along with Climate data. Four states of India have been selected for this analysis namely Delhi, Maharashtra, Kerala and Bihar. The COVID-19 data has been collected for these states from the dataset available at https://www.covid19india.org/ [23]. Following attributes has been fetched date wise per state – a) Total number of confirmed cases b) Total number of Recovered cases c) Total number of Deceased Count. Based on these data, the following data fields are calculated and populated d) Total Active Cases (Daily) e) Total Deceased Count (Daily) f) Total New cases (Daily). Further, climate data was collected from https://www.wunderground.com/. [24] Following is collected about the climate attributes date wise per state a) Maximum Temperature b) Average Humidity. Air Pollution data has been fetched from https://aqicn.org [25]. While Fetching the climate and pollution data one specific region is selected from each state as the representative for that state. Care is taken to select a region which has reported maximum percentage of COVID cases. Hence Patna from Bihar, Pune from Maharashtra and Thiruvananthapuram from Kerala has been selected. The complete data is collected date wise starting from the day the first COVID case reported in that state till January 2021. Approximately the dataset contains 300 days of COVID data along with climate parameters for all 4 states.

B. Selection Criteria of Multiple States

While analyzing COVID-19 spread in India, it has been figured out that different regions of India has shown different trends. This may be because of different demographics or different climate conditions prevailing. So, it was obvious to take one representative state from each corner of India – North, South, West and East. While selecting a particular state in each region, following parameters have been considered – a) Population Density b) Number of COVID tests performed c) Case Fatality percentage d) Recovery Percentage e) Test Positivity percentage. In North region Delhi was selected because it has a very high number of testing performed (62.4 Lakhs) [23] and having the highest population density (11297) [26]. Also, Delhi has shown multiple waves of COVID-19 during these months. In Western region, Maharashtra was selected because of being a major state having population density (365) [23] as good as national average (382) [26]. Also, it has highest COVID-19 test positivity percentage (16.8%) [23] in this region and also having the higher case fatality percentage (2.6%) [23] among the major states of India. In Eastern region, Bihar was selected because of its high population density (1102) [23] and having lowest case fatality percentage (0.5%) [23], lowest test positivity percentage (1.6%) [23] and good recovery percentage (97.1%) [23]. In Southern region, Kerala was selected because of high population density (859) [26], Lowest case fatality percentage (0.4%) [23] still having good test positivity percentage (9.6%) [23]. The idea was to select states with diverse nature both in terms of COVID-19 parameters and Climate conditions.

C. Selection of Climate Parameters

While selecting the Climate parameters, those parameters have been considered which can have an impact on spread of viral disease like COVID-19. In discussion with Medical professionals and Experimental research on related viruses found indeed an impact of virus transmission on temperature and humidity (Chan et al., 2011) [27]. There are studies Prata et al., 2020 [28] which has indicated significant impact of Temperature in sub-tropical countries of COVID-19 transmission. Pulmonary disorders like pneumonia are the prominent symptoms of Sars-COV-2 virus. Pulmonary disorders are directly related to Air pollution and AQI (Air Quality Index). Studies like Naqvi et al., 2021 [18] have shown a correlation between the COVID-19 vulnerable regions and AQI hotspots, thereby suggesting that air pollution may exacerbate clinical manifestations of the disease. AQI, PM2.5, NO2, and temperature are figured out as the four important parameters that could promote the transmission of COVID-19 as per the study of Li et al., 2020 [29]. Due to these reasons, Temperature, Humidity and AQI is selected as important parameters for ascertaining the spread and increase / decrease of total number of COVID-19 cases.

D. Selection of Regression Algorithms

The prepared unique COVID-19 dataset along with Climate parameters can be viewed as a typical regression problem. E.g. The daily new cases of COVID can be think of dependent of temperature, humidity and AQI (Air quality Index) prevailing on that day. Hence in present study following regression algorithms are considered for predicting COVID-19 parameters. a) Support Vector Machine Regressor, b) Random Forest Regressor, c) Multiple linear Regressor and d) kNN Regressor

Support Vector Machine Regressor

Support vector Machines are well known for the classification problems. SVM tries to find a line/hyperplane (in multidimensional space) that separates the two classes. Then it classifies the new point depending on whether it lies on the positive or negative side of the hyperplane depending on the classes to predict. Support Vector Regression (SVR) uses the same principle as SVM, but for regression problems.

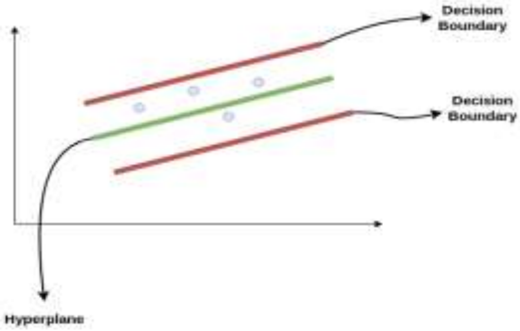


Figure 1: Basic SVM boundaries and hyperplane

Consider these two red lines as the decision boundary and the green line as the hyperplane. In SVR, basically the points that are within the decision boundary line are considered. The best fit line is the hyperplane that has a maximum number of points. SVR gives us the flexibility to define how much error is acceptable in the model and will find an appropriate line (or hyperplane in higher dimensions) to fit the data. The objective function and constraints are as follows:

$$\text{Minimize} \quad : \text{Min } \tfrac{1}{2} \|w\|^2$$
$$\text{Constraints} \quad : |y_i - w_i.x_i| \leq \mathcal{E} \tag{1}$$

Further support vector regression (SVR) is characterized by the use of kernels. Support Vector Machines are used for time series prediction and compared to radial basis function networks by Muller et al., [30]. In this work, SVR is used after the evaluation and measuring the confidence scores from linear, poly, rbf and sigmoid kernels.

KNN Regressor

KNN algorithm [31] can be used for both classification and regression problems. The KNN algorithm uses ‘feature similarity’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. As a method, the distance between the new point and each training point is calculated. Then the closest k data points are selected (based on the distance). Afterwards k value is selected. This determines the number of neighbors should be looked at when a value to any new observation is assigned. In regression problem, the average of k data points is taken. In order to select optimum value of k, consider minimizing the error between train and validation set. Following distance metrics can be used

1. Minkowski distance

$$d(X,Y) = (\sum_{i=1}^{n} | x_i - y_i |^p)^{1/p} \tag{2}$$

2. Euclidean Distance

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

3. Manhattan Distance

$$d(X,Y) = \sum_{i=1}^{n} | x_i - y_i | \tag{4}$$

Random Forest Regressor

A Random Forest [32] is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement. And hence Random Forest has a high accuracy than other algorithms. Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x' [33]

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{5}$$

Multilinear Regression

Multiple linear regression (MLR), also known simply as multiple regression [34], is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon \tag{6}$$

y = the predicted value of the dependent variable
B_0 = the y-intercept (value of y when all other parameters are set to 0)
B_1X_1= the regression coefficient (B_1) of the first independent variable (X_1)
B_nX_n = the regression coefficient of the last independent variable
e = model error

E. **Methodology**

First of all, dataset is prepared for all four states as described in section II A. Dataset contains almost 300 days of COVID-19 data for each state from the day the first case has been reported in that state. Climate data – Temperature, Humidity and AQI are also captured for all these dates. Dataset is further divided into Training (90%) and Test (10%) sets. Data partitioning is done randomly.

After data set preparation, Models have been trained using all four selected algorithms – SVR, MLR, RFR and kNN Regressor. The effectiveness of model is measured with confidence score. Model is evaluated with following regression parameters - R- Squared score, Root mean square error, Mean Absolute error. The prediction of selected parameters has been done for future dates using the trained model.

Below parameters has been chosen for Prediction of COVID -19 along with Climate conditions (Temperature, Humidity and AQI)

a) CASE-1- Measuring Accuracy of models by predicting "Total Number of COVID -19 Positive cases" per region wise

b) CASE-2- Prediction of "Total COVID-19 Active cases on daily basis" per region wise

c) CASE-3- Prediction of "Total Daily Deaths due to COVID-19" per region wise dependent on temperature and humidity

d) CASE-4- Prediction of "Total Daily Deaths due to COVID-19" per region wise dependent on Air Quality Index

e) CASE-5- Prediction of "Total New COVID-19 cases arriving on daily basis" per region

The overall methodology of learning and prediction is depicted in figure (2). The Climate and Air quality data has been taken from weather websites on per day basis. These data fields (Maximum temperature, Average Humidity and AQI) is then merged with COVID-19 parameter already arranged in chronical sequence. The data is preprocessed and divided into Training and validation sets. Regression model is trained using the training dataset and confidence score is measured in each iteration.

Once the model confidence score is attained at best level, test data set id fed into the model. The predicted value is compared with actual values and different regression metrics like MAE (Mean Absolute error), R2 score and RMSE (Root Mean Square Error) is calculated. Post that a dataset is created for future dates e.g. a data record of 1st March 2021 is created with historical Climate and AQI parameters for that location. Now this dataset is used for predicting COVID-19 parameters on future dates against the already trained machine learning models.

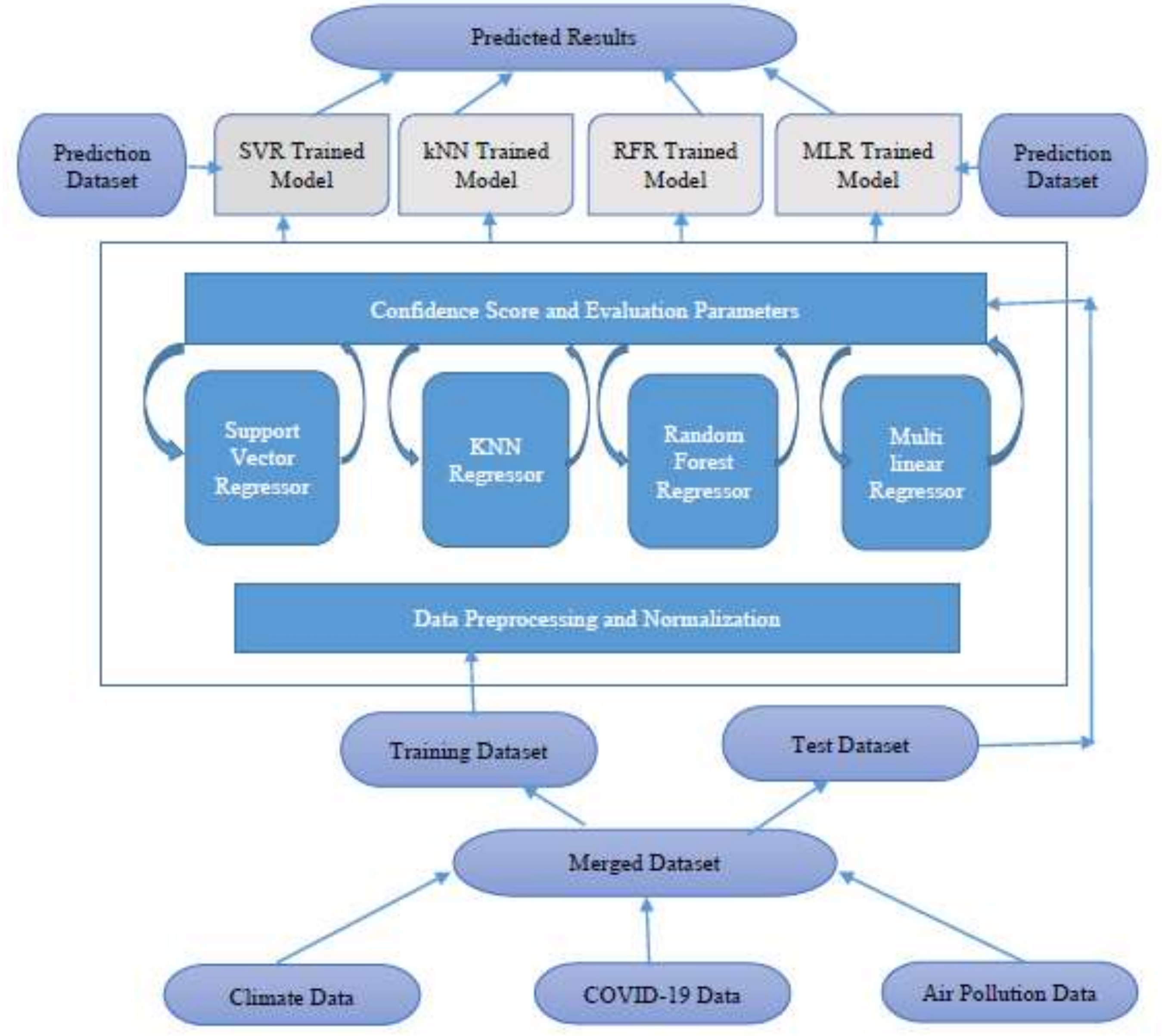


Figure 2: Learning and prediction process of Machine learning regressors

II. RESULT AND DISCUSSIONS

A. Dataset

Separate dataset has been prepared for all four selected regions. For state of Delhi the date ranges for dataset is from 2nd March 2020 till 26th Dec 2020. The same for the state of Maharashtra is from 09th March 2020 till 02nd January 2021, state of Kerala is from 30th January 2020 till 26th January 2021 and for the state of Bihar is from 22nd March 2020 till 21st January 2021. For CASE-1, following were the data fields – Date, Days Count, Maximum Temperature, Average Temperature, AQI and Total Number of COVID -19 Positive cases on that date. For CASE-2, the data fields were - Date, Days Count, Maximum Temperature, Average Temperature, AQI and Total COVID-19 Active cases on that date. For CASE-3, the data fields were - Date, Days Count, Maximum Temperature, Average Temperature and Total Daily Deaths due to COVID-19. For CASE-3, the data fields were - Date, Days Count, AQI and Total Daily Deaths due to COVID-19. For CASE-5, the data fields were - Date, Days Count, Maximum Temperature, Average Temperature, AQI and Total New COVID-19 cases arriving on that date.

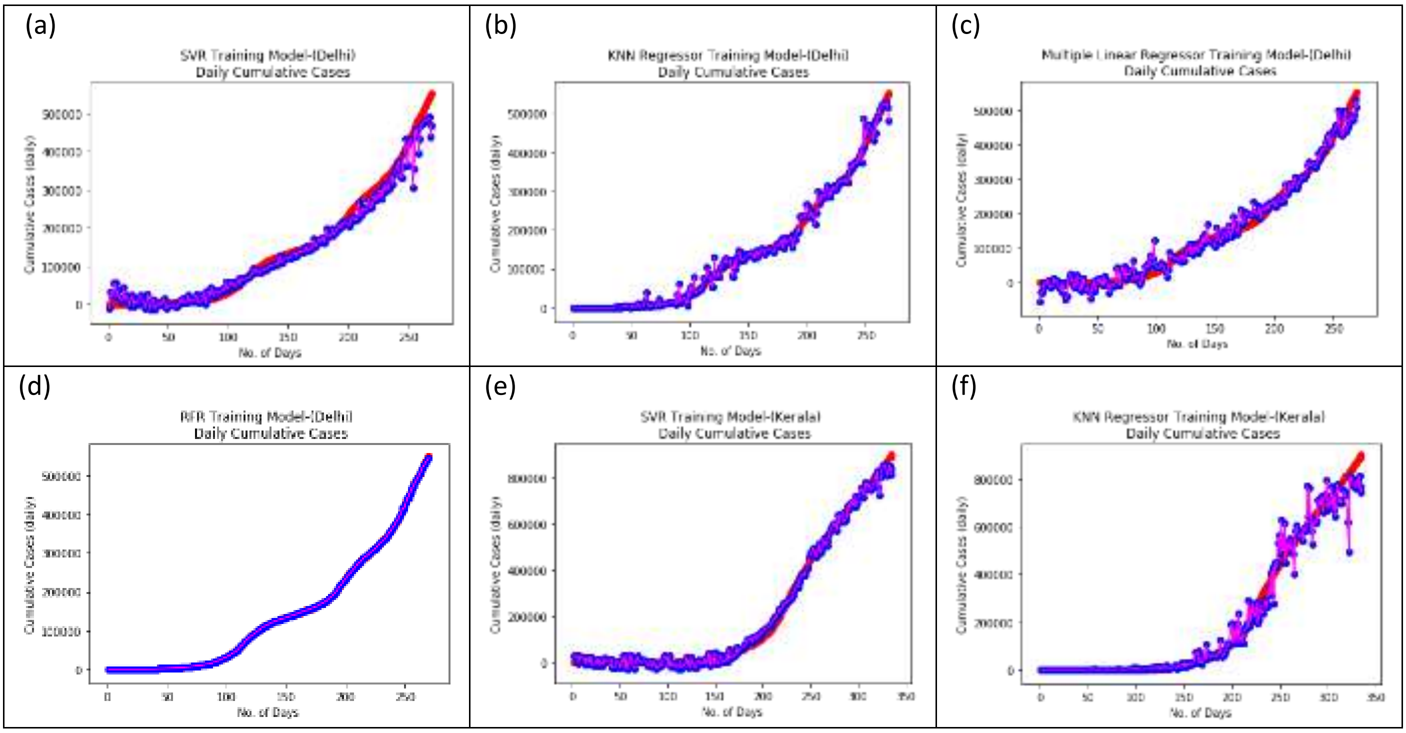
B. Results

The dataset has been created for all four states per case wise as a comma separated files. The programming language used in python. The dataset is pre-processed and divided into Training and Test Sets. The data division was kept at random and shuffled. After that data is normalized and trained. The confidence score of the regressor is measured in each case. Test data is fed into the model and regression metrics – MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), R2 Score is calculated for each case.

This state-of-the-art study of unique and extensive COVID-19 dataset along with Climate parameters reveals few important aspects of COVID-19 spread and future predictions. It is quite evident that RFR and kNN Regressor has out-performed SVR and MLR in most of the cases. The reason for kNN performing better can be because of COVID-19 spread and infection rate which depends on number of active cases in a region. More the number of active cases more will be number of people infected next day. It follows some sort of chronology where the number of COVID-19 cases depends on number of cases found in few days before and after. kNN as an algorithm also polls the values of nearest neighbors in this case nearby dates to predict the future value and hence it has performed better as compared to other algorithms. RFR on the other hand is considered as accurate classifier in most of dataset because of its ensemble characteristics, unbiased estimate of the generalization error and non-linear nature (evident in case of COVID-19 spread).

CASE-1 - Measuring Accuracy of models by predicting "Total Number of COVID -19 Positive cases" per region wise

In this case, the total number of COVID-19 Positive case is predicted on date basis along with temperature, humidity and AQI taken into considerations. Below figure 3 depicts the model performance and metrics for all four states. It is quite evident that RFR has outperformed the other algorithm and has given highest accuracy for all 4 states. For measuring accuracy, the actual data for total number of COVID-19 active cases in that state is compared with predicted values for 10 consecutive days. This average match percentage is captured in table 1 below. RFR has given the best result and its match percentage value hovered between 96% to 99.54%, followed by kNN (85.24% to 99.07%) and SVR (84.76% - 97.2%). MLR has always predicted a greater number of cases as that of actual and hence its match percentage value is more than 100%.



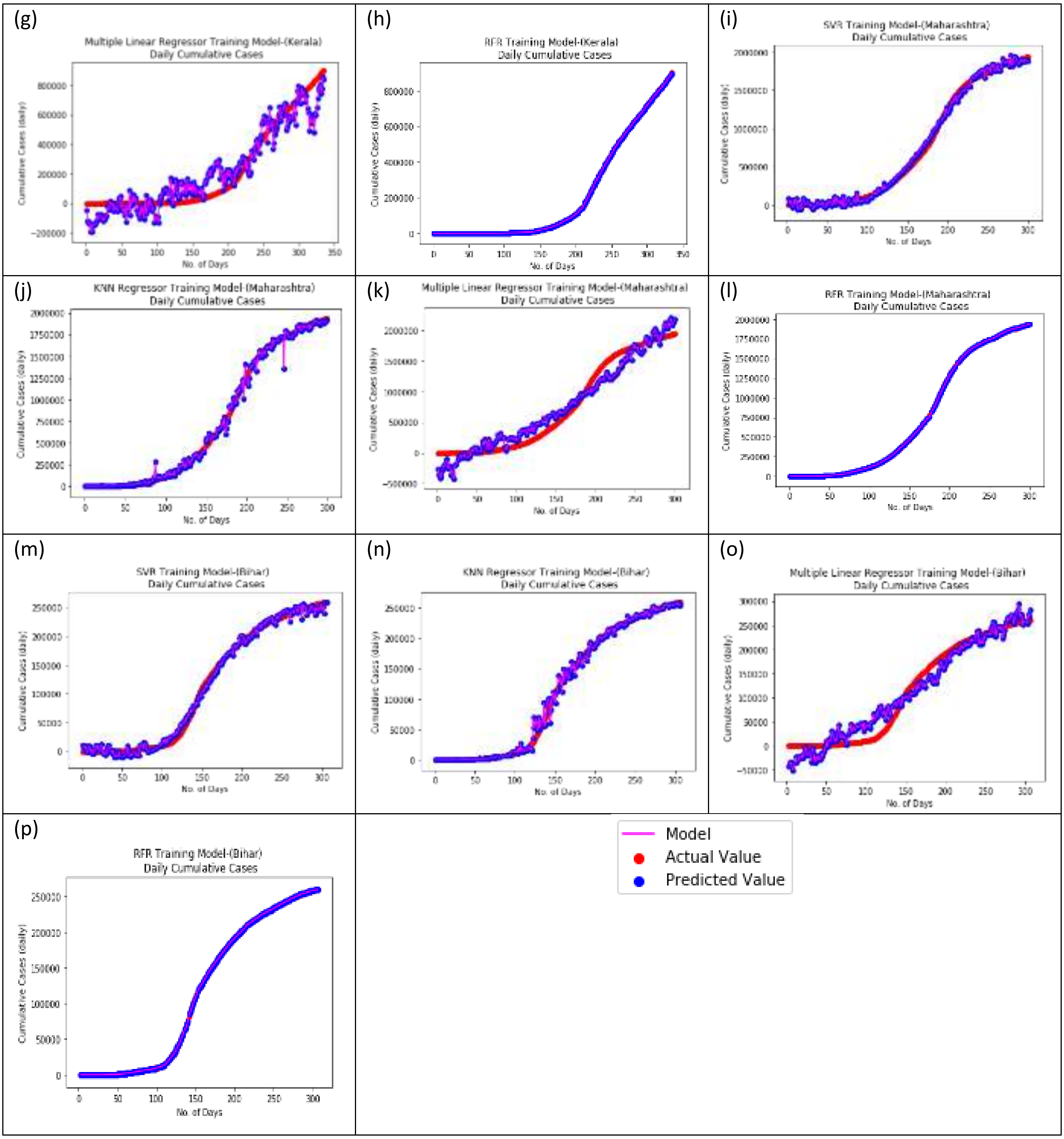


Figure 3: Regression Models (Predicted vs Actual) – CASE-1 for Delhi (a-d), Kerala (e-h), Maharashtra (i-l) and Bihar (m-p)

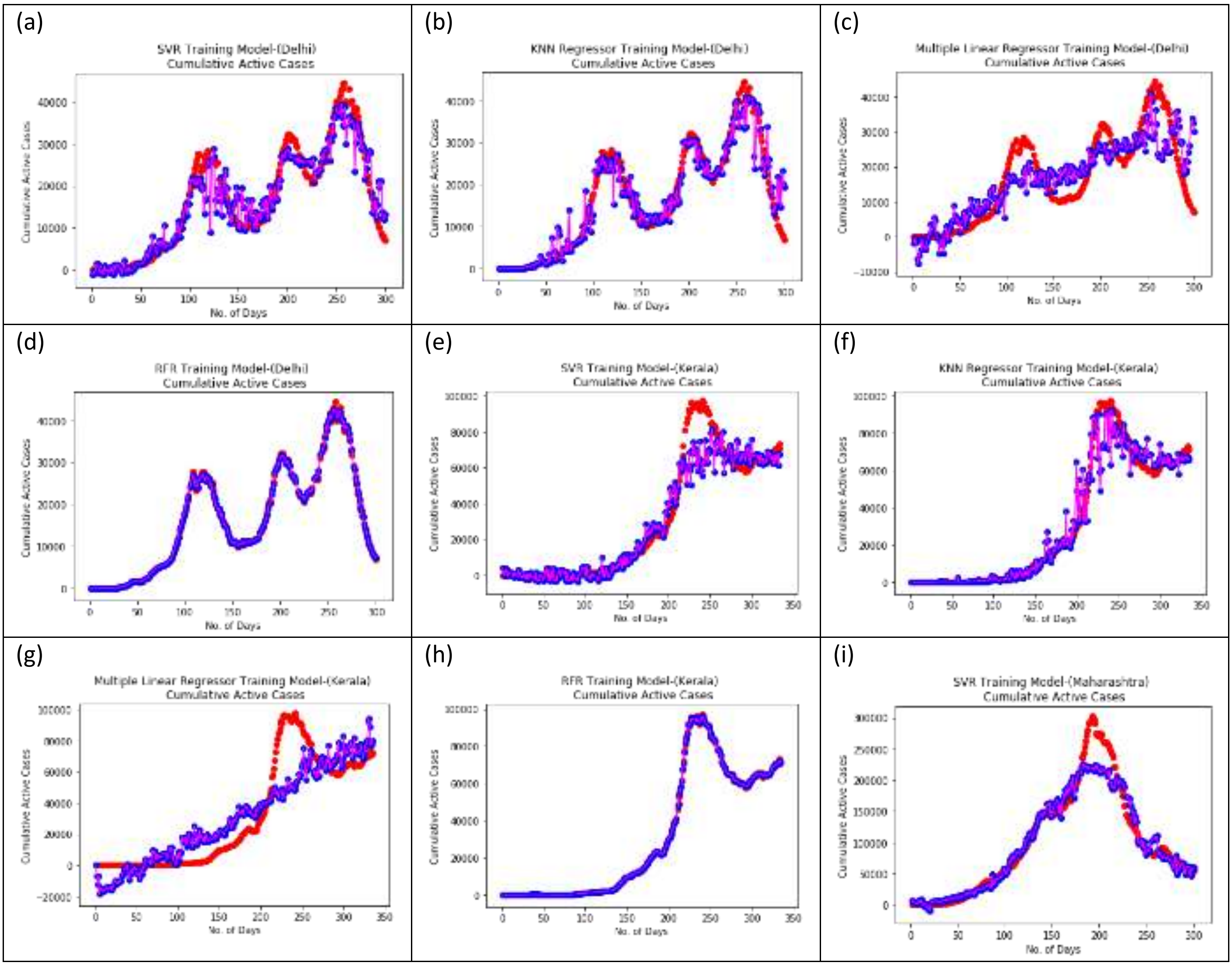
Table- 1 - Regression Metrics – Case - 1

State	Metrics	SVR	KNN-	MLR	RFR
Delhi	MAE	8024.14	8119.111	5371.816	9280.679
	R2-Score	0.6945	0.7238	0.88945	0.6734
	RMSE	10565.06	10046.703	6356.162	10924.88
	Match Percentage	89.86%	96.53%	106.54%	98.96%
Kerala	MAE	28769.31	44787.4	99006.97	37912.71
	R2-Score	0.9765	0.9461	0.8666	0.97768

	RMSE	49602.54	62286.96	118732.4	43190.49
	Match Percentage	84.76%	85.24%	95.91%	95.99%
Maharashtra	MAE	88290.82	83403.909	114757.9	86340.99
	R2-Score	0.977	0.9678	0.96484	0.97922
	RMSE	105232.2	117782.39	139820	101451.1
	Match Percentage	91.6%	97.88%	109.67%	98.72%
Bihar	MAE	10507.5	10660.377	21035.78	7549.614
	R2-Score	0.98759	0.9869	0.93324	0.99367
	RMSE	12170.56	11536.032	25813.44	8157.506
	Match Percentage	97.2%	99.07%	106.24%	99.54%

CASE-2- Prediction of "Total COVID-19 Active cases on daily basis" per region wise

In this case, the total number of COVID-19 Active case on daily basis is predicted with Climate parameters in consideration. The below figure depicts the model performance and also presents the regression metrics. It is quite evident from the graph that there had been different peaks of COVID-19 cases in different states. In Delhi, there indeed are few peaks during these days and SVR has given the best results. Similarly, for Kerala, there was one peak and now it is again going towards another peak due to more cases getting surfaced. KNN regressor has performed best for the case of Kerala, Maharashtra and Bihar.



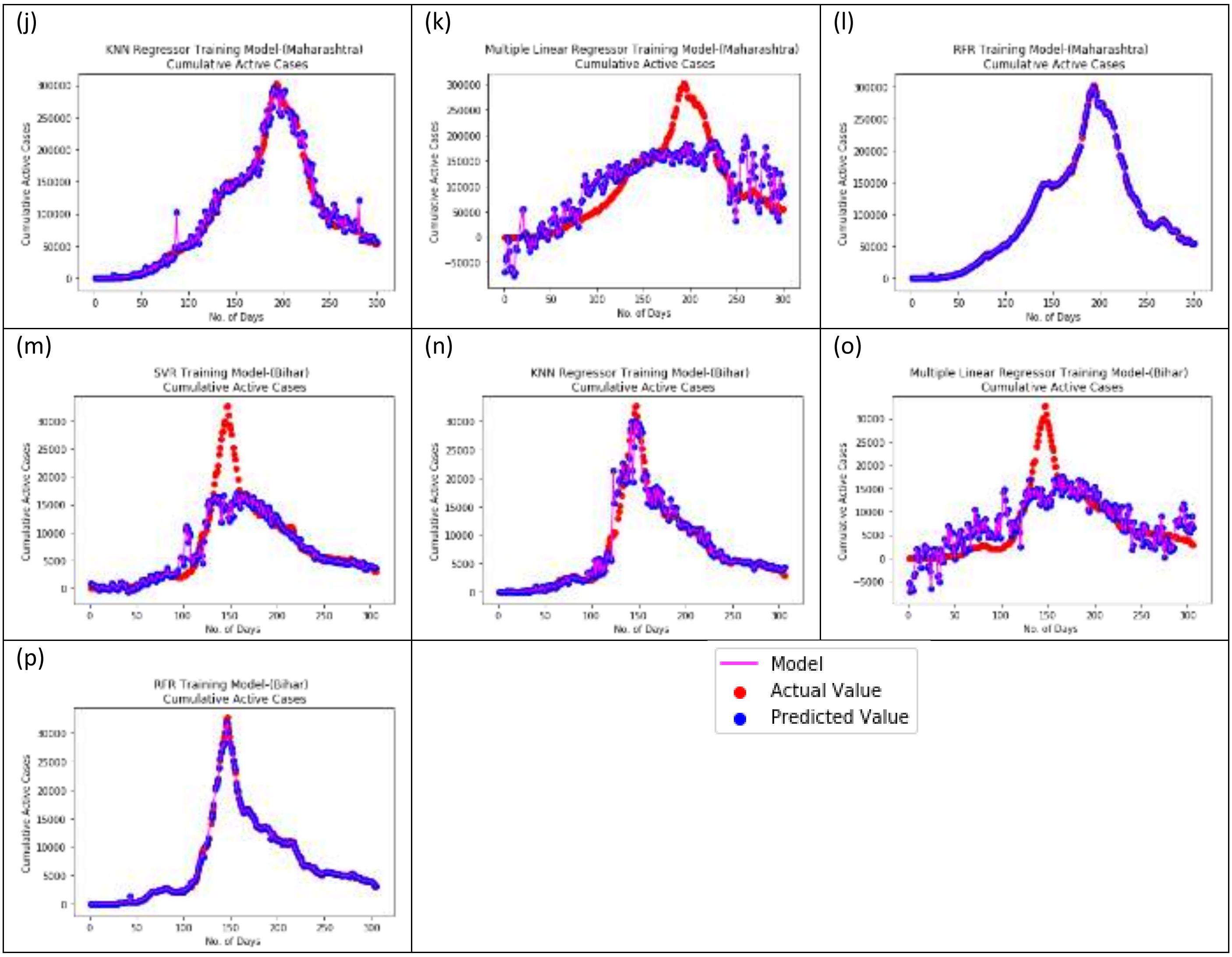


Figure 4: Regression Models (Predicted vs Actual) – CASE-2 for Delhi (a-d), Kerala (e-h), Maharashtra (i-l) and Bihar (m-p)

Table- 2 - Regression Metrics – Case - 2

State	Metrics	SVR	KNN-Regressor	MLR	RFR
Delhi	MAE	4732.131	12666.822	6553.688	5319.684
	R2-Score	0.7846	0.6958	0.4754	0.6264
	RMSE	6438.935	6975.369	8599.456	7039.891
Kerala	MAE	6561.03	5690.203	15380.21	7208.58
	R2-Score	0.9036	0.9135	0.696	0.9212
	RMSE	10436.15	9603.88	19749.18	10155.41
Maharashtra	MAE	28050.06	19738.59	30804.01	11883.55
	R2-Score	0.8177	0.8914	0.6573	0.9733
	RMSE	36045.54	28062.018	41530.72	14593.16
Bihar	MAE	2278.495	1967.037	3638.861	2493.761
	R2-Score	0.7324	0.7013	0.4849	0.6148
	RMSE	4114.071	3540.836	5075.582	3824.69

Further, the prediction for future months has been taken for all regions (Refer to figure 5). If prediction from best model is considered for each region, e.g., SVR for Delhi, it seems that after the initial decline in number of active cases in Jan and Feb, it will start increasing from March and will stabilize in May. For Kerala (as per kNN model), the prediction remains the constant over these months, so there will be consistent number of cases coming all these months. Similarly, for Maharashtra (kNN being the best model), there is an increasing trend of COVID-19 cases from January end onwards, although the increase is linear. In case of Bihar, kNN model predicts, sudden increase from the month of March.

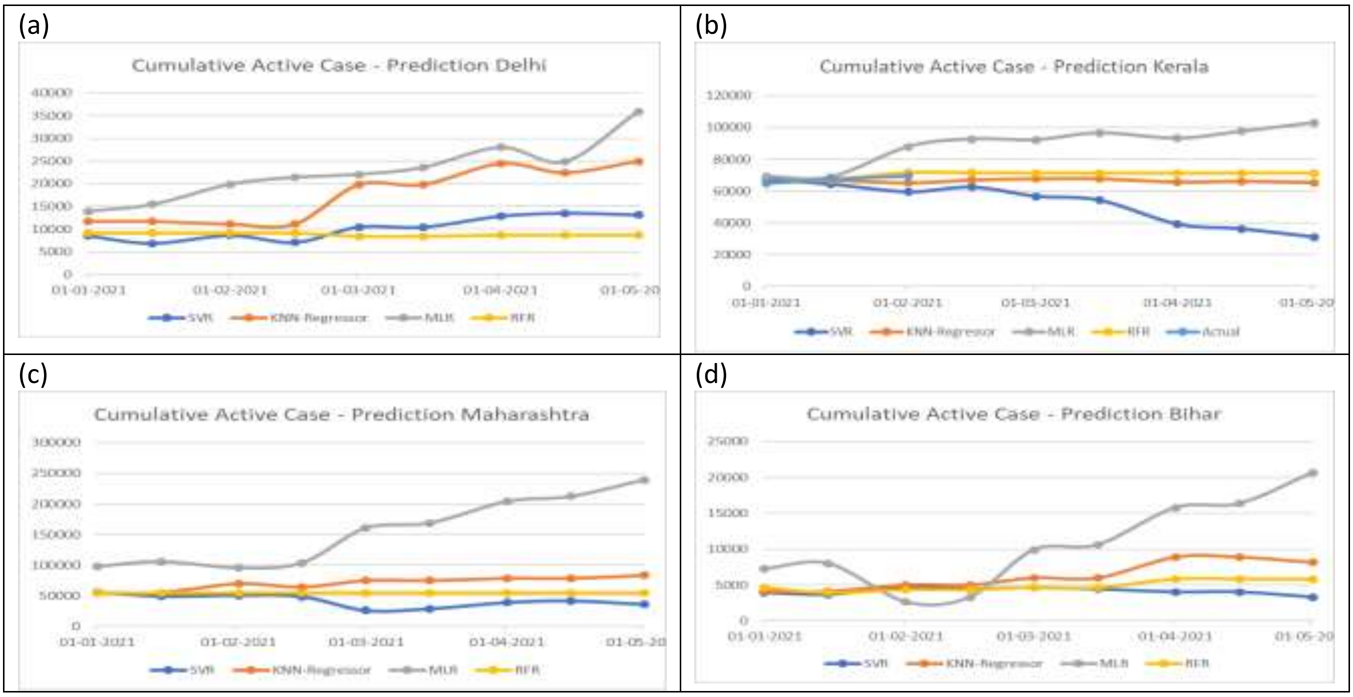
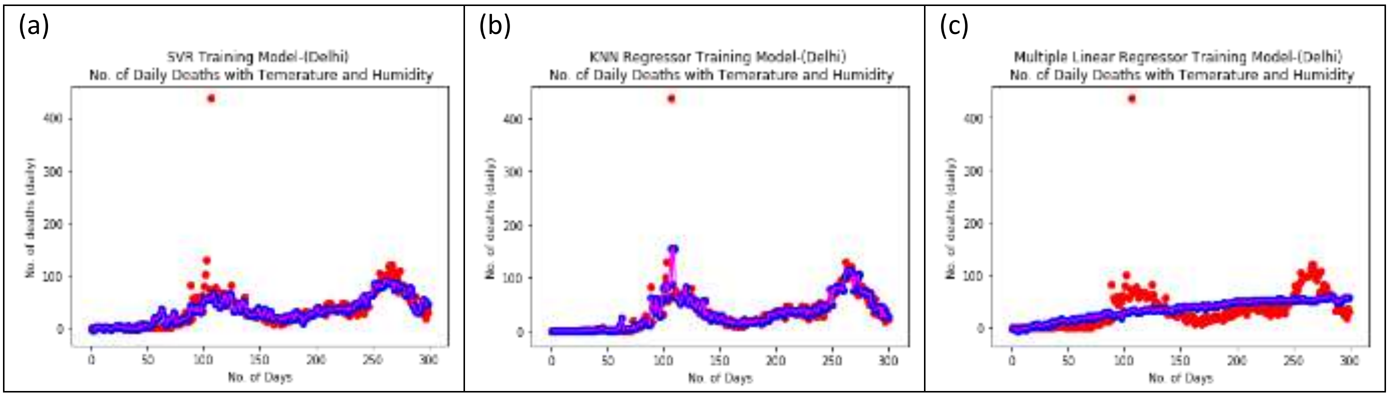


Figure 5: Prediction for Cumulative Active COVID -19 cases - Delhi (a), Kerala (b), Maharashtra (c) and Bihar (d)

CASE-3- Prediction of "Total Daily Deaths due to COVID-19" per region wise dependent on temperature and humidity
In this case, the total daily death count due to COVID-19 was predicted with temperature and humidity in consideration. As per the metrics, kNN Regressor has given best results for Delhi and Bihar, RFR for Kerala and SVR for Maharashtra (Refer to Figure -6). Further based on these models, the total death count on daily basis is predicted for coming months (Figure-7). If we consider prediction for Delhi, all four models are predicting raise in death count from March month. For Kerala, best model RFR, is predicting the constant level of death count in coming months. For Maharashtra, SVR is the best model and it is predicting increasing death count from February onwards. For Bihar, kNN being the best model is predicting sharp increase in death count by March mid to April mid, after that it is declining. The impact of rising temperature on death count is also measured to figure out which temperature range is susceptible of more death count per region (Refer figure 8). kNN being the best model for Delhi is predicting increased death count in the temperature range of 25°C to 30°C. For Kerala, this temperature range is from 25°C to 40°C. For Maharashtra, the temperature range which should be kept in watch is 20°C to 30°C. For Bihar, the best model kNN predicts 30°C to 45°C as the temperature range where death count is more.



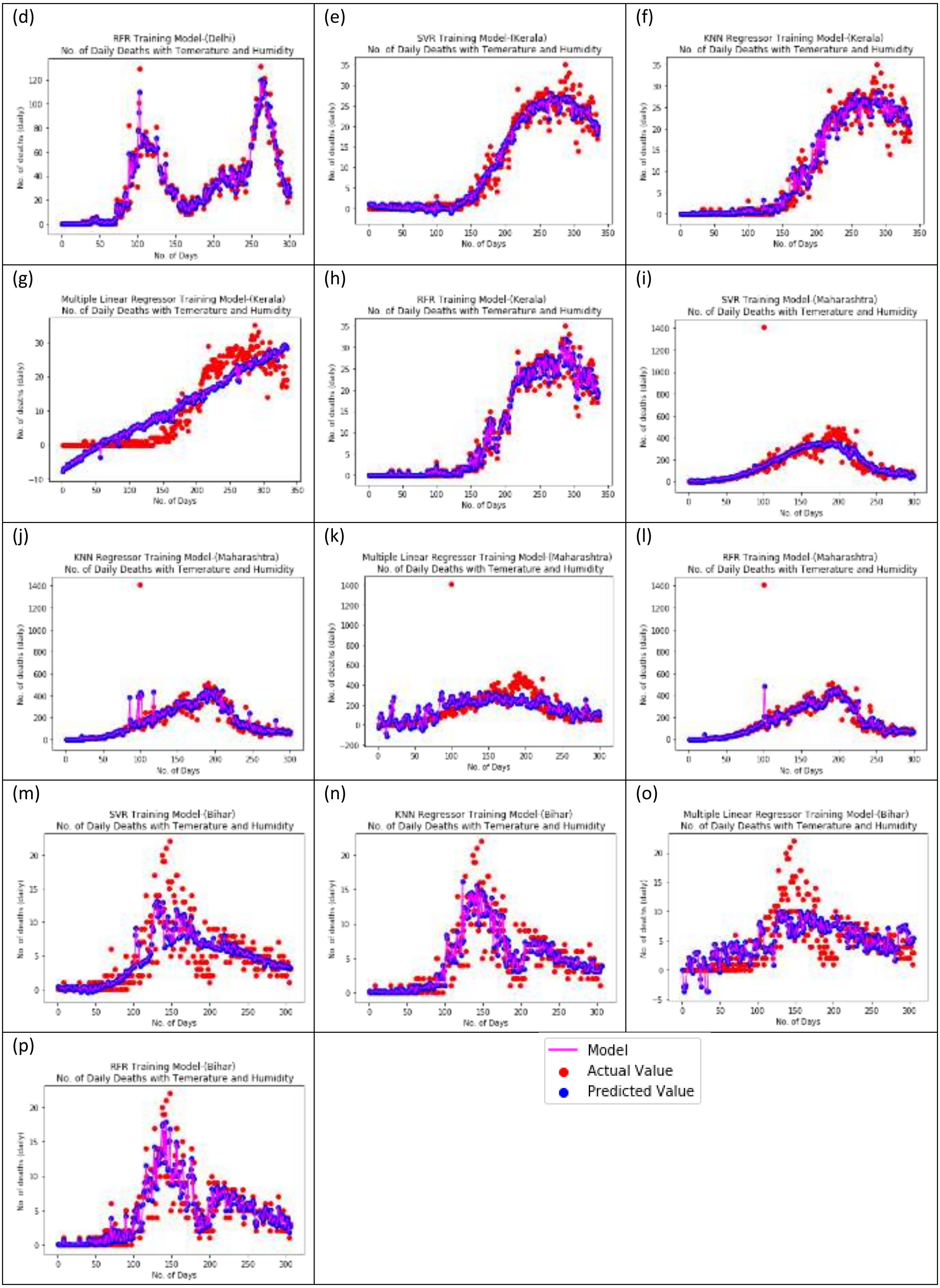

(d)
RFR Training Model-(Delhi)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(e)
SVR Training Model-(Kerala)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(f)
KNN Regressor Training Model-(Kerala)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(g)
Multiple Linear Regressor Training Model-(Kerala)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(h)
RFR Training Model-(Kerala)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(i)
SVR Training Model-(Maharashtra)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(j)
KNN Regressor Training Model-(Maharashtra)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(k)
Multiple Linear Regressor Training Model-(Maharashtra)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(l)
RFR Training Model-(Maharashtra)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(m)
SVR Training Model-(Bihar)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(n)
KNN Regressor Training Model-(Bihar)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(o)
Multiple Linear Regressor Training Model-(Bihar)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
(p)
RFR Training Model-(Bihar)
No. of Daily Deaths with Temerature and Humidity
No. of deaths (daily)
No. of Days
Model
Actual Value
Predicted Value

Figure 6: Regression Models for no. of daily death count (Predicted vs Actual) - Delhi (a-d), Kerala (e-h), Maharashtra (i-l) and Bihar (m-p)

Table- 3 - Regression Metrics – Case - 3

State	Metrics	SVR	KNN-Regressor	MLR	RFR
Delhi	MAE	16.0768	12.0926	24.6389	36.7494
	R2-Score	0.4806	0.6065	0.196	0.15043
	RMSE	21.413	15.9442	32.1526	70.242
Kerala	MAE	2.101	2.632	4.365	1.1897
	R2-Score	0.938	0.9155	0.7896	0.9772
	RMSE	2.875	3.4209	5.037	1.696
Maharashtra	MAE	39.863	42.78	64.296	83.21
	R2-Score	0.7588	0.7804	0.5415	0.25294
	RMSE	59.873	68.78	87.749	111.6
Bihar	MAE	1.9227	1.669	2.161	2.5963
	R2-Score	0.6776	0.4769	0.35094	0.5191
	RMSE	2.5418	2.096	3.0582	3.256

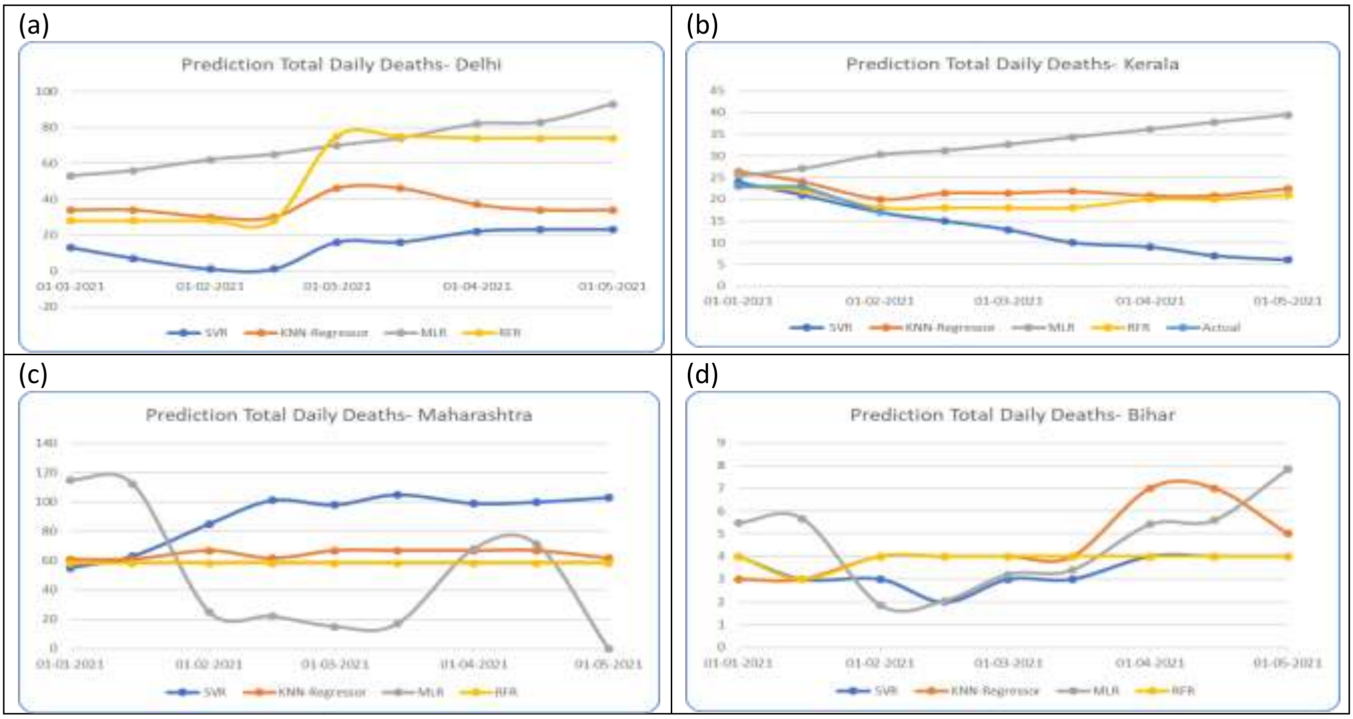


Figure 7 Caption: Prediction of total death count for four states as per the regression models – Delhi (a), Kerala (b), Maharashtra (c) and Bihar (d)

It seems that for India, the moderate to high temperature range (between 25°C to 40°C) is “temperature to watch” where the incidence of COVID-19 cases are predicted to increase.

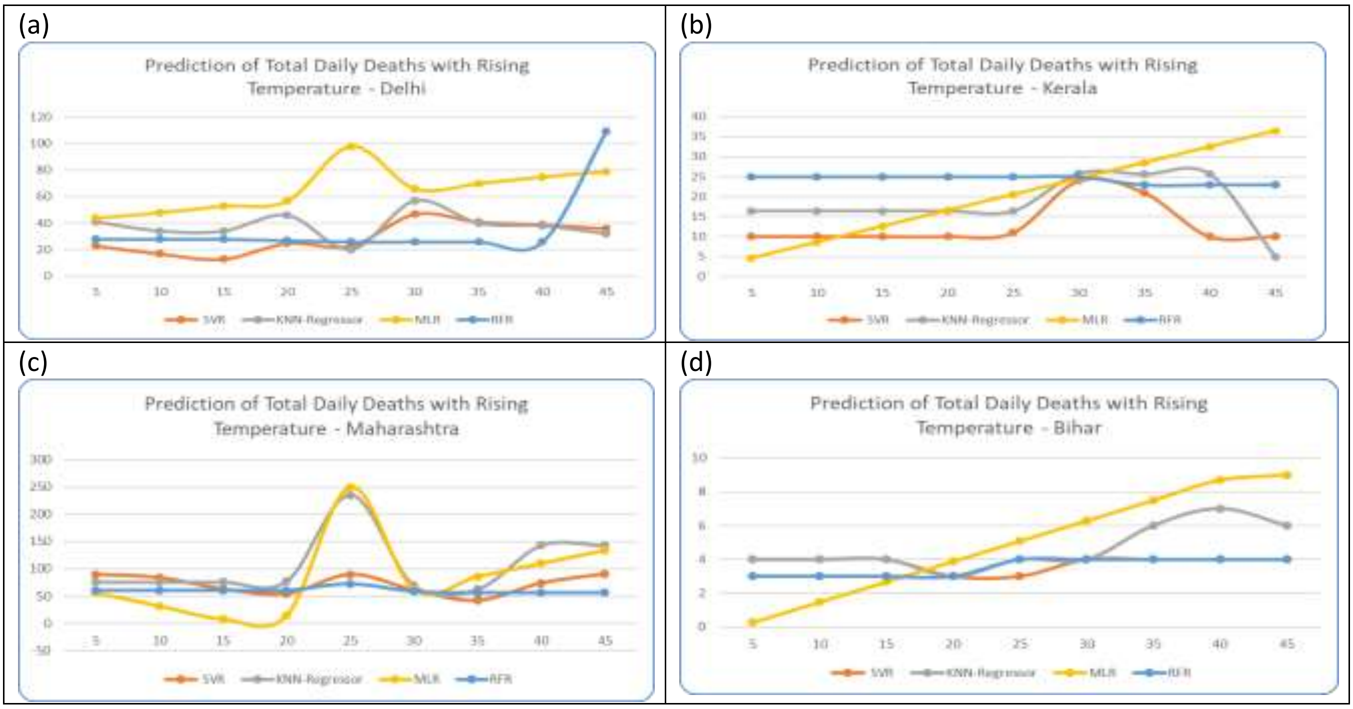


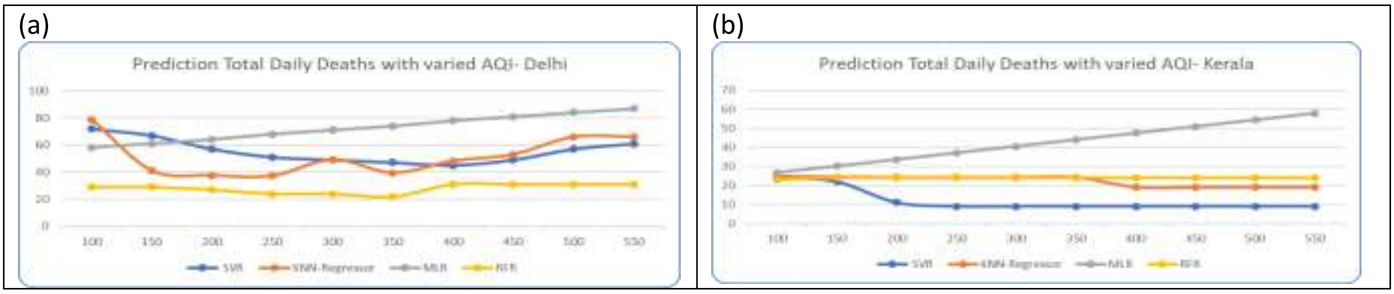
Figure 8: Impact of Daily death count with rising temperature in four states– Delhi (a), Kerala (b), Maharashtra (c) and Bihar (d)

CASE-4 - Prediction of "Total Daily Deaths due to COVID-19" per region wise dependent on Air Quality Index

In this case the death count is tried to be predicted in conjunction of AQI. The regression metrics are given in table below. Further the impact of AQI and total death count is tried to be figured out by predicting against increasing AQI (Figure-9). It is quite evident that except Delhi which has shown increased death count with increasing levels of AQI (250+), rest all states have shown little variance suggesting very less impact of AQI on COVID-19 based death count.

Table- 4 - Regression Metrics – Case - 4

State	Metrics	SVR	KNN-	MLR	RFR
Delhi	MAE	17.835	12.132	15.12407	9.106
	R2-Score	0.5218	0.67	0.4292	0.79717
	RMSE	25.436	19.0273	21.858	12.2553
Kerala	MAE	2.3069	3.688	4.0513	1.5658
	R2-Score	0.9275	0.793	0.8315	0.9606
	RMSE	3.172	4.649	4.7773	2.2175
Maharashtra	MAE	49.4546	12.132	88.923	107.25
	R2-Score	0.7897	0.67	0.5309	0.1208
	RMSE	66.1611	19.0273	102.007	139.9047
Bihar	MAE	2.1272	2.575	3.275	2.0322
	R2-Score	0.5148	-0.1863	0.224	0.4999
	RMSE	3.2784	4.696	4.176	3.1249



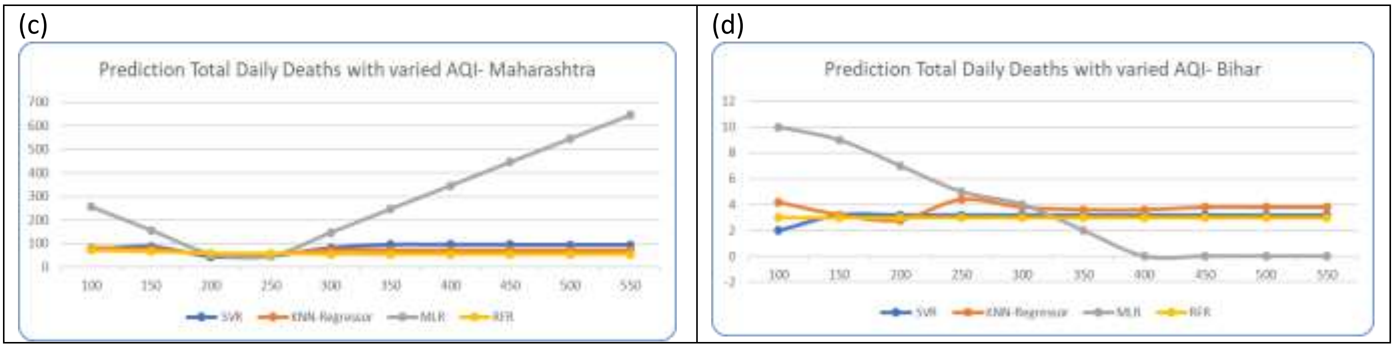
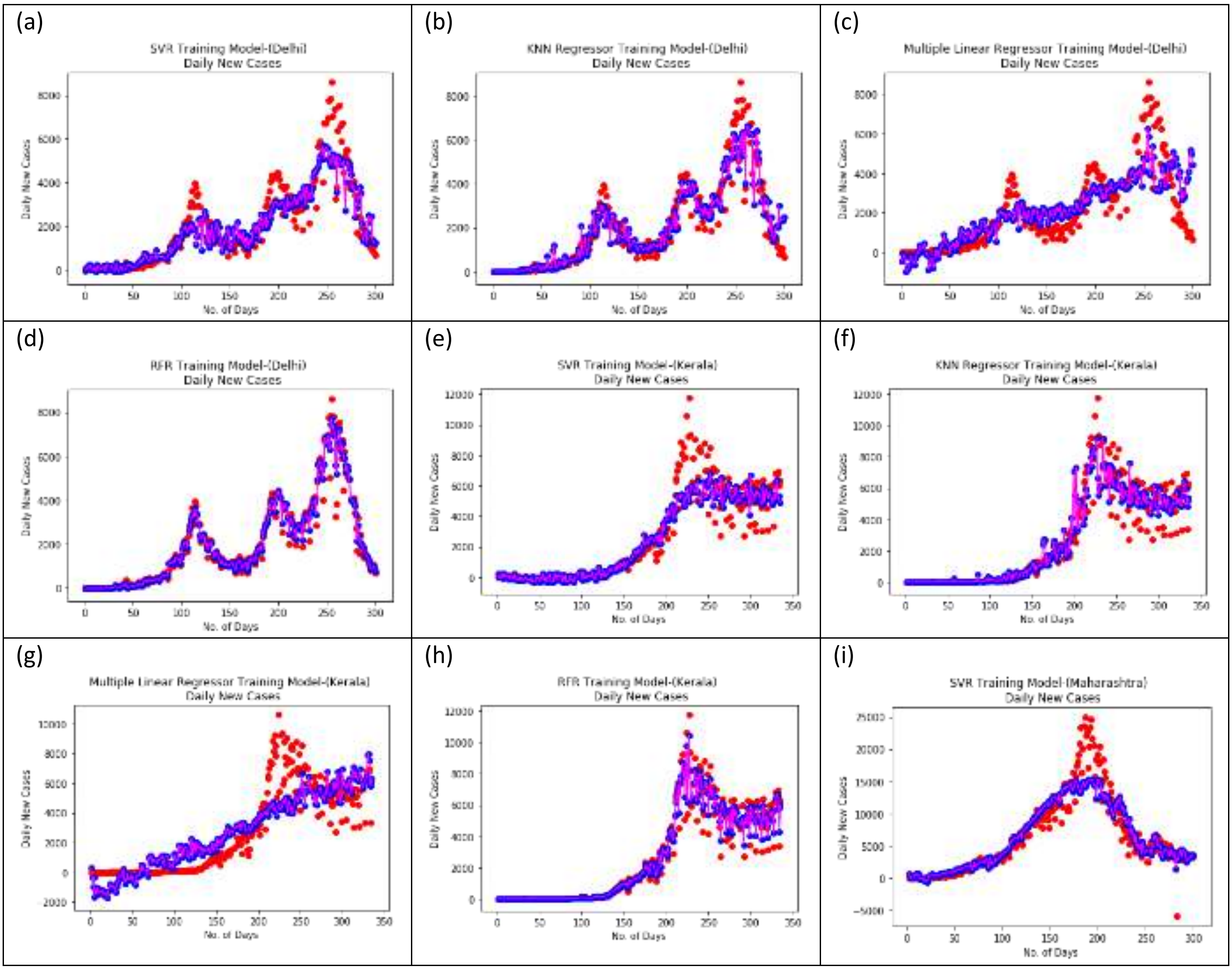


Figure 9: Impact of AQI on daily death count for various states– Delhi (a), Kerala (b), Maharashtra (c) and Bihar (d)

CASE-5 - Prediction of “Total New COVID-19 cases arriving on daily basis” per region

In this case, the daily new COVID-19 cases are modeled along with days count and climate parameters (Figure 10). As per the results, SVR and RFR have performed better in the case of Delhi. kNN has performed best in the case of Maharashtra, Kerala and Bihar. Based on these models, the total new COVID-19 cases arriving on daily basis is predicted for future months (Figure 11). For Delhi (best model was SVR), prediction says that after consistent decline in cases till February, there are chances of having increased number of cases March onwards.



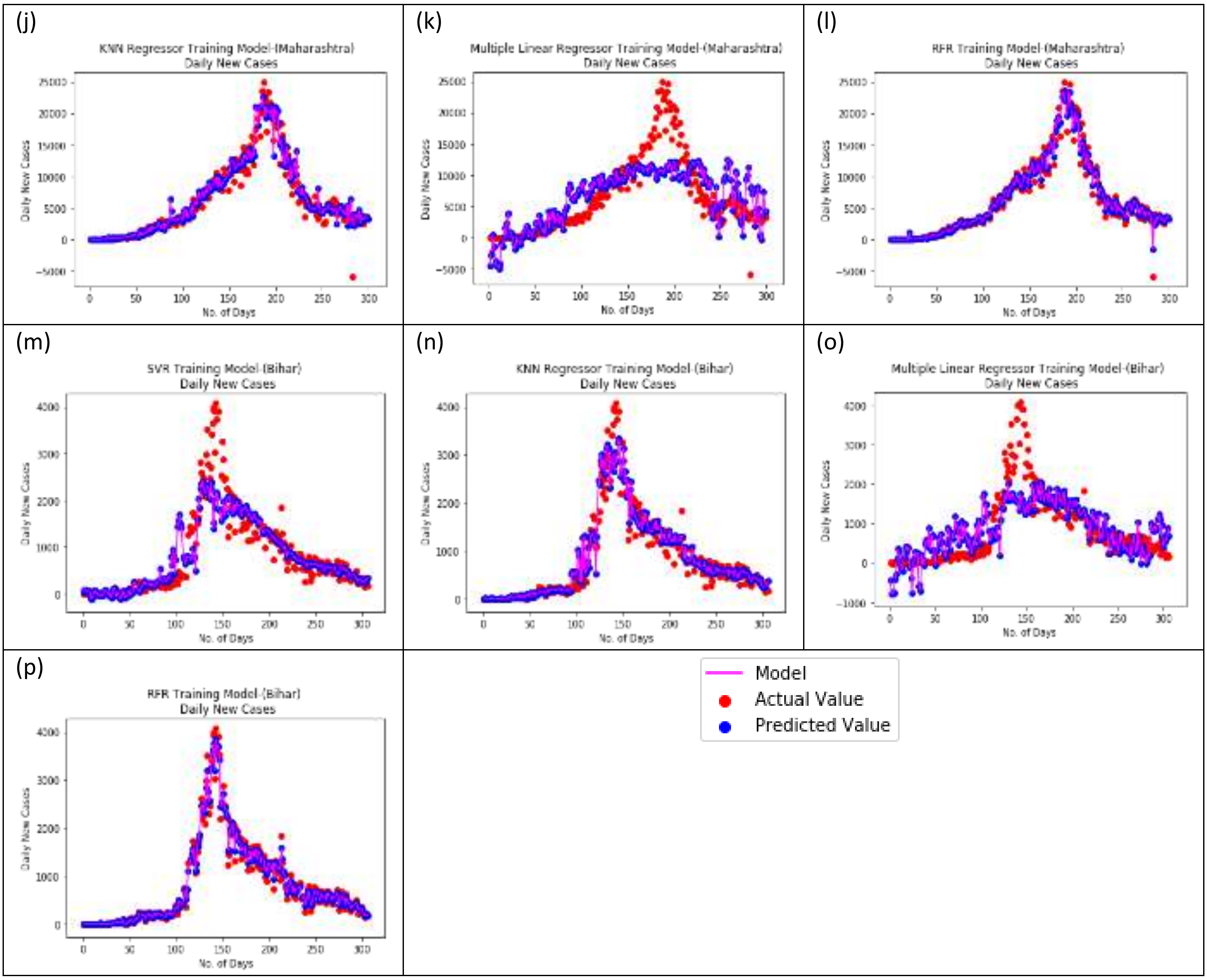


Figure 10: Regression Model for Daily new COVID-19 cases - Delhi (a-d), Kerala (e-h), Maharashtra (i-l) and Bihar (m-p)

Table- 5 - Regression Metrics – Case - 5

State	Metrics	SVR	KNN-Regressor	MLR	RFR
Delhi	MAE	652.162	749.02	1065.576	582.2567
	R2-Score	0.8781	0.5833	0.5179	0.7926
	RMSE	814.144	1321.549	1448.017	768.601
Kerala	MAE	905.275	666.503	1192.625	702.325
	R2-Score	0.8238	0.8972	0.5466	0.87266
	RMSE	1319.03	904.64	1889.704	1123.776
Maharashtra	MAE	2105.219	1095.676	2660.7	1557.583
	R2-Score	0.6422	0.95033	0.5024	0.8853
	RMSE	3134.915	1488.222	3642.392	2071.04
Bihar	MAE	251.525	243.419	509.263	699.729
	R2-Score	0.6627	0.39193	0.4142	-0.25748
	RMSE	452.704	551.6647	765.628	1014.933

For the case of Kerala (Both kNN and RFR have shown better accuracy), models predict that cases will keep on getting reported at the constant pace as it is getting reported at the start of 2021. In the case of Maharashtra (kNN being the best model) shows an increment in cases starting February 2021. For the case of Bihar (kNN being the best model) predicts increase in cases from Mid-March.

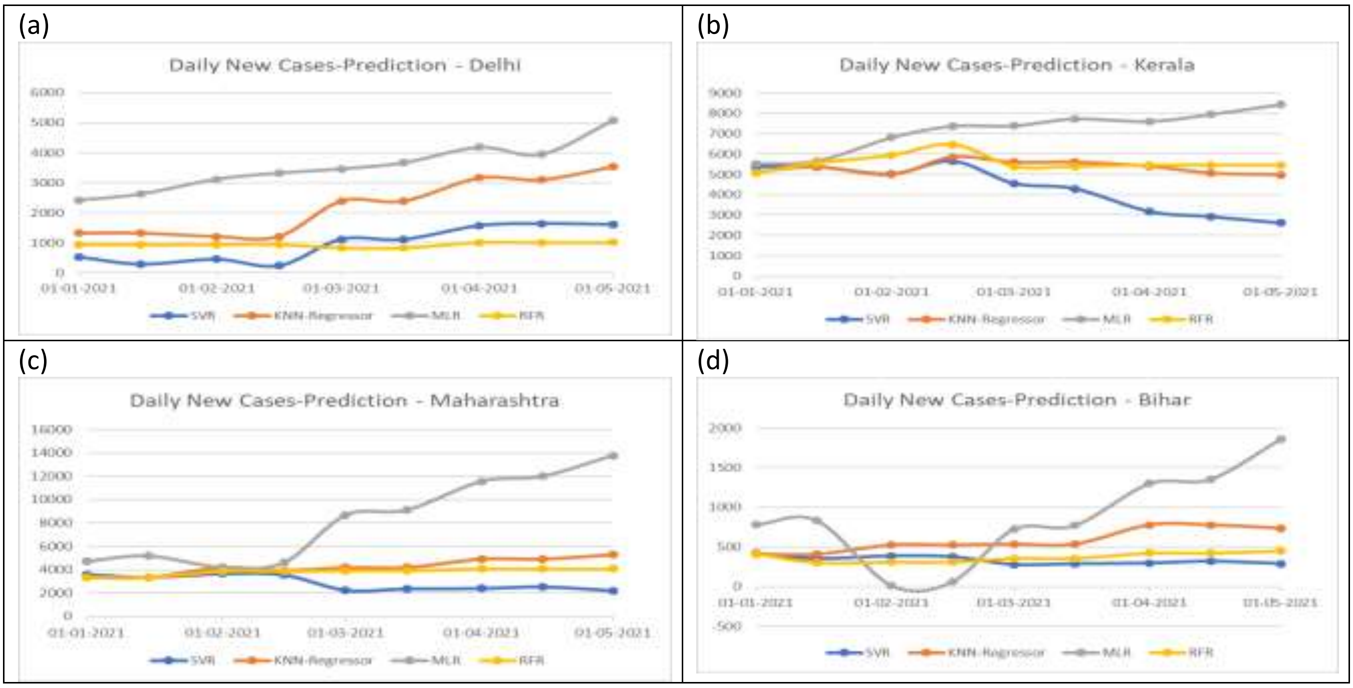


Figure 11: Prediction of Daily new cases in four states– Delhi (a), Kerala (b), Maharashtra (c) and Bihar (d)

III. CONCLUSION

Study is able to predict a long-term prediction (180 days) for COVID-19 cases spread and fatalities. It is also able to correlate the impact of weather conditions and Air quality Index on COVID-19 spread. The next 180 days prediction revealed that a surge of COVID-19 cases will be witnessed in the start of February 2021 for Maharashtra, End of February 2021 for Delhi and End of March 2021 for Bihar. For the state of Kerala, the COVID-19 cases will remain at the constant level which was there in January 2021. Air Quality Index seems to be of least effect on COVID-19 fatalities and COVID-19 cases count is showing increased trend for moderate temperature range of 25-40 °C for India. As with all studies, this study also has some assumptions and limitations. Study has only taken Maximum temperature and Average Humidity as climate parameters, but there can be other parameters like Wind speed, Precipitations, Rain fall etc. which can have impact on virus spread. On the algorithm side it will be worth important and meaningful to check with Deep Learning models like Recurrent Neural Network and LSTM. Also, some constraint-based algorithm with ML should be tried for better accuracy. This study and its accurate prediction can help Government agencies to intervene in timely manner and take adequate known methods like social distancing, Immunization to control the spread of COVID-19 in particular zone.

Compliance with ethical standards

Conflict of interest: The authors declare that they have no conflict of interest.

REFERENCES

[1] Byass, P. Eco-epidemiological assessment of the COVID-19 epidemic in China, January-February 2020. Glob. Health Action 2020, 13, 1760490, doi:10.1101/2020.03.29.20046565. [CrossRef] [PubMed]

[2] https://www.who.int/news/item/29-06-2020-covidtimeline

[3] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_India

[4] Ivanov, D. Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transp. Res. Part E Logist. Transp. Rev. 2020, 136, doi:10.1016/j.tre.2020.101922.

[5] Koolhof, I.S.; Gibney, K.B.; Bettiol, S.; Charleston, M.; Wiethoelter, A.; Arnold, A.L.; Campbell, P.T.; Neville, P.J.; Aung, P.; Shiga, T., et al. The forecasting of dynamical Ross River virus outbreaks: Victoria, Australia. Epidemics 2020, 30, doi:10.1016/j.epidem.2019.100377.

[6] Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188.

[7] Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (2020). COVID-19 Pandemic Prediction for Hungary; a Hybrid Machine Learning Approach. Mathematics, 8(6), 890.

[8] Chaurasia, V., & Pal, S. (2020). Covid-19 Pandemic: Application of Machine Learning Time Series Analysis for Prediction of Human Future. Available at SSRN 3652378.

[9] Sujath, R., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. Stochastic Environmental Research and Risk Assessment, 1.

[10] Dhanwant, J. N., & Ramanathan, V. (2020). Forecasting COVID 19 growth in India using Susceptible-Infected-Recovered (SIR) model. arXiv preprint arXiv:2004.00696.

[11] Malavika, B., Marimuthu, S., Joy, M., Nadaraj, A., Asirvatham, E. S., & Jeyaseelan, L. (2020). Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models. Clinical Epidemiology and Global Health.

[12] Pereira, I. G., Guerin, J. M., Junior, A. G. S., Distante, C., Garcia, G. S., & Goncalves, L. M. (2020). Forecasting Covid-19 dynamics in Brazil: a data driven approach. arXiv preprint arXiv:2005.09475.

[13] Wang, P., Zheng, X., Ai, G., Liu, D., & Zhu, B. (2020). Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. Chaos, Solitons & Fractals, 140, 110214.

[14] Yadav, M., Perumal, M., & Srinivas, M. (2020). Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos, Solitons & Fractals, 139, 110050

[15] Chan, K. H., Peiris, J. M., Lam, S. Y., Poon, L. L. M., Yuen, K. Y., & Seto, W. H. (2011). The effects of temperature and relative humidity on the viability of the SARS coronavirus. Advances in virology, 2011.

[16] Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., ... & Liu, M. (2020). Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Science of the Total Environment, 729, 139051.

[17] Behnood, A., Golafshani, E. M., & Hosseini, S. M. (2020). Determinants of the infection rate of the COVID-19 in the US using ANFIS and virus optimization algorithm (VOA). Chaos, Solitons & Fractals, 139, 110051.

[18] Naqvi, H. R., Datta, M., Mutreja, G., Siddiqui, M. A., Naqvi, D. F., & Naqvi, A. R. (2021). Improved air quality and associated mortalities in India under COVID-19 lockdown. Environmental Pollution, 268, 115691.

[19] Xu, H., Yan, C., Fu, Q., Xiao, K., Yu, Y., Han, D., ... & Cheng, J. (2020). Possible environmental effects on the spread of COVID-19 in China. Science of the Total Environment, 731, 139211.

[20] Heidari, A., Jafari Navimipour, N., Unal, M., & Toumaj, S. (2022). Machine learning applications for COVID-19 outbreak management. Neural Computing and Applications, 34(18), 15313-15348.

[21] Vega, R., Flores, L., & Greiner, R. (2022). SIMLR: Machine Learning inside the SIR model for COVID-19 Forecasting. Forecasting, 4(1), 72-94.

[22] Martínez-Fernández, P., Fernández-Muñiz, Z., Cernea, A., Fernández-Martínez, J. L., & Kloczkowski, A. (2023). Three Mathematical Models for COVID-19 Prediction. Mathematics, 11(3), 506.

[23] COVID19INDIA - https://www.covid19india.org/

[24] Weather Underground - https://www.wunderground.com/.

[25] Air Quality Data - https://aqicn.org

[26] Population Density India - https://censusindia.gov.in/2011-prov-results/data_files/india/Final_PPT_2011chapter7.pdf

[27] Chan, K. H., Peiris, J. M., Lam, S. Y., Poon, L. L. M., Yuen, K. Y., & Seto, W. H. (2011). The effects of temperature and relative humidity on the viability of the SARS coronavirus. Advances in virology, 2011.

[28] Prata, David N., Waldecy Rodrigues, and Paulo H. Bermejo. "Temperature significantly changes COVID-19 transmission in (sub) tropical cities of Brazil." Science of the Total Environment 729 (2020): 138862.

[29] Li, H., Xu, X. L., Dai, D. W., Huang, Z. Y., Ma, Z., & Guan, Y. J. (2020). Air pollution and temperature are associated with increased COVID-19 incidence: a time series study. International journal of infectious diseases, 97, 278-282.

[30] Müller, K. R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., & Vapnik, V. (1997, October). Predicting time series with support vector machines. In International Conference on Artificial Neural Networks (pp. 999-1004). Springer, Berlin, Heidelberg.

[31] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[32] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[33 Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[34] Tranmer, M., & Elliot, M. (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), 1-5.