BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 06/07 pp116–123
Volume 4, Number 2, June 2021
DOI: 10.26599/BDMA.2020.9020016

Prediction of COVID-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model

Vishan Kumar Gupta*, Avdhesh Gupta, Dinesh Kumar, and Anjali Sardana

Abstract: A novel coronavirus (SARS-CoV-2) is an unusual viral pneumonia in patients, first found in late December 2019, latter it declared a pandemic by World Health Organizations because of its fatal effects on public health. In this present, cases of COVID-19 pandemic are exponentially increasing day by day in the whole world. Here, we are detecting the COVID-19 cases, i.e., confirmed, death, and cured cases in India only. We are performing this analysis based on the cases occurring in different states of India in chronological dates. Our dataset contains multiple classes so we are performing multi-class classification. On this dataset, first, we performed data cleansing and feature selection, then performed forecasting of all classes using random forest, linear model, support vector machine, decision tree, and neural network, where random forest model outperformed the others, therefore, the random forest is used for prediction and analysis of all the results. The K-fold cross-validation is performed to measure the consistency of the model.

Key words: coronavirus; COVID-19; respiratory tract; multi-class classification; random forest

1 Introduction

The virus of coronaviruses (CoV) is a special kind of virus that itself is a disease and it enhances the existing disease in humans body which makes it a very dangerous virus. This virus results in wheezing, hard to breathe, bad digestive system, and liverwort, effects badly human nervous system (center), and also harms animals like cows, horses, and pigs that are kept, raised, and used by people and different wild animals. In 2002–2003 the epidemic of Severe Acute Respiratory Syndrome (SARS) and in 2012 the burst of the Middle East Respiratory Syndrome (MERS) have illustrated the probability of transferrable newly arrived COVID-19 in human to human and animal to human and vice versa, though there are very fewer cases of this kind, they do exists. In late December 2019, the effect of secret pneumonia in the whole world is a noticeable topic of study[1].

In India, the first case of coronavirus disease 2019 (COVID-19) was announced on 30th January 2020. This virus extends to the whole of India (in their different districts) till April 2020 end. In India, the total cases announced were 5734 in which 472 were recovered and 166 people were dead till 9th April 2020. In India, the total cases announced were 236 184 in which 113 233 were recovered and 6649 people were dead till 6th June 2020. After this date, fresh cases are still coming into light daily which is around 10 000. In India, the infection rate of COVID-19 is lesser than that in some other countries till date. The website worldometers[2] gives us all these details in a precise manner. Figure 1 is

• Vishan Kumar Gupta is with Department of Computer Science and Engineering (CSE), Graphic Era Deemed to be University, Dehradun 248002, India. E-mail: vishangupta@gmail.com.
• Avdhesh Gupta and Anjali Sardana are with Department of CSE, IMS Engineering College, Ghaziabad 201009, India. E-mail: avdhesh.gupta@imsec.ac.in; anju.sardana@gmail.com.
• Dinesh Kumar is with Department of CSE, KIET Group of Institutions, Ghaziabad 201206, India. E-mail: dineshvashist@gmail.com.
* To whom correspondence should be addressed.
Manuscript received: 2020-06-17; revised: 2020-08-10; accepted: 2020-08-21

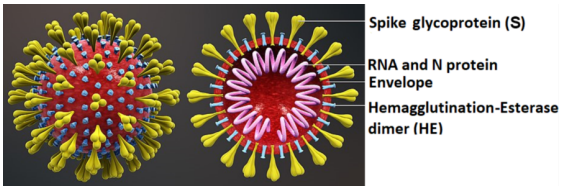


Fig. 1 Structure of coronavirus.

showing the structure of COVID-19, this structure looks like a crown. The different parts of this virus are also introduced in this diagram[3].

The objectives of this surveillance are the following:

(1) Monitor trends in COVID-19 disease at national levels.

(2) Rapidly detect new cases in countries where the virus has started to circulate and monitor cases in countries where the virus is not circulating.

(3) Provide epidemiological information to conduct risk assessments at the national and state level.

(4) Provide epidemiological information to guide preparedness and response measures.

1.1 Transmission

In China, COVID-19 first case was reported in Huanan Seafood Wholesale Market, Wuhan. The main reason which was supposed for the spread of this virus is the transmission from animal-to-human. Even so, the upcoming COVID-19 cases were not related to the subjection method. Hence the conclusion is that virus transmission is from humans to humans, and people with viruses indicative are the main recurrent reason for the spread of COVID-19. Before the symptoms progress, the transmission probability of COVID-19 appears to be very rare, even though, this virus transmission can not be prohibited. Besides these, the advice for every person is that the people who are symptomless or asymptomatic could pass on the virus and social distancing is the only way to be secure from this virus[4].

Including rhinovirus and flu, additional wheezing bacterium, it is believed that the droplets of sneeze and cough of a person are the main reason for virus imparting. In closed places, aerosol transmission is also possible in case of long exposure to deep-mouthed aerosol concentrations. In China, the result of data analysis of SARS-CoV-2 spread is that the close contact of two people is the demanded condition for the spread of the virus. The virus extension is mainly restricted to a person's family members, other nearly contacted people and healthcare experts[4].

1.2 Treatment and prevention

Currently, there is no isolated particular antiviral treatment for COVID-19 virus and their treatments are reassuring. The effects of recombination of IFN with ribavirin are very less against the infection of COVID-19. After the SARS and MERS pandemic, several valuable efforts have been provided for the development of new antivirals targeting the CoV proteases, polymerases, and entry proteins, nevertheless, none of them has been proven to be worthwhile in clinical trials, nevertheless, of them has been proven to be worthwhile in clinical trials. The patient who can already be recovered from COVID-19 can donate their plasma and antibodies, because it has been proved beneficial in the treatment of COVID-19[1].

As well, diverse vaccine schemes, like the use of disabling viruses, live attenuated viruses, a vaccine based on viral vector, subunit injection, recombinant proteins, and DNA vaccines, have been evolved, but they are tested only in the animals so far.

Till now there is not any effective injection or therapy available for COVID-19, but only the finest measures are to control the source of infection, early diagnosis, reporting, isolation, supportive treatments, and on-time producing outbreak details to keep away from inessential anxiety. For every person, a fine exclusive hygiene, wearing a shaped or suitable mask, ventilation, and keeping away from massed areas will assist to block COVID-19 virus or its inflammation[1].

The guidance and directions issued by the World Health Organization (WHO) and other corporations are as follows:

- Keep away from adjacent correspondence with people suffering from serious CoV inflammation.
- Clean your hands regularly, mainly when you come in close contact with CoV-infected people and the place where they live.
- Keep away from unsafe connections with wild and farm animals.
- Persons having symptoms of critical air shaft inflammation should maintain a distance from other peoples, enfold wheeze or sneezes with a throwaway paper napkin or material, and clean their hands from time to time.
- Specifically, in the department of a medical emergency, proper arrangement of strict hygiene measures are required for the prevention and control of infections.
- Individuals that are immunocompromised should

avoid public gatherings.

This paper proposes machine learning schemes based on a data-driven approach. This approach gives a prediction about the number of infected people with COVID-19 in the upcoming days using the available data. This paper proposes a model, which can easily forecast the count of fresh COVID-19 cases, so that the management can make a preparation to handle these cases.

Figure 2 shows the general diagram of the prediction model, where the various features are taken, and their multiple cases (classes) are predicted through random forest prediction model.

This paper is organized as follows. Section 2 explains methodology and materials for the prediction of COVID-19, where we present dataset and its features, feature selection, and all the classes. The procedure of the prediction model is clarified in Section 3. The description of various machine learning models used in this work and their performance metric is presented in Section 4. Sections 5 present the result analysis, comparison of reported and estimated cases. At long last, the conclusion is exhibited in Section 6.

2 Methodology and Material

2.1 Dataset and its features

Coronaviruses are a large family of viruses that may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases, such as MERS and SARS. The most recently discovered coronavirus causes coronavirus disease in 2019 (COVID-19)[5].

The number of new cases is increasing day by day around the world. This dataset has information from the states and union territories of India daily. The effect of preventing measures, like social distancing, face mask, and the lockdown, has also been considered.

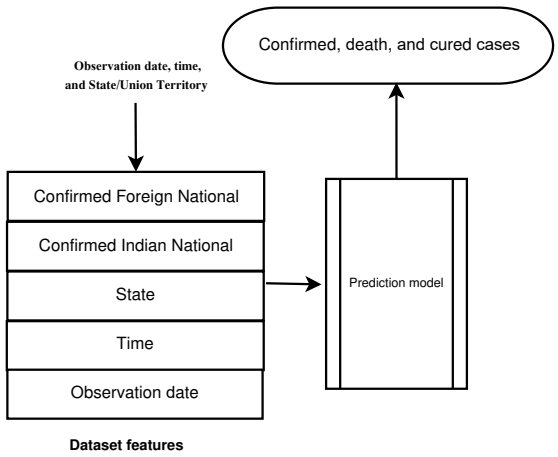


Fig. 2 Prediction method.

The dataset consists of features of COVID-19 data which are taken from https://www.kaggle.com/sudalairajkumar/covid19-in-india/ and also from the Ministry of Health & Family Welfare. The dataset consists only of 2342 samples of COVID-19 cases in India from 30 January 2020 to 26 May 2020. Table 1 shows the attributes/features used in this study and glimpse of dataset is presented in Table 2.

2.2 Feature selection

During the process of model building, feature selection is used to select most relevant features out of all the features. It reduces the complexity of the prediction model. Here, we performed feature selection using random forest importance algorithm in R programming language[6]. The classification model features are calculated using the above algorithm, whose input parameters are all the features of dataset of COVID-19 cases in India.

So, we got three features, which were used for the building of multi-class classification model using a random forest importance algorithm. These are "Obervation date", "Time", and "State/Union Territory" out of five, only two features have been discarded, that are "Confirmed Indian National" and "Confirmed Foreign National". These features are discarded, because they impact only at the beginning of COVID-19 infection, when patients were coming from abroad, later

Table 1 Feature for the prediction of COVID-19 cases in India.

Name	Description
Observation date	It is the date on which how many COVID-19 positive cases have occurred.
Time	It is the time of that particular date at which how mang COVID-19 positive cases have occurred.
State/Union Territory	It is the name of the state/union territory of India where COVID-19 cases were found.
Confirmed Indian National	It is the total number of confirmed COVID-19 cases found in India itself at the starting of SARS-CoV-2 in India.
Confirmed Foreign National	It is the total number of confirmed COVID-19 cases found in India, which came from any foreign countries at the beginning of SARS-CoV-2 cases in India.

Table 2 Dataset on SARS-CoV-2 in India.

Date	Time	State/Union Territory	"Confirmed Indian National" case	"Confirmed Foreign National" case	Cured case	Death case	Confirmed case
30-01-2020	6:00 PM	Kerala	1	0	0	0	1
04-03-2020	6:00 PM	Rajasthan	1	14	0	0	15
07-03-2020	6:00 PM	Telengana	1	0	0	0	1
07-03-2020	6:00 PM	Tamil Nadu	1	0	0	0	1
08-03-2020	6:00 PM	Ladakh	2	0	0	0	2
08-03-2020	6:00 PM	Telengana	1	0	0	0	1
10-03-2020	6:00 PM	Jammu and Kashmir	1	0	0	0	1
11-03-2020	7:00 PM	Maharashtra	2	0	0	0	2
11-03-2020	7:00 PM	Delhi	5	0	0	0	5
29-03-2020	7:30 PM	Andhra Pradesh	–	–	1	0	19
10-04-2020	5:00 PM	Maharashtra	–	–	125	97	1364
29-04-2020	5:00 PM	Gujarat	–	–	434	181	3774
01-05-2020	5:00 PM	Madhya Pradesh	–	–	482	137	2719
26-05-2020	8:00 PM	West Bengal	–	–	1414	278	3816

CoV cases are arisen only based on internal infection due to COVID-19's communicable property. Therefore, the values of these fields are not considered.

2.3 Target classes used in prediction dataset

Our dataset contains three target classes, which have multiple discrete instances. These target classes are the following:

(1) **Confirmed cases:** Number of confirmed cases at any particular date. It may be increased or decreased according to next date, time, and location-specific to the Indian states only.

(2) **Death cases:** Number of death cases at any particular date. It may be increased or decreased according to next date, time, and location-specific to the Indian states only.

(3) **Cured cases:** Number of cured cases at any particular date. It may be increased or decreased according to the next date, time, and location-specific to the Indian states only.

3 Procedure of Prediction Model

We are developing a machine learning-based methodology, which has the following four steps. This methodology is also depicted in Fig. 3.

Step 1: Building multi-class classification model using the training-testing concept. The dataset of COVID-19 features of date-wise, time-wise, and state-wise were taken from Kaggle and then trained and tested at 70% and 30%, respectively.

Step 2: Feature selection. Before going to the model formation, we selected only important features for the reduction of the complexity of the model. For the same,

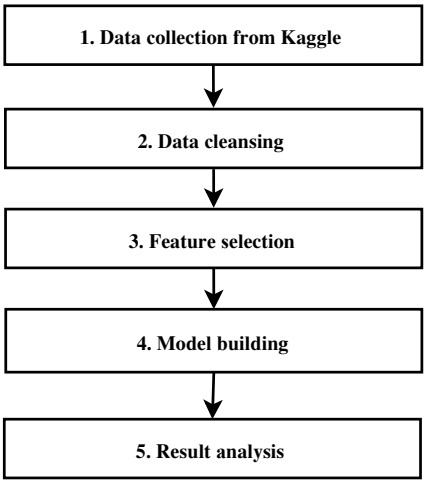


Fig. 3 Methodology of work.

we applied the random forest importance algorithm. Section 2.2 describes it in detail. The formulas for the prediction model in the confirmed, death, and cured cases are the following:

$$\text{Confirmed} \sim f(\text{Observation Date} + \text{Time} + \text{State/Union Territory}) \quad (1)$$

$$\text{Death} \sim f(\text{Observation Date} + \text{Time} + \text{State/Union Territory}) \quad (2)$$

$$\text{Cured} \sim f(\text{Observation Date} + \text{Time} + \text{State/Union Territory}) \quad (3)$$

Step 3: Training the dataset using the multi-class classification. The dataset is then modeled using random forest, Support Vector Machine (SVM), decision tree, multinomial logistic regression, and neural network at 70% training dataset.

Step 4: Testing the dataset using the multi-class

classification: 30% data are tested using these all five models, the results from all the five models are collected and found that the random forest model outperformed the other models for the prediction of confirmed, death, and cured cases, individually. Therefore, we have considered the random forest model for the prediction of this multi-class classification model.

4 Machine Learning Models Used in This Study and Their Performance Metrics

These are the following models used for the prediction of the cases of COVID-19 using multi-class classification:

(1) Decision tree (rpart): To build decision trees, we used rpart() method of R programming language[7, 8].

(2) Random forest (randomForest): It is an ensemble tree-based learning algorithm. The random forest classifier is a set of decision trees from a randomly selected subset of the training set. It aggregates the votes from different decision trees to decide the final class of the test object. We used randomForest() method of R programming language for this algorithm[9, 10].

(3) Multinomial logistic regression (multinome): In statistics, multinomial logistic regression is a classification method that generalizes logistic regression to multi-class problems, i.e., with more than two possible discrete outcomes. We used multinome() method of nnet package of R programming language for this algorithm[11].

(4) Neural networks (nnet): Neural networks are used just for classification as well as for regression. We are using here feed-forward neural networks with a single hidden layer, possibly with skip-layer connections. We used nnet() method of R programming language for this algorithm[7, 11].

(5) Support vector machine (svm): SVM can be used for classification or regression. It represents the input features as vectors, which are projected onto higher-dimensional space. An optimal hyperplane is then constructed for separating the different instances of confirmed, death, or cured cases. We have used svm() method of e1071 package of R programming language for this algorithm[7, 12].

4.1 Performance tuning of the prediction models

Table 3 shows the popular prediction models, which are used in our study, and the packages used by these models are open source libraries in R programming language, licensed under GNU GPL. All packages are used here having some appropriate method for model formation, which are tuned for better results[13].

Table 3 Machine learning models and their tuning parameters.

Model	Method	Required package	Tuning parameter
Random forest	randomForest	randomForest	mtry=2, ntree=500
SVM	svm	e1071	kernal=radial, degree=3
Decision tree	rpart	rpart	usesurrogate=0
Neural network	nnet	nnet	size=10
Multinomial logistic regression	multinome	nnet	maxit=1000

4.2 Accuracy

The accuracy is computed as the percentage deviation of the predicted target concerning the actual target with some acceptable error. It is the main performance evaluation parameter of any machine learning model[7, 14].

$$\text{Accuracy} = \frac{100}{n}\sum_{i=1}^{n} q_i,$$
$$q_i = \begin{cases} 1, & \text{if } \text{abs}(p_i - a_i) \leqslant 2; \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where p_i is a predicted target, a_i is an actual target, and q_i is an arbitrary variable, which contains the absolute difference of predicted target value and actual target value.

5 Result Analysis and Comparison of Reported and Estimated Cases

The number of the total sample for training and testing is 2342 according to different date, time, and states, which are taken from the website of Kaggle. This is the dataset of multi-class classification to foresee confirmed, death, and recovered/cured cases calculated through various decision models, like decision tree, multinomial logistic regression, neural network, SVM, and random forest.

The distribution of data in the training and testing experiments has been set to 70% and 30%, respectively, for all the methods used. Comparative performance of all the methods in the prediction of confirmed, death, and cured cases on accuracy has been highlighted. Accuracy is computed as the percent deviation of the predicted target concerning the actual target. The accuracy has been calculated using Eq. (4), and Table 4 lists the accuracy of all the models. The results show that

Table 4 Multi-class classification accuracy calculated by various machine learning models.

(%)

Model name	Confirmed cases	Death cases	Cured cases
Random forest	83.54	72.79	81.27
Decision tree	77.69	69.11	79.62
Multinomial logistic regression	67.69	66.52	71.96
Neural network	70.28	63.18	69.16
SVM	71.35	70.12	68.27

the random forest method outperforms other machine learning models. Random forest is an ensemble model that uses bagging for sampling, therefore, we found its overwhelming performance in comparison to other models.

In the prediction of confirmed, death, and cured cases on the testing dataset, random forest has the highest accuracy of 83.54%, 72.79%, and 81.27% on confirmed, death, and cured cases, respectively.

Figures 4, 5, and 6 show the histogram for the comparison of accuracy of confirmed, death, and cured cases, respectively, using the random forest model as well as some other models. These results show that the random forest model has outperformed the other machine learning models.

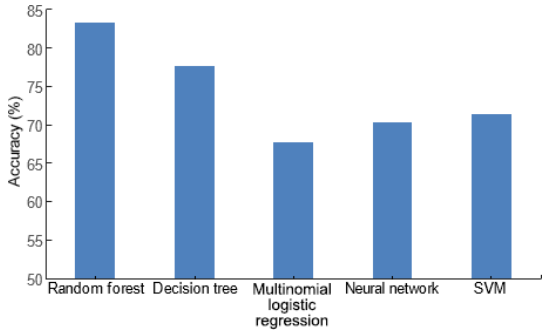


Fig. 4 Performance comparison of random forest model with other models in confirmed cases prediction.

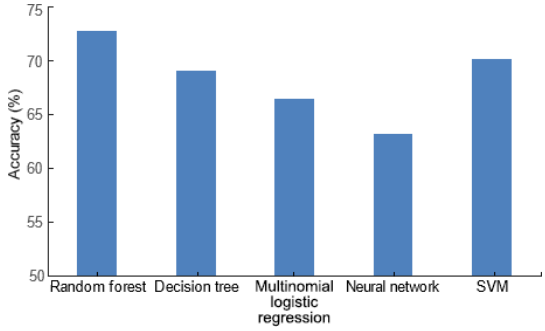


Fig. 5 Performance comparison of random forest model with other models in death cases prediction.

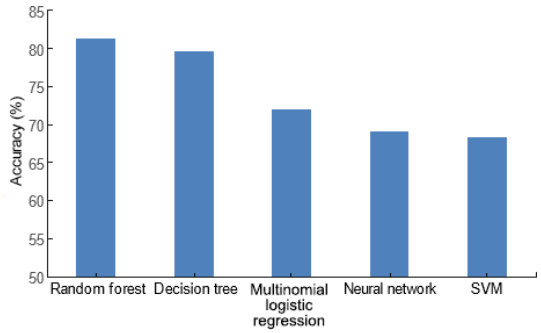


Fig. 6 Performance comparison of random forest model with other models in cured cases prediction.

5.1 K-fold cross-validation

The K-fold cross-validation technique shows the robust performance for the accuracy of any machine learning model[7]. Here, we have used 7-fold cross-validation for the prediction of confirmed, death, and cured cases. In this case, at a time six data frames are used for training and one data frame is used for testing. Table 5 describes the accuracy of random forest model for the different folds of dataset, and Fig. 7 shows the accuracy of the random forest model in the form of a line graph for the prediction of all the target classes, which depicts the consistent performances of random forest model[15].

5.2 Comparison of total reported and estimated confirmed, death, and cured cases

For this data-driven estimations, the data has been taken from 30 January 2020 to 26 May 2020 from different states of India. The comparison has also been made for the daily reported positive confirmed cases with estimated cases (by data-driven model) for some dates and states. Tables 6, 7, and 8 are showing the comparison made by us for the confirmed, death, and cured cases, respectively.

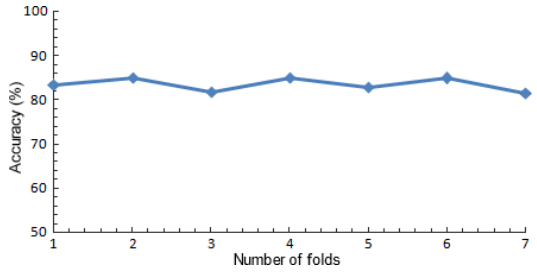
6 Conclusion

We tend to explore five machine learning models with three important features for estimating the confirmed,

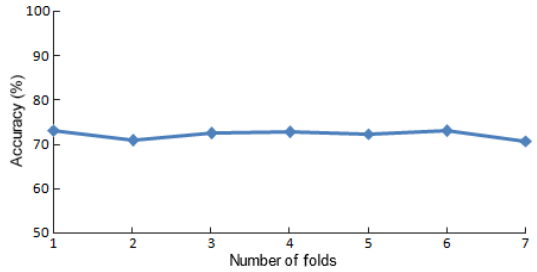
Table 5 Accuracy provided by 7-fold cross-validation.

(%)

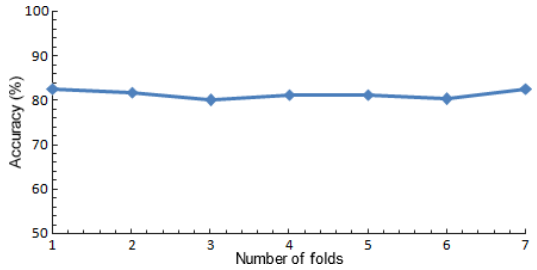
Fold	Confirmed cases	Death cases	Cured cases
1	83.29	72.97	82.52
2	84.98	70.81	81.63
3	81.71	72.35	79.92
4	84.83	72.67	81.17
5	82.65	72.19	81.06
6	84.72	72.88	80.22
7	81.40	70.63	82.44



(a) Confirmed cases



(b) Death cases



(c) Cured cases

Fig. 7 Results of K-fold cross-validation.

Table 6 Comparison of total reported and estimated confirmed cases.

Date	State	Official data	Estimation	Error (%)
08-05-2020	Rajasthan	1596	1520	−4.76
09-05-2020	Bihar	297	305	2.62
21-05-2020	Maharashtra	10 318	10 386	0.60
22-05-2020	Gujarat	5488	5403	−1.50
23-05-2020	Delhi	5897	5912	0.25

Table 7 Comparison of total reported and estimated death cases.

Date	State	Official data	Estimation	Error (%)
08-05-2020	Rajasthan	97	88	−9.27
09-05-2020	Bihar	5	5	0
21-05-2020	Maharashtra	1390	1376	−1.01
22-05-2020	Gujarat	773	740	−4.26
23-05-2020	Delhi	208	256	2.30

death, and cured cases of COVID-19 in the various states of India. The qualitative measures are the confirmed, death, and cured cases. Here, machine learning methods do not embody any additional information from different models or different templet structures. All the models are evaluated on accuracy. Through the intensive experiments, it is found that the random forest method

Table 8 Comparison of total reported and estimated cured cases.

Date	State	Official data	Estimation	Error (%)
08-05-2020	Rajasthan	3427	3514	−2.53
09-05-2020	Bihar	571	601	5.25
21-05-2020	Maharashtra	39 297	38 920	0.90
22-05-2020	Gujarat	5488	5403	−1.50
23-05-2020	Delhi	12 319	12 356	0.30

outperforms other machine learning methods, therefore, we considered it as a final prediction model for the prediction of our various cases. The K-fold cross-validation is used to measure the consistency of random forest model, which provided nearly linear performance to the prediction of all these cases.

Acknowledgment

We are very much thankful to the Indian Ministry of Health and Family Welfare (MoHFW) for making the data available to the general public. Thanks to covid19india.org for providing the individual states level details to the general public. We are also thankful for Kaggle and the worldometer website, which provide huge data in date-wise to perform data analytics.

References

[1] Y. Chen, Q. Liu, and D. Guo, Emerging coronaviruses: Genome structure, replication, and pathogenesis, *Journal of Medical Virology*, vol. 92, no. 4, pp. 418–423, 2020.

[2] Coronavirus cases, https://www.worldometers.info/coronavirus/country/india/, 2020.

[3] Diagram of coronavirus virion structure, https://commons.wikimedia.org/wiki/File:3D_medical_animation_corona_virus.jpg, 2020.

[4] M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn, and R. D. Napoli, *Features, Evaluation and Treatment Coronavirus* (*COVID-19*). Treasure Island, FL, USA: StatPearls Publishing, 2020.

[5] Kaggle dataset for COVID-19 in India, https://www.kaggle.com/sudalairaj kumar/covid19-in-India, 2020.

[6] V. K. Gupta and P. S. Rana, Ensemble technique for toxicity prediction of small drug molecules of the antioxidant response element signalling pathway, *The Computer Journal*, doi: 10.1093/comjnl/bxaa001.

[7] J. Han, J. Pei, and M. Kamber, Data mining: Concepts and techniques, *Data Mining Concepts Models Methods & Algorithms Second Edition*, vol. 5, no. 4, pp. 1–18, 2006.

[8] rpart–the r package for decision tree, https://cran.rproject.org/web/packages/rpart/rpart.pdf, 2020.

[9] Random forest model, https://towardsdatascience.com/random-forest-classification-and-itsimplementation-d5d840dbead0, 2020.

[10] randomforest–the r package for statistical computing, https://cran.rproject.org/web/packages/randomforest/randomforest.pdf, 2017.

[11] nnet–the r package for neural network, https://cran.rproject.org/web/packages/nnet/nnet.pdf, 2017.

[12] e1071–the R package for statistical computing, https://cran.rproject.org/web/packages/e1071/e1071.pdf, 2019.

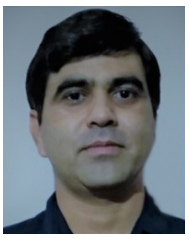
[13] V. K. Gupta and P. S. Rana, Activity assessment of small drug molecules in estrogen receptor using multilevel prediction model, *IET Systems Biology*, vol. 13, no. 3, pp. 147–158, 2019.

[14] V. K. Gupta and P. S. Rana, Toxicity prediction of small drug molecules of androgen receptor using multilevel ensemble model, *Journal of Bioinformatics and Computational Biology*, vol. 17, no. 5, pp. 1–26, 2019.

[15] V. K. Gupta and P. S. Rana, Toxicity prediction of small drug molecules of aryl hydrocarbon receptor using a proposed ensemble model, *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 24, no. 4, pp. 2833–2849, 2019.

Vishan Kumar Gupta received the PhD degree from Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab, India in 2020. He received the BEng degree in computer science & engineering from Rajiv Gandhi Technical University, Bhopal, MP, India in 2005, and the master degree in information technology from ABV-Indian Institute of Information Technology and Management, Gwalior, MP, India in 2007. Currently, he is working as an assistant professor at Graphic Era Deemed to be University, Dehradun, Uttarakhand, India. His areas of research are computational biology, data mining, and machine learning.

Avdhesh Gupta received the PhD degree from Gurukula Kangri Deemed to be University, Haridwar, Uttarakhand, India in 2012. Currently, he is working in Institute of Management Studies (IMS) Engineering College, Ghaziabad, India. He has authored/co-authored several research papers and guided two PhD candidates. He is a reviewer of various journals and conferences. He has more than 20 years of experience in academics. His research areas are network security, data mining, data analytics, and image processing.



Dinesh Kumar received the MTech and BTech degrees from Kurukshetra University, Kurukshetra, Haryana, India in 2006 and 2004, respectively. Currently, he is working in KIET Group of Institutions, Ghaziabad, India. His research areas are machine learning, data base security, and theory of computer science.

Anjali Sardana received the MTech degree from Department of Computer Science and Engineering, GJUS&T, Hisar, Haryana, India in 2001, and the BTech degree in information technology from TITS Bhiwani, Haryana, India in 2008. Currently, she is working in IMS Engineering College, Ghaziabad, India. Her areas of research are computational biology, machine learning, and artificial intelligence.