Applied Intelligence (2021) 51:2714–2726
https://doi.org/10.1007/s10489-020-01948-1

Kalman filter based short term prediction model for COVID-19 spread

Koushlendra Kumar Singh[1] · Suraj Kumar[1] · Prachi Dixit[2] · Manish Kumar Bajpai[3]

Accepted: 12 September 2020 / Published online: 3 November 2020

Abstract

Corona Virus Disease 2019 (COVID19) has emerged as a global medical emergency in the contemporary time. The spread scenario of this pandemic has shown many variations. Keeping all this in mind, this article is written after various studies and analysis on the latest data on COVID19 spread, which also includes the demographic and environmental factors. After gathering data from various resources, all data is integrated and passed into different Machine Learning Models in order to check its appropriateness. Ensemble Learning Technique, Random Forest, gives a good evaluation score on the tested data. Through this technique, various important factors are recognized and their contribution to the spread is analyzed. Also, linear relationships between various features are plotted through the heat map of Pearson Correlation matrix. Finally, Kalman Filter is used to estimate future spread of SARS-Cov-2, which shows good results on the tested data. The inferences from the Random Forest feature importance and Pearson Correlation gives many similarities and few dissimilarities, and these techniques successfully identify the different contributing factors. The Kalman Filter gives a satisfying result for short term estimation, but not so good performance for long term forecasting. Overall, the analysis, plots, inferences and forecast are satisfying and can help a lot in fighting the spread of the virus.

Keywords COVID19 · Kalman filter · Pearson correlation · Random Forest

1 Introduction

Corona virus belongs to a group of viruses that cause disease in mammals and birds. The viral agent (that is, the coronavirus) is SARS-Cov-2 (severe acute respiratory syndrome coronavirus 2). The name ‘corona virus’ is derived from Latin word corona meaning “crown”. The name refers to the characteristic appearance of virions by electron microscopy. This is due to the ‘viral’ spikes which are proteins on the surface of the virus. It causes respiratory throat infections in humans. SARS-Cov-2 is a large spherical particle with surface projections. The average diameter is approximately 120 nm (nm) [1]. Envelope diameter is 80 nm while spikes are 20 nm long. The viral envelope consists of a lipid out layer where the membrane, envelope and spike structural proteins are anchored. Inside the envelope, there is the nucleocapsid, which is formed of proteins, single stranded RNA genome. SARS-Cov-2 contains SSRNA [2]. The genome size ranges from 26.4–31.7 ab. It also contains “polybasic cleavage site” which increases pathogenicity and transmissibility. The spike protein is responsible for allowing the virus to attach and fuse with the membrane of a host cell. SARS-Cov-2 has sufficient affinity to the receptor angiotensin converting enzyme on human cells to use them as mechanism of cell entry [3]. The cell’s transmembrane protease serine cuts open the protein exposing peptide, after a SARS-Cov-2 virion attaches to a target cell. The virion then releases RNA into the cell and forces the cell to produce and disseminate copies of the virus [4]. The RNA genome then attaches ribosome to host cells for translation. The basic reproductive number (RO) of the virus has been estimated to be between 1.4 to 3.9. The infection transmits mainly through droplets released from the infected person or surfaces containing these infected droplets [5].

The detection and controlling of any infection based disease outbreaks have been major concern in public health [2]. The data collection from different sources such as health departments, emergency department, weather information etc. plays vital role in decision making to control the epidemic. It has been well established in literature that data sources contains important information that helps the current public health status [6, 7]. Now, it is very crucial that the world health

✉ Koushlendra Kumar Singh
koushlendra.cse@nitjsr.ac.in

[1] National Institute of Technology, Jamshedpur, India

[2] Jai Narayan Vyas University, Jodhpur, India

[3] Indian Institute of Information Technology Design and Manufacturing, Jabalpur, India

authorities and the global population remain vigilant against any resurgence of the virus. Situational Information in social media during COVID19 has been studied [8]. Markov switching model has been used to detect the outbreak of the disease [9]. Prediction model of HIV incidence has been proposed using neural network [10]. There are many methods existing in literature to detect various diseases at an early stage [11–13].

In case of an epidemic, prediction at an initial stage plays a pivotal role in order to control the epidemic. The Government agencies and the public health organizations can prepare accordingly as per the prediction. There are many prediction algorithms available in the literature for different types of data [2, 9]. There are various Fuzzy based model available in the literature to model different non-linear systems. The evolving fuzzy models are developed for prosthetic hand myoelectric based control system [14]. The wavelet neural networks are used to predict the muscle force, and to predict the handgrip force of a human [15]. The TakgiSugeno-Kang (TSK) fuzzy models have been used to model nonlinear dynamic mechanisms of hand finger [16]. The fuzzy based model is developed to predict the optimal green period ratios and the traffic light cycle times of a road [17]. As discussed in the literature, Artificial Intelligence techniques also include the neural network to model the nonlinear system [18]. Deep neural network based approach is developed to model the heart failure of a patient with the use of ECG signals [19]. The Random forest model is also applicable to predict the heart failure of a patient [20].The forecasting of power systems have been done with the help of Random forest model [21]. The random forest algorithm provides an alternative approach to predict the uncertain data [22]. Kalman filter is a popular filter that is used to study multivariable systems, highly fluctuated data, and time varying systems. They are also suitable to forecast random processes [23]. Yang and Zhang used Kalman filter for prediction of stock price [24]. They have used Changbasihan as a test case to predict the stock price [24]. The extended Kalman filter in nonlinear domain has been studied by Iqbal et al. [25]. Kalman filter is very useful in the field of Robotics [26]. It is extensively used to optimize the robotic movements, tracking of robots and their localization [26]. Kalman filter is used in supply chain as abstraction [27]. It is also useful in manufacturing process to improve the capability of overall process [28]. Another useful application of Kalman filter was reported in literature to estimate parameters of train i.e. coating on a flat track [26]. ANIMA based model has been used to forecast COVID19 [29].

The structural and replication mechanism followed by SARS-Cov-2 have motivated the study of demogramic and environmental factors, affecting its spread. The present paper has included many such factors for the study; for e.G. *minimum* Temperature, Maximum Temperature, Humidity, and Rain Fall have been used. Effects of these factors have been studied thoroughly on the spread rate. The effect is also studied on death rate and active cases rate.

Kalman filter has been used in order to formulate the present paper. The future forecasting of COVID19 has been done through the proposed model. The Pearson correlation has been used to find the dependency of different features of the data. The importance of the feature in the proposed model has been calculated through random forest algorithm.

The novelty and contribution of the paper are as follows. Many other engineering applications have already used Kalman filter. This paper uses a well-established concept of Kalman filter for the forecast of the epidemic for the first time. Thorough study has been done on different parameters affecting the growth of SARS-Cov-2. Paper has also compared the performance of Kalman filter based prediction algorithm with other state-of-the-art methods present in literature and have been used for prediction purpose. Different performance metrics have been used to validate the claims made in the paper. The performance of proposed method established the claims and have scientific validation.

The paper is divided into various sections: Section-2 describes the proposed methodology of prediction. Section-3 outline the details of data collection and data processing. Section- 4 deals in the discussion of results and the conclusions are pointed out in section-5.

2 Proposed methodology

The proposed prediction method is divided into three sections. In the first section we developed a model with Kalman filter to forecast the confirmed cases of COVID19. The second section of the proposed approach uses the Pearson correlation to find the correlation among the features of data. The random forest algorithm is used in the last section of the proposed approach to find the dependency among the features of confirmed cases of SARS-Cov-2.

The Kalman filter tries to predict the state $X \in R^n$ of a discrete-time dependent process that is controlled by linear stochastic difference equation:

$$x_i = Ax_{i-1} + Bj_i + l_{i-1} \tag{1}$$

Here A is a matrix that denotes the relation between states of previous and current time step and B is also a matrix which denotes the relationship between input and state function.

With a measurement$Y \in R^m$ which is

$$y_i = Gx_i + m_i \tag{2}$$

The random variable l_i represents the process noise and $m_i$represents the measurement noise. Both of these variables are assumed to be independent of each other with normal probablity distributions

$$p(l)\approx N(0,Q) \tag{3}$$

$$p(m)\approx N(0,R) \tag{4}$$

Here, Q is the process noise covariance matrix and R is the measurement noise covariance matrix. A is nxn matrix which establishes relationship between the state at the previous time step and the state at the current time step, when the driving function or process noise is absent. B is $nx1$ matrix which establishes relationship between the optional control input $j\in R^1$ and the state x. G is mxn matrix which establishes relationship between the state and the measurementy_i.

2.1 The discrete Kalman filter algorithm

We can narrow the focus to the specific equations and their use in this version of the filter. The Kalman filter does estimation of a process by a kind of feedback control: the filter predicts the process state at some time and then accepts feedback in the form of measurements. Hence, the Kalman filter algorithm can be divided into two parts: *(i)* Time Update Equations (*Prediction Phase*) and *(ii)* Measurements Update Equations (*Feedback Phase*).

The time update equations are accountable for advancing the current state and error covariance forward in time. To obtain a priori estimates of the upcoming time step. The measurement update equations are accountable for the feedback, i.e., fetching the actual measurement and changing the parameters to improve the Kalman Filter. In order to improve the posteriori estimates. Hence, in simple words, Time Update Equations can be considered as predictor equations, while the Measurement Update Equations can be considered as corrector equations, and the algorithm works in these two steps.

Time Update Equations:

$$\widetilde{X}_i = A_{i-1}X_{i-1} + B_iJ_i \tag{5}$$

$$\widetilde{C}_i = A_{i-1}C_{i-1}A_{i-1} + Q_{k-1} \tag{6}$$

Measurements Update Equations:

$$M_i = Y_i - G_i\widetilde{X}_i \tag{7}$$

$$S_i = G_i\widetilde{C}_iG_i^T + R_i \tag{8}$$

$$K_i = \widetilde{C}_iG_k^TS_i^{-1} \tag{9}$$

$$X_i = \widetilde{X}_i + K_iM_i \tag{10}$$

$$C_i = \widetilde{C}_i - K_iS_iK_i^T \tag{11}$$

where

- $\widetilde{X}_i$ is the predicted mean on time step i before seeing the measurement and X_i is the estimated mean on time step i after seeing the measurement.
- $\widetilde{C}_i$ is the predicted covariance on time step i before seeing the measurement and $C_i$is the estimated covariance on time step i after seeing the measurement.
- A_i is the state relationship on time step i.
- B_i is control input relationship on time step i.
- Y_i is mean of the measurement on time step i .
- M_i is the measurement residual on time step i .
- S_i is the measurement prediction covariance on time step i .
- K_i Reflects how much prediction needs correction on time step i.

Here, $\widetilde{X}_i$ is the predicted mean on time step i before seeing the measurement and X_i is the estimated mean on time step i after seeing the measurement. $\widetilde{C}_i$ is the predicted covariance on time step i before seeing the measurement and C_i is the estimated covariance on time step i after seeing the measurement. Y_i is mean of the measurement on time step i. M_i is the measurement residual on time step i. S_i is the measurement prediction covariance on time step i. K_i reflects how much prediction needs correction on time step i.

2.2 Pearson correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. We have used Pearson Correlation Coefficient for analyzing the relationship between the generated features and the number of confirmed cases. Pearson correlation coefficient is a statistic which lies between −1 to +1, and measures linear correlation between any two variables Z and V. +1 means total positive linear correlation. -1 means total negative linear correlation and 0 means no correlation.

Pearson's correlation coefficient can be presented as a ratio of the covariance of the two variables Z and V and the product of their standard deviations.

$$r_{zv} = \frac{\sum z_iv_i - n\overline{z}\,\overline{v}}{(n-1)\sigma_z\sigma_v} \tag{12}$$

Here, n is sample size. z_i and v_i are individual sample points indexed with i. $\overline{z}$ and $\overline{v}$ are sample mean. σ_z and σ_v are sample standard deviation.

After cleaning and preparing the dataset, *Pearson correlation* has been calculated for each pair of features using Eq. 12 and results have been analyzed using *Heat map*. Each column of the heat map represents the dependency of the X-

```
Algorithm (Y,T, X)
Input:
Y : The original measurement matrix.
T :The number of time steps.
Output:
X : The mean matrix.
C : Covariance matrix
A : State relationship matrix
Q : Process noise covariance matrix
B : Control input relationship matrix
J : Control input
M: Measurement residual
begin
    initialise matrices $X, C, A, Q, B, J, M$
    fori←1 to T
        Update $X_i$and $C_i$ as
        $X_i = A_{i-1}X_{i-1} + B_iJ_i$
        $C_i = A_{i-1}C_{i-1}A_{i-1}^T + Q_{k-1}$
        Perform measurement updates as:
        $M_i = Y_i - G_iX_i$
        $S_i = G_iC_iG_i^T + R_i$
        $K_i = C_iG_k^TS_i^{-1}$
        $X_i = X_i + K_iM_i$
        $C_i = C_i - K_iS_iK_i^T$
    end for
end
```

axis parameter on the Y-axis parameters. Heat map of the Pearson Coefficient for pair of features of Covid-19 cases has been shown in Fig. 2. It has been observed from Fig. 2 that confirmed cases have strong positive correlation with growth in 1 day, growth in 3 days, growth in 5 days, and growth in 7 days. Total confirmed cases are highly dependent on the cases reported in previous 7 days.

It has been also noted that prediction model is also having positive correlation with previous day cases. The effect of historical data about the spread has less correlation in prediction as compared to previous day data. It can be seen from Fig. 2 that minimum temperature and maximum temperature are having a weak positive correlation in the spread. It has been seen from heat map that minimum temperature is more crucial than maximum temperature in spread analysis.

2.3 Feature importance through ensemble learning models

Decision Trees are an important type of algorithm for predictive modelling machine learning. The classical decision tree algorithms have been around for decades and modern variations like random forest are among the most powerful techniques available. A decision tree is created by dividing up the input space. A greedy approach is used to divide the space called recursive binary splitting. This is a numerical procedure where all the values are lined up and different split points are tried and tested using a cost function. The split with the best cost is selected. All input variables and all possible split points are evaluated and chosen in a greedy manner. The cost function that is minimized to choose split points, for regression predictive modelling problems, as in our case, is the sum squared error across all training samples that fall within the rectangle:

$$\sum_{i=1}^{n}(y_i–\text{prediction}_i)^2 \tag{13}$$

Here, y is the output for the training sample and prediction is the predicted output for the rectangle.

The problem with decision trees are that they are sensitive to the specific data they are trained in. Changing the training data changes the structure of the tree and hence the predictions differ. They are computationally expensive to train, with high chance of overfitting. A Random Forest is a bagging technique, where the trees run in parallel, without any interaction. It operates by constructing a multitude of decision trees at training time and outputting mean prediction of the individual trees. A random forest is a meta-estimator which aggregates many decision trees. Here, the number of features that can be split on at each node is limited. It ensures that the ensemble model doesn't rely too much on any individual feature. Each tree draws a random sample from the original dataset which prevents over fitting.

Ensemble Learning models have been used to test the importance of various generated features. The generated data has been used to train a Random Forest Estimator, which gave satisfactory test accuracy. Training data has been prepared by combining all the data from various states till 01/05/20 and testing data included data from 02/05/20 till 09/05/

Table 1 Temperature and Humidity data of 15 States of India

S. No	State	Average min Temp	Average max Temp	Average Humidity
01	Andhra Pradesh	27	31	73
02	Delhi	28	37	24
03	Gujarat	28	41	31
04	Haryana	20	33	32
05	Jammu & Kashmir	7	19	62
06	Karnataka	24	35	33
07	Madhya Pradesh	25	39	21
08	Maharashtra	30	34	57
09	Punjab	20	33	32
10	Rajasthan	24	35	29
11	Tamil Nadu	26	32	69
12	Telengana	29	40	27
13	Uttar Pradesh	27	37	24
14	West Bengal	14	37	66
15	Kerala	35	27	71

20.Importance of the various features have been analysed using the Random Forest Estimator model [30]. Random forests have been used because they produce better results by accumulating the relatively uncorrelated decision trees and avoiding over-fitting [31].

Every decision tree of the forest chooses a subset of the feature set and produces a result accordingly. By analysing the result of every such decoupled decision tree, it can be decided that which feature has more ability to guide us to the required result. The more a feature contributes towards minimizing the error, the more its importance. The minimization effect of a feature can be calculated by taking mean of the error values from the trees in which it appears. It can also be said that features that appear on higher levels of the tree are more important as they contribute to relatively higher information gain.

Table 2 COVID-19 data of 15 states of India

Date (2020)	State	Confirmed Cases	Cases 1 day ago	Growth in 1 day	Growth in 3 days	Growth in 5 days	Growth in 7 days	Growth rate for 1 day	Growth rate for 3 days	Growth rate for 5 days	Growth rate for 7 days
04–21	Andhra Pradesh	757	722	35	154	223	284	4.84	25.53	41.76	60.04
04–21	Delhi	2081	2003	78	374	503	571	3.89	21.90	31.87	37.81
04–21	Gujarat	2066	1851	215	794	1195	1449	11.61	62.42	137.19	234.84
04–21	Haryana	254	233	21	29	49	55	9.01	12.88	23.90	27.63
04–21	Jammu & Kashmir	368	350	18	40	68	98	5.14	12.19	22.66	36.29
04–21	Karnataka	415	395	20	44	100	157	5.06	11.85	31.74	60.85
04–21	Kerala	408	402	6	12	20	29	1.49	3.03	5.15	7.65
04–21	Madhya Pradesh	1540	1485	55	185	420	810	3.70	13.65	37.5	110.95
04–21	Maharashtra	4669	4203	466	1346	1750	2332	11.08	40.50	59.95	99.78
04–21	Punjab	245	219	26	43	59	69	11.87	21.28	31.72	39.20
04–21	Rajasthan	1576	1478	98	347	553	697	6.63	28.23	54.05	79.29
04–21	Tamil Nadu	1520	1477	43	197	278	347	2.91	14.89	22.384	29.58
04–21	Telengana	919	873	46	128	221	295	5.26	16.18	31.661	47.275
04–21	Uttar Pradesh	1294	1176	118	325	521	637	10.03	33.53	67.39	96.95
04–21	West Bengal	392	339	53	105	161	202	15.63	36.58	69.69	106.31

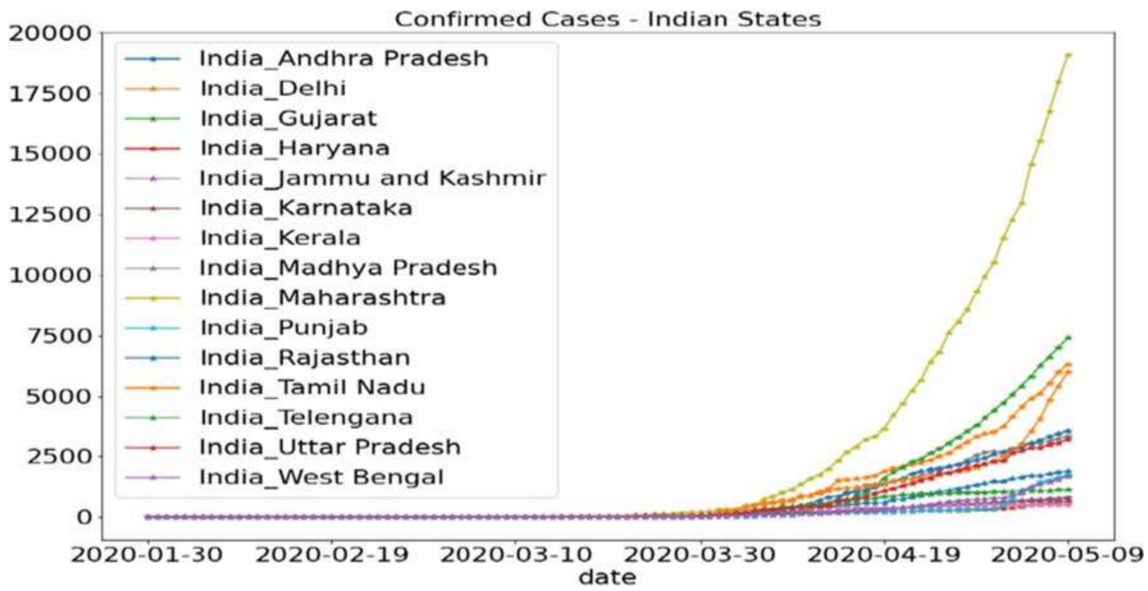


Fig. 1 COVID-19 Spread Scenario of different Indian States

It can be said that random forests provide better results for feature importance as it include results of several decision trees and thus is more generalised than other approaches.

Steps followed:

1. Data collection from Kaggle and world weather online
2. Cleaning the data and preparing a time-series dataset as suitable for our model
3. Generating new features on the dataset to analyse the relation
4. Analyse the correlation between features using Pearson Correlation Coefficient
5. Prepare a train-test dataset (Train- 30/1/20–01/05/20, Test- 02/05/20–09/05/20)
6. Train a random forest regression model on the training set and check for errors using the test set.
7. Extract and Analyse the Scaled Importance value of features using the trained model
8. Apply Kalman filter on the training set for future forecast of next 7 days
9. Evaluate the forecast using Mean Absolute Error

3 Data collection

Study of COVID19 is based on the data available in the open domain. Data of number of confirmed cases of COVID19 in India and different Indian states were collected from Kaggle website (https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset). The website provides the data of COVID19 for every country (https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset). The positive cases of COVID19 in India are also collected from the same (https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset). The experimental data of different Indian states of COVID19 has been taken from website (https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset). It provides different attributes like, Province/State, Country/Region, Confirmed COVID19 cases, Death, and Cured. We collected the data between 31 and 01-2020 to 22-04-2020 for the above study. Various data of weather are collected in order to find out their effects on COVID19 cases. The maximum temperature, minimum temperature, humidity etc. were fetched from available API in python. Pandas and Matplotlib in Python are used to plot these statistics. The source code of this paper is uploaded in the open domain (https://github.com/shazz10/kalman-covid).

3.1 Data processing

Kalman future forecast algorithm was used to predict the future growth of number of cases in India as a whole, as well as Indian state wise. Kalman algorithm requires Time-Series Data as input hence some data pre-processing was done on the collected data. Pandas in Python are used to pre-process these .csv files.

As per record, the first case of Corona in India was reported on 30/01/2020 in Indian state Kerala. Corona virus affects other states of India after few days of the 1st case. The present paper 30/1/2020 is chosen as the starting date and 22/04/2020 as the ending date for data consistency. New row data with default value for confirmed cases as zero was generated for states which reported their first case after few days. From the available features in the data, only Date and Confirmed Cases were selected. Setting the State name as index and increasing

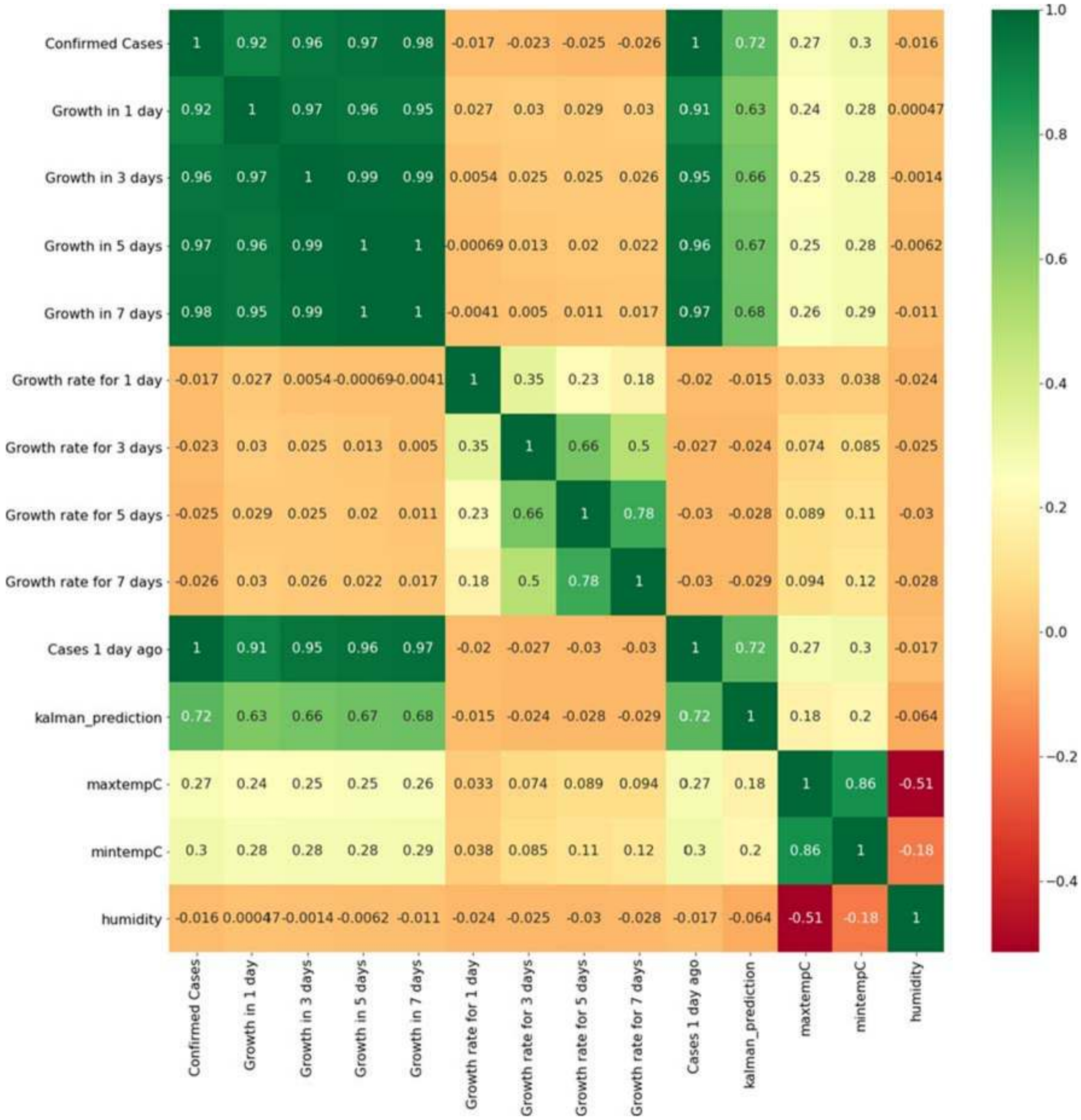


Fig. 2 Heat Map of Pearson Coefficients for COVID-19 Spread

dates as columns; a time series data frame is generated which is ready to be used as input in the Kalman Algorithm. Same process was repeated for Death and Cured Cases.

The statistical relationship between some of self-generated features and number of confirmed cases is established with the help of ensemble-learning models. The features generated to analyse the effect of different parameters in present paper are Confirmed Cases a day ago, Growth in a day, Growth rate in a day $(100 * (Growthin1day) \div (Cases1dayago))$, Growth in 3 days, Growth rate in 3 days, Growth in 5 days, Growth rate in 5 days, Growth in 7 days, Growth rate in 7 days, Max Temperature in C, Min Temperature in C and Humidity. Table 1 represents the temperature and humidity data, during the study period of 15 states of India chosen for this study. Table 2 represents different features of the data collected for spread scenario.

In order to validate the proposed methodology in paper data has been collected from https://www.worldweatheronline.com/ API in python. All the major COVID19 hotspots cities were considered from each states and an average of their data was allotted to the corresponding state.

4 Results and discussion

The proposed methodology has been validated by the cases reported in 15 different states of India. These states of India have been chosen for this study because of the large number of

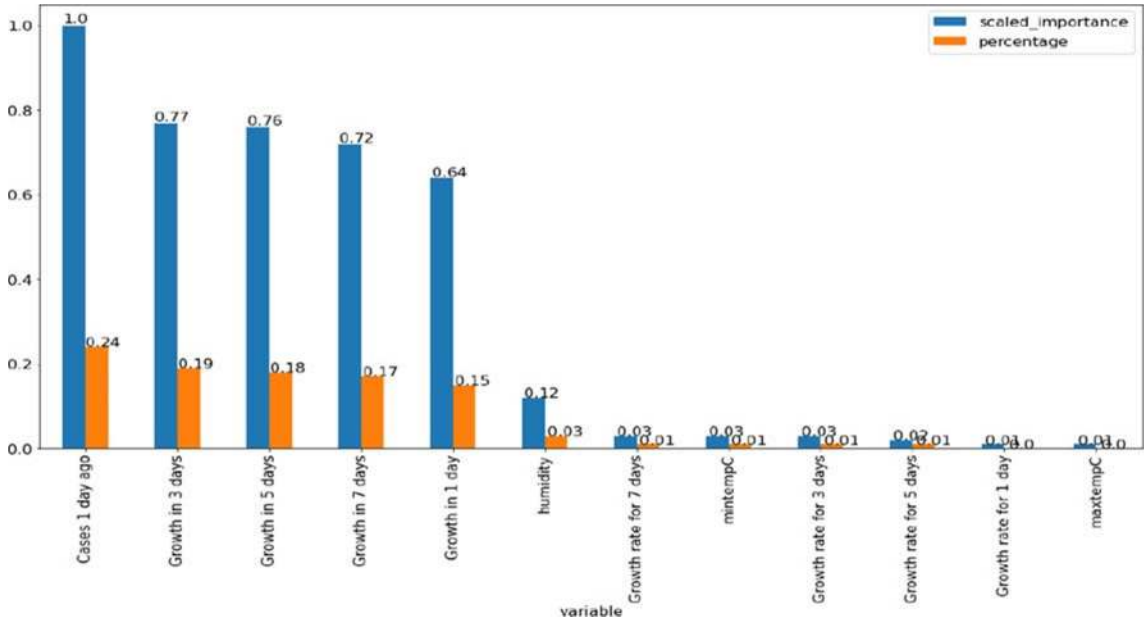


Fig. 3 Scaled Importance of features in COVID-19 spread using Random Forest

COVID19 cases being reported over there. We have chosen top 15 states based on the number of cases reported over there. These states are Andra Pradesh, Delhi, Kerala, Madhya Pradesh, Jammu and Kashmir, Haryana, Karnataka, Gujarat, Maharashtra, Punjab, Rajasthan, Telengana, Tamil Nadu, Uttar Pradesh and West Bengal. We have taken number of total positive cases of COVID19 from all these states for our analysis as well as for the purpose of prediction. Training of the proposed model has been done by using the data between January 30 and May 09, 2020. Fig. 1 shows the spread scenario of COVID19 cases in these states during this period.

After cleaning and preparing the dataset, Pearson correlation has been calculated for each pair of features using Eq. 12 and the results have been analysed using *Heat map*. Each column of the heat map is representing the dependency of the X-axis parameter on the Y-axis parameters. Heat map of the Pearson Coefficient for pair of features for COVID19 cases has been shown in Fig. 2. It has been observed from Fig. 2 that confirmed cases have strong positive correlation with growth in 1 day, growth in 3 days, growth in 5 days, and growth in 7 days. Total confirmed cases are highly dependent on the cases reported in previous 7 days. Hence, it is supporting the well existing idea of COVID19 spread chain i.e., spread chain is dependent on the number of previous cases [32, 33].

It has been also noted that prediction model is also having positive correlation with previous day cases. The effect of historical data about the spread has less correlation in prediction as compared to the previous day data. It can be seen from Fig. 2 that minimum temperature and maximum temperature are having weak positive correlation in the spread. It has been seen from the heat map that minimum temperature is more crucial than maximum temperature in spread analysis.

Train-test split is done on the dataset based on date. Training data belongs to data from January 30, 2020 to May 01, 2020. Validation data has been chosen from May 02–09, 2020. The input features are Confirmed Cases of a day ago, Growth in a day, Growth in 3 days, Growth in 5 days, Growth in 7 days, Growth rate in a day, Growth rate in 3 days, Growth rate in 5 days, Growth rate in 7 days, Maximum Temperature in Centigrade, Minimum Temperature in Centigrade and Humidity and the output in Confirmed Cases. This dataset is then fitted into the Random Forest Regression Model. On evaluating, test dataset gives a Mean Absolute Error of 109.85. This model is then used to

Table 3 Mean Average Error in Validation of prediction model state wise

S. No	State	MAE(1 day)	MAE(7 day)	MAE(15 days)
01	Andhra Pradesh	12.00	46.42	55.336
02	Delhi	7.00	42.42	1599.27
03	Gujarat	35.00	137.42	195.46
04	Haryana	24.00	151.00	68.27
05	Jammu &Kashmir	1.00	24.71	131.6
06	Karnataka	25.00	45.71	38.8
07	Madhya Pradesh	26.00	141.42	162.2
08	Maharashtra	128.00	337.14	3527.06
09	Punjab	133.00	814.14	440.80
10	Rajasthan	29.00	46.85	340.93
11	Tamil Nadu	372.00	1297.71	1214.20
12	Telengana	1.00	52.57	145.27
13	Uttar Pradesh	7.00	38.57	245.40
14	West Bengal	16.00	82.71	375.20
15	Kerala	2.00	35.57	38.80

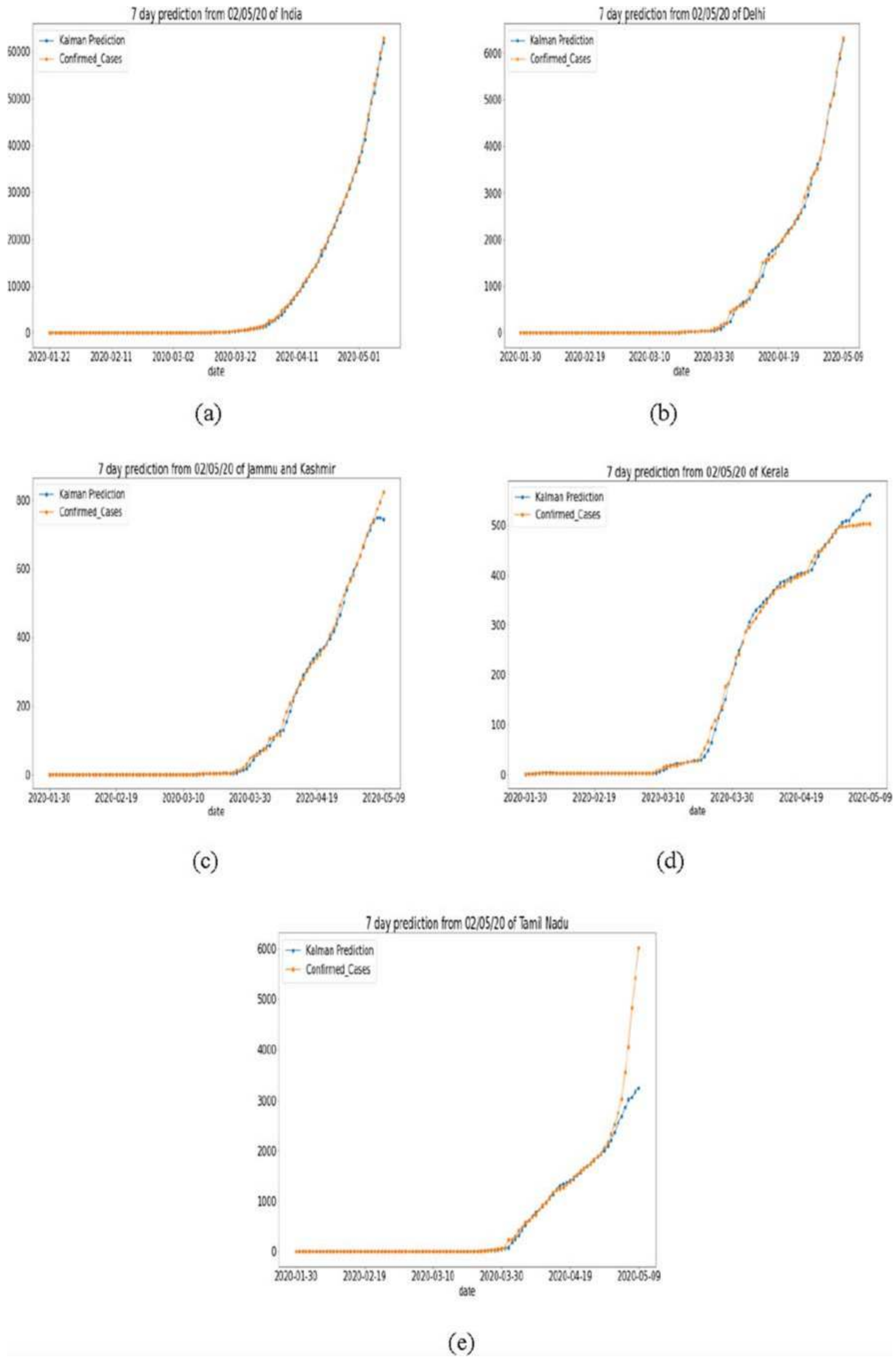


Fig. 4 Validation Results of Prediction model for different states

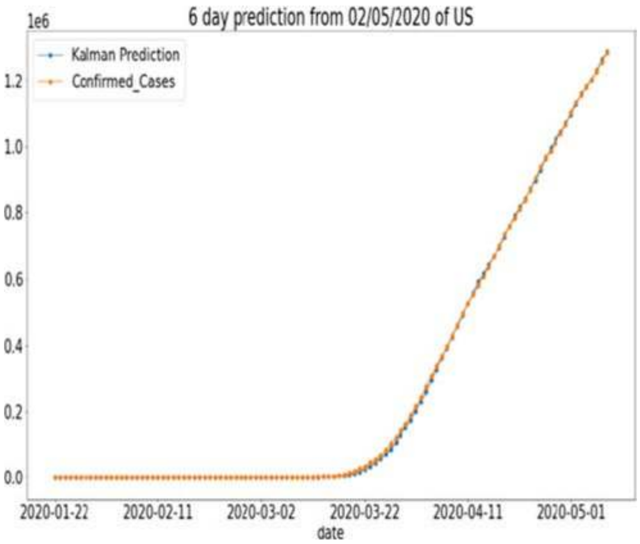


Fig. 5 Validation Results of Prediction model for USA

Table 5 The predicted and actual confirmed cases of China

Date(D/M/Y)	Actual Confirmed	Predicted Confirmed
11/2/2020	42,708	44,032
12/2/2020	44,730	45,461
13/2/2020	46,550	47,105
14/2/2020	48,548	48,632
15/2/2020	50,054	50,513
16/2/2020	51,174	51,128
17/2/2020	70,635	54,219
18/2/2020	72,528	59,703

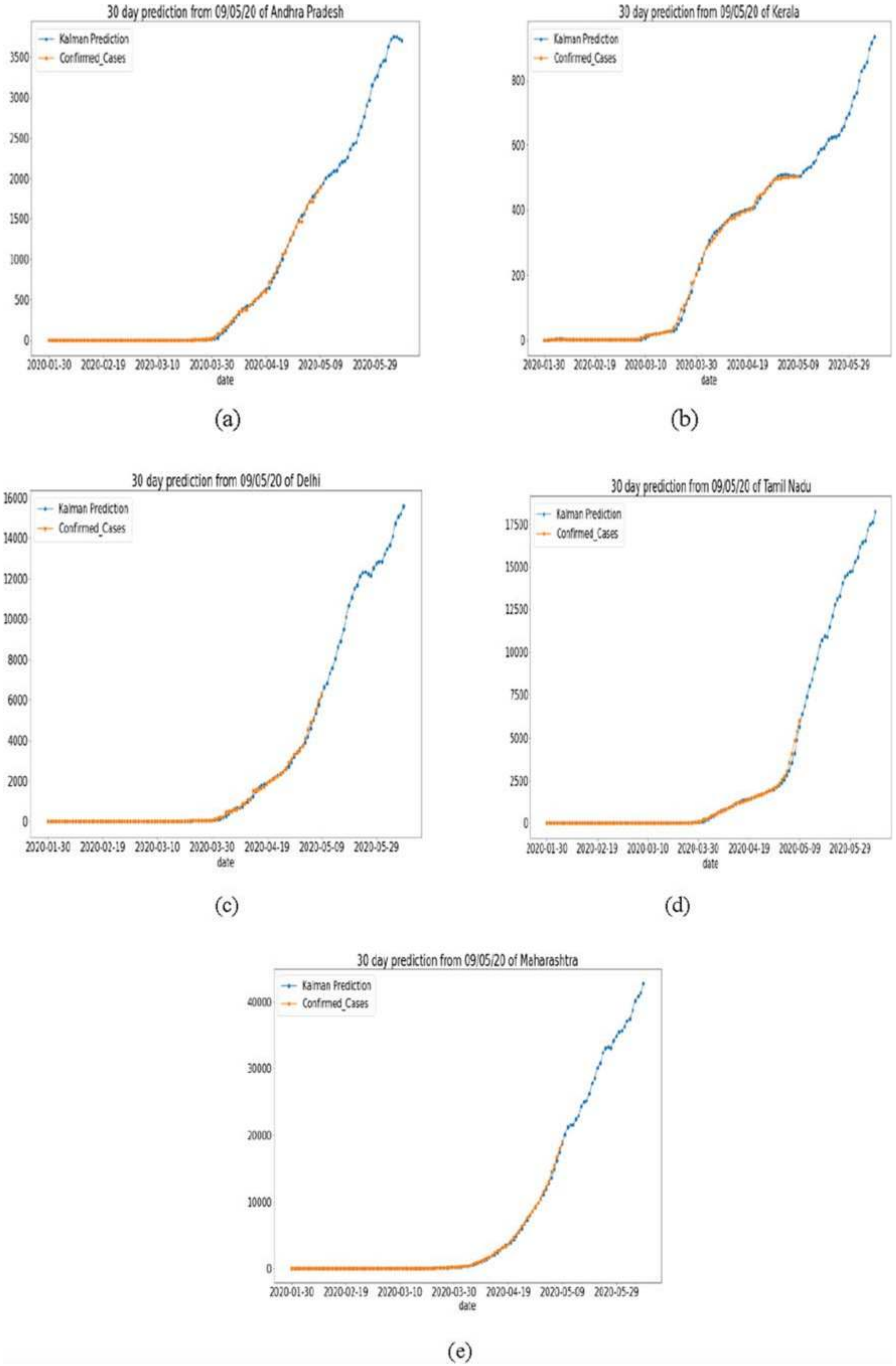
analyse the feature importance of different features on the target variable.

Figure 3 represents the importance of different features on total confirmed cases by using random forest method. It has been noted in Fig. 3 that historical spread data has high importance for the prediction. It has also been seen in Fig. 3 that maximum temperature is having very less importance as compared to minimum temperature. Humidity is also playing a crucial role in the spread and is well noted from the figure. It has been pointed out in Figs. 2 and 3 that humidity is having negative correlation with confirmed cases and with the prediction model as well. It has higher importance for prediction than temperature.

The time-series dataset prepared is then used for future forecast of COVID19 cases; state-wise first then for India. Firstly, our Kalman Filter is applied for the training till May 01, 2020. Validation of our prediction model has been performed by using the dataset from May 02–09, 2020. Predicted data has been compared with the real data for same time period. Table 3 shows the average mean error reported for different states of India in prediction. This error is absolute difference between predicted values and real data. It can be noted that mean average error is varying in the range of 24 to 1297 for different states. It has been pointed out in Table 3 that the validation results are very good except few states like Tamilnadu, Panjab, Maharastra and Gujrat. These states are showing different behaviour of COVID19 spread from the prediction. The reason behind this deviation is the delay in declaration of testing results. Mass results have been declared in one day. Hence, they are showing different behaviour. Fig. 4 shows the validation results of prediction model with state wise data. Some random states have been chosen to show the graphical results. All state data has been shown in Table 3. Fig. 5 shows the validation results of prediction model with the United States of America's (USA) COVID-19 data. The model is trainedwith the confirmed cases of COVID19 in USA between 30/01/2020 to 02/05/2020. The proposed predication model has predicted the confirmed cases of COVID19 from 03/05/2020 to 08/05/2020. The mean average error (MEA) for this case is 2353. The proposed model is compared with the different algorithms existing in the literature. Forecasting of COVID19 cases in China has been done with the help of neuro-fuzzy inference system (ANFIS). They have used enhanced flower pollination algorithm (FPA) by using the salp swarm algorithm (SSA).The main dataset of this

Table 4 The dataset of the COVID-19 used in the paper [34]

Date (D/M/Y)	Confirmed	Date (D/M/Y)	Confirmed	Date (D/M/Y)	Confirmed
21/1/2020	278	31/1/2020	9720	10/2/2020	40,554
22/1/2020	309	1/2/2020	11,821	11/2/2020	42,708
23/1/2020	571	2/2/2020	14,411	12/2/2020	44,730
24/1/2020	830	3/2/2020	17,283	13/2/2020	46,550
25/1/2020	1297	4/2/2020	20,471	14/2/2020	48,548
26/1/2020	1985	5/2/2020	24,363	15/2/2020	50,054
27/1/2020	2741	6/2/2020	28,060	16/2/2020	51,174
28/1/2020	4537	7/2/2020	31,211	17/2/2020	70,635
29/1/2020	5997	8/2/2020	34,598	18/2/2020	72,528
30/1/2020	7736	9/2/2020	37,251		



(a)

(b)

(c)

(d)

(e)

Fig. 6. prediction results for next 30 days for different states of India

Table 6 Comparison of Result obtained from various models on the same dataset

Method	Mean Average Error
ANN	5413
KNN	7671
SVR	5354
ANFIS	5523
PSO	4559
GA	4963
ABC	6066
FPA	4379
FPASSA	4271
Proposed	**4055**

paper was COVID19 dataset of China. It was collected from the website of WHO (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/).

It contains the daily confirmed cases in China from 21 January 2020 to 18 February 2020. 75% from the dataset is used to train the model and rest is used to test. The results were compared to various models such as ANN, KNN, SVR, ANFIS, PSO-ANFIS, ABC-ANFIS, FPA-ANFIS and FPASSA-ANFIS. After training our Kalman Filter model on the same dataset, we made a comparative study based on Mean Average Error; our model performed much better than any models cited in the paper [34]. Table 4 describes the data of confirmed cases of China.

Here training has been done on data till 10/2/2020 and testing on rest. Table 5 describes the predicted values by proposed Kalman Filter based model and actual confirmed cases of China between 11/2/2020 to 18/2/2020. The comparison of the proposed model has been done through different algorithm and is reported in Table 6. It is clear from the table that proposed Kalman filter based model is better for forecasting the COVID19 with existing approaches.

Figure 6 shows the results obtained by prediction model for the next 30 days cases of COVID19 in different states of India. It has been observed that Kalman filter based prediction model shows higher deviation from the real data in long term prediction. Kalman filter based prediction is more accurate for short term prediction. This phenomenon is well supported by Pearson coefficient and random forest based study. Both the studies show that confirmed cases have strong positive correlation and are highly significant for historical spread data. Hence, any error in prediction for a single day can be propagated and can produce the larger error after a few days. Hence, Kalman filter based prediction model is good for short term prediction i.e. Daily and Weekly.

5 Conclusions

The present manuscript presented a prediction model based on Kalman filter. The correlations between different features of COVID19 spread have been studied. It has been found that previous spread data has strong positive correlation with the prediction. The importance of different features in prediction model has also been studied in the present manuscript. It has been noted that historical spread scenario has large impact on the current spread. Hence, it can be concluded that COVID19 spread is following a chain. Hence, to reduce the spread; this chain has to be broken. The proposed prediction model is providing encouraging results for the short term prediction. It has been noted that for a long term prediction, Kalman filter based proposed model is showing large mean average error. Hence, it can be concluded that proposed prediction model is good for short term prediction i.e. daily and weekly. The proposed prediction model can be updated to accommodate long term and medium term series prediction in future.

Funding Work is supported by Ministry of Education, Govt. of India under SPARC scheme with project Id 1396.

Compliance with ethical standards

Conflicts of interest/competing interests Authors have no conflict of interest.

References

1. Hongzhou CWS, Lu YT (2020) Outbreak of pneumonia of unknown etiology in Wuhan. Themystery and the miracle, J Med Virol, China, pp 401–402
2. Zhong L, Mu L, Li J, Wang J, Yin Z, Liu D (2020) Early prediction of the 2019 novel coronavirusoutbreak in the mainland China based on simple mathematical model. IEEE Access 8:51761–51769. https://doi.org/10.1109/access.2020.2979599
3. Zou HML, Ruan F (2020) Sars-cov-2 viral load in upper respiratory specimens of infected patients. New England J Med 382:11771179
4. Williamson G (2020) Covid-19 epidemic editorial. Open Nurs J 14: 37–38. https://doi.org/10.2174/1874434602014010037
5. W. H. Organization, (2020), Novel Coronavirus (2019-nCoV) Advice for the public, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public
6. ZAK Mizumoto, Kagaya K, (2020) Estimating the asymptomatic proportion of coronavirus disease 2019(covid-19) cases on board the diamond princess cruise ship, Yokohama, Japan. Eurosurveillance
7. K Kwok, F Lai, W Wei, S Wong, J Tang, Herd immunity estimating the level required tohalt the covid-19 epidemics in affected countries, J Inf Secur https://doi.org/10.1016/j.jinf.2020.03.027
8. Li L, Zhang Q, Wang X, Zhang J, Wang T, Gao T-L, Duan W, Kam-faiTsoi K, Wang F-Y (2020) Characterizing the propagation of situational information in social media during COVID-19 epidemic: a case study on Weibo. IEEE Trans Computational Social Syst 7(2):556–562

9. Lu H-M, Zeng D, Chen H (2010) Prospective infectious disease outbreak detection using Markov switching models. IEEE Trans Knowledge Data Eng 22(4):565–577
10. Xiaoming LI, XU XI, Wang JIE, Jing LI, Qin S, Yuan J (2020) Study on Prediction Model of HIV Incidence Based on GRU Neural Network Optimized by MHPSO. IEEE Access 8:49574–49583. https://doi.org/10.1109/ACCESS.2020.297985
11. Chandra SK, Bajpai MK (2019) Mesh free alternate directional implicit method based three dimensional super-diffusive model for benign brain tumor segmentation. Comput Mathematics Appl 77(12):3212–3223
12. Kashyap KL, Bajpai MK, Khanna P, Giakos G (2017). Mesh Free based Variational Level Set Evolution for Breast Region Segmentation and Abnormality Detection using Mammograms, Int J Numerical Methods Biomed Eng 34 (1)
13. Singh KK, Bajpai MK (2019). Fractional Order Savitzky-Golay Differentiator based Approach for Mammogram Enhancement, 2019 IEEE International Conference on Imaging Systems and Techniques (IST)
14. Precup R-E, Teban T-A, Albu A, Borlea A-B, Zamfirache IA, Petriu EM (2020) Evolving fuzzy models for prosthetic hand myoelectric-based control. IEEE Trans Instrum Meas 69(7):4625–4636
15. Zhuojun X, Yantao T, Yang L (2015) SEMG pattern recognition of muscle force of upper arm for intelligent bionic limb control. J Bionic Eng 12(2):316–323
16. Precup R-E, Teban T-A, Alves de Oliveira TE, Petriu ElM (2016). Evolving Fuzzy Models for Myoelectric-based Control of a Prosthetic Hand, IEEE International Conference on Fuzzy Systems (FUZZ) 2016
17. Gil RPA, Johanyák ZC, Kovács T (2018) Surrogate model based optimization of traffic lights cycles and green period ratios using microscopic simulation and fuzzy rule interpolation. Int J Artificial Intell 16(1):20–40
18. Precup R-E, Teban T-A, Albu A, Borlea A-B, Zamfirache IA, Petriu EM (2019) Evolving fuzzy models for prosthetic hand myoelectricbased control using weighted recursive least squares algorithm for identification. Proc. IEEE Int. Symp. Robotic Sensors Environ. (ROSE), Ottawa, pp 164–169
19. Acharya UR, Fujita H, Oh SL, Tan YHJH, Adam M, Tan RS (2019) Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. J Appl Intell 49:16–27
20. Ashir J, Zhou S, Yongjian L, Qasim I, Noor A, Nour R (2019) An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection. IEEE Access 7:180235–180243
21. Yin L, Sun Z, Gao F, Liu H (2019) Deep Forest Regression for Short-Term Load Forecasting of Power Systems. IEEE Access 8: 49090–49099
22. Favieiro GW, Balbinot A (2019) Paraconsistent Random Forest: An Alternative Approach for Dealing With Uncertain Data. IEEE Access 7:147914–147927
23. Sharath Srinivasan, The Kalman Filter: An algorithm for making sense of fused sensor insight, https://towardsdatascience.com/kalman-filter-an-algorithm-for-making-sense-from-the-insights-of-various-sensors-fused-together-ddf67597f35e
24. Caron F, Duflos E, Pomorski D, Vanheeghe P (2006) Gps/imu data fusion using multisensory kalman filtering: introduction of contextual aspects. Inf Fusion 7(2):221–230
25. Azeem Iqbal (2019). Applications of an Extended Kalman filter in nonlinear mechanics, PhD Thesis, University of Management and Technology. https://www.physlab.org/wp-content/uploads/2019/06/Thesis-compressed.pdf
26. Phil Howlett, Peter Pudney, Xuan Vu, et al. (2004). Estimating train parameters with an unscented kalman filter. PhD thesis, Queensland University of Technology
27. Wu T, O'Grady P (2004) An extended Kalman filter for collaborative supply chains. Int J Prod Res 42(12):2457–2475
28. Oakes T, Tang L, Landers RG, Balakrishnan SN (2009) Kalman Filtering for Manufacturing Processes, Kalman Filter Recent Advances and Applications, Victor M. Moreno and Alberto Pigazo (Ed.), ISBN: 978-953-307-000-1, InTech, Available from: http://www.intechopen.com/books/kalman_filter_recent_adavnces_and_applications/kalman_filtering_for_manufacturing_processes
29. Hernandez-Matamoros A, Fujita H, Hayashi T, Perez-Meana H (2020) Forecasting of COVID19 per regions using ARIMA models and polynomial functions. App Soft Comp J 96:106610
30. Tony Yiu (2019). Understanding Random Forest, https://towardsdatascience.com/understanding-random-forest-58381e0602d2
31. Synced, How Random Forest Algorithm Works in Machine Learninghttps://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning3c0fe15b6674
32. Avaneesh Singh, Saroj Kumar Chandra, Manish Kumar Bajpai, (2020). Study of Non-Pharmacological Interventions on COVID-19 Spread, medRxiv 2020.05.10.20096974, https://doi.org/10.1101/2020.05.10.20096974
33. Saroj Kumar Chandra, Avaneesh Singh, Manish Kumar Bajpai, (2020). Mathematical Model with Social Distancing Parameter for Early Estimation of COVID-19 Spread, medRxiv 2020.04.30.20086611, https://doi.org/10.1101/2020.04.30.20086611
34. AI-Qaness MAA, Ewees HA, HoryFan MAEIA (2020) Optimization method for forecasting confirmed cases of COVID-19 in China. J Clin Med 9(3):674

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr. Koushlendra Kumar Singh is working as a Assistant Professor, Department of Computer Science and Engineering National Institute of Technology Jamshedpur, India. He received the Ph.D. degree in 2016 from Indian Institute of Information Technology, Design & Manufacturing, Jabalpur, India. He received the master degree from the same institute in computer science and engineering discipline. He graduated the B.Tech in CSE Department from Bhagalpur College of Engineering, Bhagalpur. His current research interests include image processing, epidemic forcasting, biometrics, and different applications of fractional derivatives, computational modeling etc.