Analysis of origin, risk factors influencing COVID-19 cases in India and its prediction using ensemble learning

N. Rajesh · Prasanna Ranjith Christodoss

Abstract A genetically heterogeneous infectious virus recognized as COVID-19 (CoV-19) has a significant influence on worldwide health. COVID-19 is an infection transmitted by a viral disease known as Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS-CoV-2). In the current period, the instances of COVID-19 outbreaks are spreading significantly over the globe. Swallowed (through inhalation) or contact with contaminated droplets transmits the illness. Symptoms appear from two to fourteen days after exposure. Many individuals have a symptomless infection. The emphasis of therapy is on fundamental human needs; further research is needed on the risk analysis of the disease and use of antiviral medications. This research focuses on the case instances through the chronological dates in various parts of Indian states. The primary analysis of case instance is dependent on three stages; they are confirming stage, fatality stage, and recovered stage. The relevant datasets are obtained through Kaggle (from January 2020 to May 2021). The dataset has multiple categories that further intensify to perform high-end classifications. Initially, data preprocessing is done that include cleaning and feature selection on the prepared dataset, and then the prediction process is carried out using an Ensemble training approach. Approaches of Ensemble learning utilize many learning algorithms to achieve improved prediction efficiency. The stacking mechanism is used for aggregating accuracy. From the experimental observations, the suggested learning strategy is found to attain the maximum degree of precision for each stage (confirmed: 84.37%, fatality: 82.13%, and recovered: 88.67%) compared to the other approaches that are chosen to perform alone. Assessing the stability of the learning model is done by bootstrapping validation.

Keywords COVID-19 · Ensemble learning · Bootstrapping validation · K-fold cross-validation · Risk analysis · Fatality

1 Introduction

In 2019 Wuhan Metropolis of China’s Hubei Region has proven to be the source of the major virus outbreak, SARS-CoV-2, referred to as the “Severe Acute Respiratory Syndrome-CoronaVirus-2” (Wang et al. 2020). The virus began with one person, has spread worldwide, and shows no signs of stopping. This study provides an eagle-eye perspective of this new pathogen. The reader and practitioners are encouraged to educate themselves often to keep abreast of the latest developments. Coronaviruses have crown-like extensions (spike) on their surface. The size of the virus ranges approximately around 0.06–0.14 microns in diameter.

1.1 Origin and transmission consequences

Till now, two rare instances have occurred in the previous couple of decades whereby wildlife beta corona pathogens have spread to people, resulting in an outbreak of acute...
sickness. The first of its kind indication was only recorded in the year 2002–03. A novel variant named β genera originated in Chiroptera (bats—the primary host of the Coronavirus). Later then traversed on to human beings carried through palm civet cats (intermediary host) in China’s Guangdong region. A new virus (that first appeared in the Republic of China and Hong Kong) afflicted around 8,422 individuals and killed 916 (11% fatality rate) until being stopped (Chan-Yeung 2003).

In the year 2012, the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) (Middle East Respiratory Syndrome Coronavirus 2020), which is believed to have come from Chiroptera, first surfaced in Saudi Arabia where the mammal hosts included dromedary camels. Two thousand four hundred ninety-four individuals were infected, and 858 individuals died (a 34% mortality rate) until being stopped (Chan-Yeung 2003).

Residents of Wuhan city, a key transit hub for China, emphasize acute illness of unknown etiology to the medical centers in the month of December 2019. The early instances might have been linked via their earlier contact to that same Huanan commercial fish market, which often dealt with life (surviving) animals. Immediately following the SARS epidemic, an observation strategy was implemented to monitor victims’ respiratory functioning and their samples for the etiology of the disease. At the end of December of 2019, China alerted the WHO (World Health Organization) about the epidemic, and on New Year’s Day, December of 2019, China alerted the WHO (World Health Organization) about the epidemic, and on New Year’s Day, the commercial fish market of Huanan was shut down. As a result, the pathogen was shown to have 70% and 95% congruence and resemblance with both the SARS-CoV and bat Coronavirus.

Additionally, samples extracted from the seafood market of Huanan showed evidence of the virus having originated locally and so were very certainly transported through seafood (Xinhua 2020). The proportions began to rise dramatically, mostly of which were not exposed to the actual livestock market, indicating that person-to-person transmission took place (Huang et al. 2020). The first recorded death instance due to the virus infection occurred on January 11, 2020. Due to the enormous number of Chinese migrants over the Lunar New Year in China, the disease spread quickly. There have also been many instances documented in persons who migrated to other regions of China and other nations. On January 20, 2020, the infection to healthcare staff/professionals was reported.

Wuhan’s inhabitants of 0.011 billion were put under complete lockdown (implemented limits on entering and departing the city) on 23rd January. Soon similar restriction was implemented in other Hubei regions. Virus-affected cases were recorded in places outside China without any immigration status into the country, indicating that in those nations, the new social transfer occurred (Rothe et al. 2020). An increasing number of airports worldwide started to implement screening measures to identify symptomatic individuals who just repatriated from overseas (especially the Chinese region), and those infected persons are eventually isolated after the resultant confirmation through their blood tests. By the time this was clear, the illness might be spread via asymptomatic persons. This level of dissipation indicates the disease had already begun spreading before any symptoms started. Most of the world nations, therefore, started to evacuate their nationals from China. Also, every tourist coming home from China is screened and isolated for around fourteen days.

Cases have progressed exponentially, and epidemic doubling time in assessments had been observed to be 1.8 days (Li et al. 2020). In reality, on 12 February, Chinese patients with a diagnostic, epidemiological, and radiographic profile of COVID-19 modified the predefined facts of positive cases since there is a steep rise in COVID-19 cases (15,000/single day) (World Health Organization 2020). It is crucial to highlight the spreading unevenness that while most nations record a decline of new instances while in other nations, the cases have surged enormously.

Many nations saw a 2-wave trend throughout a pandemic (2020 & 2021), early spring wave, and the present second phase in late summer. A wide variety of scientific evidence demonstrates that the features of the pathogen’s impacts might differ across the two time periods. The subsequent (second wave) outbreak of the pandemic across all parts of India was more destructive when compared to the first wave. To put it another way, it seems to be distinct over the previous year’s spike in various aspects, all of which serve to increase concerns and worry.

According to health professionals in India, it is estimated that most citizens, around 80 to 85 percent, are asymptomatic (undiagnosed). From the observation, it is found that asymptomatic survivors of the COVID 19 virus may still spread the infection in a confined indoor environment. Around March 2021, a faster and speedier than the first, the second wave of virus swept throughout the nation, bringing scarcity of vaccinations, hospital wards, other medical facilities, and emergency equipment like oxygen tanks. At the end of April 2021, India had surpassed all other countries in the number of active instances. The first nation to identify above 400,000 new infections in a single day occurred on the 30th of April month. The severely affected first ten states in India and their COVID-19 condition are depicted in Table 1. In the early period of May month, the top scientific experts of the Indian
government indicated that the third wave is all but certain. At this moment, it is unclear how long this third wave will last. At the community level across the nation, the implementation of proper guidance will make a difference. As of June 4, 2021, there have been approximately 171,782,908 cases recorded globally (28,574,350 in India) which is elaborately depicted in Fig. 1.

### Table 1 State level COVID-19 condition

| State/UTs        | Total cases | Active | Discharged | Deaths |
|------------------|-------------|--------|------------|--------|
| Maharashtra      | 58,05,565   | 1,99,736 | 55,07,058  | 98,771 |
| Karnataka        | 26,69,514   | 2,80,207 | 23,58,412  | 30,895 |
| Kerala           | 26,01,082   | 1,74,933 | 24,16,639  | 9,510  |
| Tamil Nadu       | 21,95,402   | 2,68,968 | 19,00,306  | 26,128 |
| Andhra Pradesh   | 17,38,990   | 1,33,773 | 15,93,921  | 11,296 |
| Uttar Pradesh    | 16,96,325   | 22,877   | 16,52,417  | 21,031 |
| Delhi            | 14,28,449   | 8,060    | 13,95,892  | 24,497 |
| West Bengal      | 14,11,448   | 53,023   | 13,42,391  | 16,034 |
| Chhattisgarh     | 9,78,220    | 26,977   | 9,38,081   | 13,162 |
| Rajasthan        | 9,44,500    | 24,004   | 9,11,897   | 8,599  |

2 Risk factors and assessments

There are three determinants in the adverse effects of SARS-CoV-2 infection:

a. **Exposure** The probability that they may come in touch with possible virus sources throughout the day-to-day activities, on a scale ranging from “not really exposed” to “totally exposed (i.e., 1 to 5).”

b. **Nearness** The fundamental characteristics of normal activities of any human cannot provide an acceptable social distance. Here, the characteristic was measured by scaling from “work performed alone in a normal day time” to “work performed closely with everyone in a single day time” (i.e., 1 to 5)

c. **Aggregation** An affliction associated with work tasks which may result in interactions with infected peers apart from coworkers highlighted in this section in two major classes:

i. Restrictive connectivity of a foreign entity as 1.00,

ii. Intrinsic existence of foreign entity as 1.15,

iii. Minimalized governable integrative processes as 1.30,

iv. Maximal ungovernable specific processes as 1.50.

Both exposure and nearness denote the risk of exposure to possible infection vectors and physical closeness to others while working. Figure 2 provides an overview of a few of the existing risk assessment tools and applications.

Tracing the afflicted individuals or detecting a suspected illness via signs are the two primary risk assessment methods were currently used. The authorities, as well as the public, may use these resources to determine whether they would have come into interaction with someone infected.
Several commercial applications have been introduced to help in personal risk evaluation and tracking even in the middle of privacy issues.

A social application called the 'Arogya Setu' (Chaturvedi 2020) was inducted by the Indian government as a COVID-19 risk assessment tool, which uses questions to enlighten people on essential information and identify and analyze risk behavior social abidance. In addition, this tool contains guidance on levels of anxiety so that users may understand the total requirement for societal care. In both the Arogya Setu application and other concern tools, there is no mention of smoking, identified as a health risk. Several unique features are found in the program, including evaluating risk depending on residence type that is lacking in other techniques and systems.

2.1 Guidance and directions

WHO and other medical organizations give the following guidelines and instructions to prevent the spread of COVID-19 infection (Dhiman et al. 2021):

i. Maintain a strategic distance from individuals with severe CoV illness, especially during communication.

ii. People should wash their hands often, as much as they come into touch with persons who are sick with the contagious virus and the locations wherever they reside.

iii. Don’t engage in unethical or dangerous connections with animals.

iv. People who are experiencing signs of vital breathing tract infection should stay far away from others, be supposed to cover their nose and mouth when they sneeze or cough, and often sanitize their hands.

v. Further mainly, strong cleanliness precautions are essential to preclude and manage infections in the sector of a medical checkup and primary conditions (emergencies).

vi. Immunodeficient persons must avoid large social/public assemblies.

This research offers a statistics-driven approach to the Ensemble learning technique. This analysis provides a forecast of the volume of COVID-19 infected individuals in the subsequent weeks using the evidence collected. This study offers a framework that can readily predict the number of new instances of COVID-19, especially for the administrations, to make them prepare for any worst scenarios.

3 Material consideration and modeling

COVID-19 is a broad group of viruses, which can infect both animals and humans. A range of Coronaviruses, from cold or flu to much more serious conditions such as SARS and MERS, are reported in humans, which can cause respiratory illnesses. New instances are on the rise all over the globe every day. The dataset utilized in this study includes detailed data about various affected states in India.

The impact of preventative actions such as social separation, lockdown, and mask-wearing has also been addressed. The data sample comprises characteristics from two different datasets: COVID-19 information obtained from the MoHFW (Ministry of Health & Family Welfare) and the Kaggle website. The dataset comprises just 5000 instances from the first wave (February 1, 2020, to May 30, 2020) and 5000 from the second wave (March 25, 2021, to May 30, 2021) of virus-affected patients in India.

3.1 Feature selection

Feature selection is utilized throughout the modeling process to choose the most appropriate characteristics. By simplifying the prediction model, it decreases the intricacy. We used Ensemble learning algorithmic in R language (Gupta and Rana 2020) to do the feature selection process. The classifier features are computed through methodology, as mentioned earlier, wherein the parameter values are all characteristics of recorded COVID-19 datasets across India. We focused on six different characteristics that led to the construction of non-linear and linear classifiers using an Ensemble learning algorithm. They are delineated in Table 2.
3.2 Target classes

The preferred dataset incorporates three different target classes with many distinct instances. They are:

**Confirmed cases (CC)** At any given day, d at time, t numbers of positive COVID-19 cases. CC might raise or lower by the following d, t, and L across India. In confirmed instances, the formulation to prediction is,

\[
CC_t \sim f \left( \frac{(n)_{d,t}}{N} \right), \quad 0 \leq N \leq \infty
\]  

where, N represents the total number of reported cases.

**Fatality cases (FC)** Rate of deaths on a given day, d at time, t. FC might raise or lower by the following d, t, and L across India.

\[
FC_t \sim f \left( \frac{(n)_{d,t}}{N} \right), \quad 0 \leq N \leq \infty
\]  

**Recovered cases (RC)** At any specific date (d) at time (t), the total number of instances recovered from COVID-19. This might vary by the following d, t, and L across India.

\[
RC_t \sim f \left( \frac{(n)_{d,t}}{N} \right), \quad 0 \leq N \leq \infty
\]  

Table 2  Feature for COVID-19 prediction in India

| Time (t) | A given time during which indicates the volume of COVID-19 positivity has happened at that particular time |
| Date (d) | It is COVID-19 specified date, indicating the confirmation of positive cases through an appropriate test on that particular date |
| Location (L) | This represents the occurrence of COVID-19 instances and its detection in the specified State/UTs of India |
| Active cases (AC) | This is defined as the amount of cases reported divided by the amount of cured patients and deaths. Currently, it is the amount of potential patients (current) that are considered infectious |
| Recovery Rate (RR) | RR is the ratio of the percentage of people recovered or discharged with certain health status (optimal) to the total number of diagnosed COVID-19 cases |
| Fatality rate (FR) | FR is the ratio of the percentage of people who died with severe health distortion to the total number of diagnosed COVID-19 cases |

The values of the top 10 affected states are taken into account*

## 4 Prediction process using ensemble learning

Meta-algorithms, such as ensemble techniques, use multiple Machine Learning (ML) algorithms and methodologies to increase predictive accuracy, referred to as stacking, to reduce variability known as bagging, and minimize biasing known to be boosting. Integrating algorithms uniquely is quite intriguing. Similarly, integrating results from several learners is what we utilize a learner for. Using one integrating learner causes the reduction of either variation inaccuracy or bias.

The initial process involves using all the primary learners in tandem, including learners who have learned anything influencing the others. The Ensemble learning approach known as stacking is used in the following stage, in which regression or classification models (Karthikeyan et al. 2019) termed “meta-classes” are stacked by using a meta-regressor/classifier. Training models is accomplished by feeding them every set of relevant training data, then training the conceptual by its outcomes as features are done. The generalized process of Ensemble learning is depicted in Fig. 4.

This research proposes an Ensemble learning approach comprised of five algorithms. They are Naïve Bayesian (NB) (Lin et al. 2017), Decision Tree (DT) (Han et al. 2006), Support Vector Machine (SVM) (e1071—the R package for statistical computing 2019; Han et al. 2006), Neural Networks (NN) (nnet—the r package for neural network 2017), and finally, K-Nearest Neighbors (KNN)
The baseline learning algorithm is often composed of several learning algorithms, and thus, stacking ensembles consequently include heterogeneity quality. Stacking is shown using the algorithmic procedure in the Table 3.

### 4.1 Accuracy estimation

The proportion difference between the predicted target and the real target is calculated to measure the accuracy (Accu). This is compared to an error threshold, after which the estimated value is termed accurate. This ML model’s critical success assessment parameter is used to judge the attainment of the goal (Han et al. 2006; Gupta and Rana 2019).

\[
Accu = \frac{100}{n} \sum_{i=1}^{n} \delta_i, \delta_i = \begin{cases} 
1, & \text{Obs}_i - RT_i \leq 2 \\
0, & \text{Otherwise}
\end{cases}
\]

In Eq. (4), \(PT_i\) represents the predicted target, whereas \(RT_i\) represents the real target, and \(\delta\) denotes the arbitrary factor, which holds the actual error between the expected target value and the true target value.

### 5 Comparative result analysis of stated and predicted cases

According to different occasions (date), periods (time), and states, the total amount of 10,000 samples (5000 from the first wave-February 1, 2020, to May 30, 2020 and 5000 from the second wave-March 25, 2021, to May 30, 2021) is provided to train and test the proposed approach.

These samples are obtained on Kaggle’s website (https://www.kaggle.com/sudalairajkumar/covid19-in-india). This dataset represents the multi-class categorization which uses several decision models, such as DT, NN, SVM, K-NN, and NB, to forecast CC, FC, and RC. The distributive sample mean in training was fixed at 75%, and for testing fixed at 25% for all techniques utilized. All of the methodologies in the estimation of CC, FC, and RC have been assessed in terms of their accuracy relative to each other. Accuracy is calculated as the percentage of expected target variation from the real target. Equation 4 was used to compute the accuracy, and the accuracy of all the models is listed in Table 4. The Proposed Ensemble model performed better than all the other ML models in terms of predicting outcomes. Ensemble model employs stacking for estimating final prediction; consequently, we noticed its astounding effectiveness in contrast to different approaches.

The Ensemble model had the best prediction accuracy of CC, FC, and RC with an outcome of 84.37%, 82.13%, and 88.67% correspondingly. All the three histograms (a, b, and c) from Fig. 5 shows the accuracy of the CC, FC,
RC instances by utilizing both Ensemble and other models. These findings suggest that the methodology of the proposed Ensemble learning model has excelled the performance of other ML models.

5.1 Consistency analysis

K-fold cross-validation (Han et al. 2006) is used to determine the robustness of any ML model’s ability to accurately predict values for any variable. For the estimation of CC, FC, and RC, we employed fivefold cross-validation. For training purposes, four dataset folds (DF1, DF2, DF3, DF4) are utilized at (t), and one dataset fold (DF5) is utilized for testing. Thus, there are eight values in maximum depth, eight values in minimum samples (leaf), and three maximum features. As a result, 192 distinct models can be trained. Mainly in the fivefold method, every option is executed five times. That’s equal to 960 iterations. The exactness of the Ensemble approach for the various dataset folds is described in Table 5. The Ensemble approach demonstrates consistent results in the trend line format depicted in Fig. 6, as it was able to forecast all of the three predefined target cases.

Statistics estimates have been made based on extracted data samples from the first (February 1, 2020, to May 30, 2020) and second waves (March 25, 2021, to May 30, 2021) that covered the top ten severely affected states in India. It was also stated that daily confirmed instances are compared to their corresponding model-estimated cases for various (d), (t), and (L). A comparison of CC, FC, and RC is shown in Fig. 6.

6 Conclusion and future plan

This work developed an Ensemble learning approach that integrates primary learners’ (NB, DT, SVM, K-NN, and NN) accuracy results. The stacking technique is utilized to consolidate the base model’s prediction outcomes. This ensemble approach is more accurate as compared to other classification methods individually. To assess the consistency of a suggested model, which offered a nearly linear ability to estimate all of those instances, the K-fold cross-validation was applied. Many algorithms and combined techniques are available for classification. Integrated learning with other classifications and merging approaches may also be used as the future investigation strategy to analyze the accuracy.

Table 5: Accuracy report via 5-Fold cross validation

| ML Model | CC     | FC     | RC     |
|----------|--------|--------|--------|
| DF1      | 81.23  | 77.64  | 85.16  |
| DF2      | 80.32  | 79.3   | 86.54  |
| DF3      | 83.21  | 78.46  | 87.13  |
| DF4      | 82.48  | 76.15  | 88.78  |
| DF5      | 84.36  | 78.37  | 90.13  |
| Error (%) (Raghunath et al. 2012; Raghunath and Rengarajan 2013) | 2.42 | 5.22 | 1.27 |

Fig. 6: Comparison of CC, FC and RC

Funding No funding.

Declarations

Conflicts of interest The authors assure don’t have any conflict of interest.

Human and animals rights The authors assure no animals and humans involved.

Informed consent Not Applicable to this research.

References

Chan-Yeung M (2003) Xu RH SARS: epidemiology. Respirology 8:S9-14
Chaturvedi A (2020) Top 10 popular smartphone apps to track COVID-19. Available from Geospatial World: https://www.geospatialworld.net/blogs/popular-apps-COVID-19/; 2020, 4 18. accessed on: 24 April, 2020.
Coronavirus (COVID-19) self-assessment tool (2021) Available at: https://www.humandx.
Dhiman G, Kumar VV, Kaur A et al (2021) DON: deep learning and optimization-based framework for detection of novel coronavirus disease using x-ray images. Interdiscip Sci Comput Life Sci. https://doi.org/10.1007/s12539-021-00418-7

e1071—the R package for statistical computing, https://cran.rproject.org/web/packages/e1071/e1071.pdf, 2019.

Gupta VK, Rana PS (2019) Toxicity prediction of small drug molecules of androgen receptor using multilevel ensemble model. J Bioinform Comput Biol 17(5):1–26

Gupta VK, Rana PS (2020) Ensemble technique for toxicity prediction of small drug molecules of the antioxidant response element signalling pathway. Comput J. https://doi.org/10.1093/comjnl/bxa001

Han J, Pei J, Kamber M (2006) Data mining: concepts and techniques, data mining concepts models methods & algorithms, 2nd edn, vol 5, no 4, pp 1–18

Health Engine. Available at: . https://healthengine.com.au/covid19/. accessed on: 26 April 2020.

Henry Ford Health System. Available at: https://www.henryford.com/services/infectious-diseases/conditions/COVID-19-coronavirus/coronavirus-assessment-tool. accessed on: 26 April 2020.

Huang C, Wang Y, Li X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395:497–506

Infermedica. Available at: https://infermedica.com/covid19. accessed on: 26 April 2020.

Karthick Raghunath KM, Rengarajan N (2013) Investigation of faults, errors and failures in wireless sensor network: a systematical survey. Int J Adv Comput Res 3(12)

Karthick Raghunath KM, Vallimayil A, Aarthy MP (2012) Self diagnosing and fault evacuation framework for Ad-Hoc network. Int J Eng Technol 4(6):785–789. https://doi.org/10.7763/ijet.2012.v4.484

Karthikeyan T, Sekaran K, Ranjith D, VinodKumar V, Balajee JM (2019) Personalized content extraction and text classification using effective web scraping techniques. Int J Web Portals (IJWP) 11(2):41–52

Li Q, Guan X, Wu P et al (2020) Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med. https://doi.org/10.1056/NEJMoa2001316

Lin W, Ziming Wu, Lin L, Wen A, Li J (2017) An ensemble random forest algorithm for insurance big data analysis IEEE. School of Computer Science Guangzhou University, China

Middle East Respiratory Syndrome Coronavirus. Available at: https://www.who.int/emergencies/mers-cov/en/. Accessed 16 Feb 2020.

nnet—the r package for neural network, https://cran.rproject.org/web/packages/nnet/nnet.pdf, 2017.

Rothe C, Schunk M, Sothmann P et al (2020) Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. N Engl J Med. https://doi.org/10.1056/NEJMc2001468

Wang C, Horby PW, Hayden FG, Gao GF (2020) A novel Coronavirus outbreak of global health concern. Lancet. https://doi.org/10.1016/S0140-6736(20)30185-9

World Health Organization (2020) Situation reports. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/. Accessed 22 Feb 2020.

Xinhua (2020) China’s CDC detects a large number of new Coronavirus in the South China seafood market in Wuhan. Available at: https://www.xinhuanet.com/2020-01/27/c_1125504355.htm. Accessed 20 Feb 2020.

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