Chapter 2
Machine Learning-Based Scheme to Identify COVID-19 in Human Bodies

Harish Kumar, Anuradha, A. K. Solanki, and Sudeep Tanwar

Abstract A virus spread from China to all around the world named COVID-19 has now become a demon. The fear of death can be easily seen in citizens of around 180 countries and fear to force us indoors. This is a demon of the twenty-first century; typically, this demon does not link with any of the evil, occultism, literature, fiction, mythology, and folklore. COVID-19 is a member of the coronavirus family and caused by the SARS-CoV-2 virus. COVID-19 was first identified in December 2019 at Wuhan, China. SARS virus is responsible for respiratory illness known as COVID-19. We have limited articles on COVID-19 with machine learning (ML) and AI. We do not have any antivirus medicine and other dataset that bring in mind about prediction, detection, and stage identification of COVID-19 in human bodies. Therefore, we decided to bring a machine learning-based technique with a list of datasets that will apply to coronavirus dataset for identification. It is believed that ML and artificial intelligence can help accelerate solutions for predicting the stage of infection. Data analysis presented in this paper helps in minimizing the virus impact with all the other research.

Keywords COVID-19 · AI · SARS · SARS-CoV-2 · CXR · WHO

H. Kumar (*) · Anuradha
Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, India
e-mail: sudeep.tanwar@nirmauni.ac.in

A. K. Solanki
Department of Computer Science and Engineering, Bundelkhand Institute of Engineering and Technology, Jhansi, India

S. Tanwar
Department of Computer Science and Engineering, Nirma University, Gujarat, India
e-mail: sudeep.tanwar@nirmauni.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021
F. Al-Turjman et al. (eds.), Emerging Technologies for Battling Covid-19, Studies in Systems, Decision and Control 324,
https://doi.org/10.1007/978-3-030-60039-6_2
2.1 Introduction

Our earth is infected by a virus that attacks our respiratory system—the nose, throat, and lungs. Commonly we called it flu, but this not the same as normal “flu” viruses that cause minimal damage. So far, there have been more than 112,000 deaths until now worldwide. The behavior of the virus is very harsh and difficult to understand. In 2020 respiratory flu is the most emerging infectious disease, causing high morbidity and mortality in society and worldwide (deadly disease). This spreading virus is an RNA virus and generally found in three types based on their protein. Generally, we termed them as A type, B type, and C type [1, 2]. The WHO publicized that the scientific name of the 2019 SARS virus is COVID-19. COVID-19 belongs to a family of coronaviruses that may cause symptoms such as sore throat, pneumonia, fever, respiratory illness, and lung infection [3–5]. In the past 3 months, the coronavirus disease 2019 (COVID-19) pandemic is spreading at a swiftly. The first genetic material (genome) of COVID-19 was given by Prof. Yong-Zhen Zhang, in January 2020 [6]. On March 13, the WHO declared it as a pandemic for the whole world [3]. Most of the devolved or developing countries are also affected by COVID-19 [3, 4, 7]. Approximately 400 billion people have become prisoners in their houses. Lockdowns are the only solution till date as a preventive measure for spreading this virus [8, 9]. COVID-19 influenza is a pandemic in human society since December 2019. At the time of writing this paper, there have been 10,000,000 confirmed cases of COVID-19 and 500,000 deaths. Using a survey of 10,000 infected people (data available online), we find that risk perception is very high in COVID-19 patients. The human body has no immunity against the coronavirus because its protein structure is similar to human body protein. Influenza pandemics have resulted in damaging human antigen and damaging human tissues of the respiratory system. Coronavirus exists in four different generations.

1. Alphacoronavirus
2. Betacoronavirus
3. Gammacoronavirus
4. Deltacoronavirus

Coronavirus is a type of RNA virus, and its mutation rate is higher than DNA viruses. The genome codes (the set of all the genes of a specific species) for at least four main structural proteins: spike, membrane, envelope, and nucleocapsid proteins [10]. Besides these proteins, the coronavirus has some accessory proteins, which helps the replicative processes and makes it possible to enter into human cells [2]. Figure 2.1 shows the basic structure of COVID-19 virus presented by Dr. Rohan Bir Singh [10] (Table 2.1).

A number of communities are working on finding the solution to COVID-19. The more fertile ground to find the solution to COVID-19 is by using machine learning algorithms [3, 11, 12]. Machine learning creates spot correlations across large amounts of data describing the viruses (Table 2.2).
Fig. 2.1 Basic structure of coronavirus. (Proposed by Dr. Rohan Bir Singh)

Table 2.1 Basic structure of coronavirus and the function of each protein [10]

| Basic structure of coronavirus | Structural protein | Function of protein |
|-------------------------------|-------------------|--------------------|
| Nucleocapsid protein (N) | Bound to RNA genome to make up nucleocapsid |
| Spike protein (S) | This protein is responsible for binding of host cell receptor to facilitate entry to host cell |
| Envelope protein (E) | Interacts with M to form viral envelope |
| Membrane protein (M) | Responsible for determining the shape of viral envelope |
| S. no | Name of virus | Symptoms | No. of country infected (approx.) | Case fatality rate (CFR) | No of persons infected | Total death | Death rate (%) |
|-------|---------------|----------|-----------------------------------|--------------------------|----------------------|-------------|---------------|
| 1     | Marburg       | Hemorrhagic fever | 11 | 50% ≤ CFR ≥ 80% | 466 | 367 | 80 |
| 2     | Nipah         | Acute respiratory infection and fatal encephalitis | 10 | 40% ≤ CFR ≥ 75% | 513 | 398 | 77.60 |
| 3     | H5N1 Bird flu | Fever (often high fever, > 38 °C) and malaise, cough, sore throat, and muscle aches | 18 | 56% | 861 | 455 | 52.8 |
| 4     | Ebola         | Hemorrhagic fever, rare but severe, often fatal illness in humans | 9 | 50% ≤ CFR ≥ 90% | 33,557 | 13,562 | 40.4 |
| 5     | H7N9          | Fever, cough that produces sputum, breathing, problem and wheezing, headache, muscle pain, general malaise | 11 | 56% | 1,632,258 | 284,500 | 17.40 |
| 6     | H1N1 Swine flu | Fever, chills, cough, sore throat, runny nose, stuffy nose watery red eyes, body aches, headache | 100+ | 1% | 8,600,000 | 420,000 | 20.47 |
| 7     | COVID-19      | Fever, dry cough, sore throat, runny nose, and stuffy nose | 176 | Not yet known | 8,600,000 | 420,000 | 20.47 |
2.1.1 Symptoms

The signs of COVID-19 are nonspecific, and the disease identification can range from no symptoms to severe pneumonia and death. As of April 13, 2020, from laboratory confirmed cases, typical signs and symptoms include the following: [3, 4, 13].

(a) Fever
(b) Cough (dry cough)
(c) Tiredness
(d) Sputum production
(e) Shortness of breath
(f) Sore throat
(g) Headache
(h) Myalgia or arthralgia
(i) Chills
(j) Nausea or vomiting
(k) Nasal congestion
(l) Diarrhea
(m) Hemoptysis
(n) Conjunctival congestion

A limited number of articles about coronavirus with artificial intelligence and machine learning are available; few have offered a truly comprehensive view. Our main aim is to design a pattern, datasets, and other aspects of analysis that will be applied to COVID-19 based on symptoms given above. In our research, decisions are according to perceived risk. Various data mining, machine learning, and AI techniques can help us to easily identify the pattern similarity in patients, which can help to accelerate solutions for predicting the stage of infection. The idea behind data mining and ML/AI techniques is to learn the hidden pattern from the available data [7, 8]. We apply various machine learning approaches to identify relationships or associations in biological data to groups with similar genetic structure, to analyze and predict infection stage. We apply machine learning techniques on sequence alignment, structure, function, and clustering structure of symptoms given in 1.1. We apply Duster algorithm which is a data filtering algorithm for refining the data. For smoothing, we use binning method. For reduction, we use parametric numerosity reduction technique and support vector machines with decision tree techniques to identify the impact of latest deadly disease damage.

2.2 Problem Definition

Fighting against COVID-19 confronts the risk of infection. Without infection, there are essentially zero risks of death. Therefore, we decide to beat COVID by prediction rules. Data mining, machine learning, and AI can contribute to the fight against COVID-19. The following are the areas where we can apply advance technology:
(a) Setting the footpath for prediction
(b) Data dashboards
(c) Before time warnings and alerts
(d) Diagnosis and prediction
(e) Treatments and cures
(f) Social control

We assume that being in a healthy state means without infection, the vector $V = (0, X, Y)$ and in an infected state with COVID-19, the vector $V = (1, X, Y)$, where $X$ is the set of individual characteristics, $Y$ is the set of immunity, and $(0, 1)$ represents the infected or not infected status. COVID-19 identifications can be categorized as follows:

(a) Medical Ground: Foretell the impact of new antiviral medicines, structure of proteins, and their impact after interactions with human bodies. This study helps to identify which antiviral medicine helps the patient recover from coronavirus. According to the results of dataset, we modify the medicine to predict protein-ligand interactions [10]. Then after modification, the medicine is used to check with RNA sequence of COVID-19 and chemical composition to forecast which medicine works best.

(b) Self-Quarantine: To predict the infection rate in a community of patients, we apply multiple techniques that help our doctors and society to better plan resourcing and response. Till now we are having various methods for predicting the spreading rate of normal flu [14]. But all these methods are not applicable on COVID-19 due to its changing nature and availability of limited datasets.

(c) Digital Image Results: Medical images like X-ray or CT scan of coronavirus-infected people help in diagnosis. Image filtering and diagnosis help our doctors to identify the infection rate in a patient [15]. The following is the process of percentage calculation in an X-ray and CT scan method. Block diagram is a part of our identification method and is well explained in Fig 2.4.

(d) Machine Learning: The main step is to mine the data to better estimate symptoms and infection rate. The method used in mining the data helps in getting relevant information about the disease. Data received from various sites is very noisy and requires an excellent refining technique. Our main work is based on the fourth part. In an analysis, 54% had positive RT-PCR results, and 86.8% had positive chest CT scans. But in countries like India, we have limited number of laboratories and have a geographical distance between all labs. This is very typical for doctors to predict the COVID-19 patient in his geographical domain. So our aim is design a system which will help and assist society to identify a COVID-19 patient. Our approach worked on the three major tests, depending on patient requirements (Figs. 2.2, 2.3, and 2.4).

1. OPD data collection and first-level identification (first-level quarantine).
2. CT image or CXR report collection and identification (second-level quarantine).
3. Main laboratory report and final prediction (third-level quarantine).
2.3 Methodology

To control spreading of COVID-19, inspecting large numbers of cases infected with the virus for appropriate quarantine and treatment is a priority. Fast and accurate pattern recognition methods are urgently needed to fight against the disease. We decided to bring a machine learning-based technique with a list of datasets that will apply to coronavirus based on available COVID-19 dataset. Our aim is to develop a method that could extract COVID-19’s features in order to provide a tentative diagnosis to our doctors. This will help in predicting the stages of infection of COVID-19.
patients. Our pattern matching can help to accelerate solutions for predicting the stage of infection. To achieve better results, we use recent datasets of 450 patients along with a history of previous illness. We collected approximately 350 X-ray and CT images of confirmed COVID-19 cases. Figure 2.5 represents the deployment model of the approach.

The following are the steps for data preprocessing, from steps 2 to 5.

1. Collection of data
2. Filtering of data
3. Smoothing of data
4. Reduction of data
5. Pattern analysis
Figure 2.6 represents the basic methodology of the detection of COVID-19. A dataset is designed for final analysis of results. Based on the results, an automated system automatically detects whether the person is infected or not and if there is any chance of recovery or not with a minimum cost and accurate results.

2.4 Dataset and Tentative Solution

Coronavirus disease 2019 (COVID-19) is a communicable disease. The virus can easily spread from one person to another. A 57-year-old Chinese woman named Wei Guixian is the patient zero of COVID-19. The disease causes respiratory illness with symptoms like cough, fever, and difficulty in breathing. A person can protect himself/herself by washing his/her hands frequently, avoiding touching the face, and avoiding close contact with infected people. On December 31, 2019, the WHO was alerted by the Chinese government of a series of corona-like cases in the city of Wuhan. The disease causes respiratory illness with symptoms like cough, fever, and difficulty in breathing. We collect all the data from various sites [3, 4, 16]. We feel that every problem seems to have a ML solution. There are various areas where ML/AI can contribute to the fight against COVID-19.

(a) OPD data collection and first-level identification: Testing is also an essential process for a suitable response to the pandemic illness [8]. Testing helps us to understand the spread processing and to take evidence-based measures to slow down the spread of the disease. OPD diagnosis allows COVID-19-infected people to know that they are infected or not. Its solution may be moved with manual
processes; every doctor must have to fill all the following data of each patient. This is a first-stage diagnosis process or clinical features identification [17] (Table 2.3).

On the basis of above diagnosis table, doctors cannot predict if a patient is infected with COVID-19 or not, and people do not know that they suffering from COVID-19 or not. If no, one might stay at home, and others might go for quarantine stage for 14 days. This approach is applicable where a number of cases are limited and in controlled population states. But for countries like India where the population is around 135 billion, this is not possible if a disease becomes a pandemic. For finding, I strongly feel that ML/AI with data mining technique is the only solution for identification of patients (Table 2.4).

The decision tree algorithm for predictive modeling can be used to explicitly represent decisions. Its graphical representation makes the use of branching techniques to exemplify all possible outcomes. For identifying whether the person is infected with COVID-19 or not, we apply DT algorithm of machine learning technique (Figs. 2.7 and 2.8).

The supervised learning (decision tree) technique can be used for prediction [11, 12, 18, 19]. This helps in solving classification and regression issues. We complete this into two steps, as follows:

1. Classification (used to find results from a set of possible values)
2. Regression (used to find the result data where results are continuous values)

Constructing optimal binary decision trees is NP-complete. Decision trees are created through a process called splitting, and this process is also known as induction. We use recursive divide and conquer strategy. To achieve better results, we apply greedy algorithm steps, as follows.

### Table 2.3 First-level clinical data of COVID-19 patients

|                        | Mild (N = 42) | Severe (N = 9) |
|------------------------|--------------|----------------|
| **Sex**                |              |                |
| Male                   | 51           | 56             |
| Female                 | 48           | 44             |
| **Travel history**     | Yes/no       | Yes/no         |
| **Age**                | Mean (SD)    |                |
| 43.70 ± 11.71          | 52.83 ± 12.96|
| Range                  | 15–78        | 31–79          |
| **Cough**              | No           | Yes            |
| 32                     | 17           |
| Yes                    | 68           | 83             |
| **Fever**              | No           | Yes            |
| 16                     | 17           |
| Yes                    | 84           | 83             |
| **Temperature**        | Male         |                |
| 98.4                   | 98.6         |
| Female                 | 100.4        | 100.4          |
| **Previous case history** | No        | Yes            |
| 85                     | 83           |
| Yes                    | 15           | 17             |
1. Trace the root node; we create a crack (splitting) for each attribute.
2. Calculate the cost of the split at each level.
3. Choose the split that costs the least.
4. Recur the process into the sub-trees and repeat step 1 until all leaf nodes are covered.

Information gain is the main factor that is used to construct a decision tree [12]. For calculating the splitting cost, we apply cost function and information gain function:

\[
I(P,n) = \frac{P}{S} \log_2 \frac{P}{S} - \frac{n}{S} \log_2 \frac{n}{S}
\]  

(2.1)

where \(S = P + n\).
The two impurity measures or splitting criteria that are commonly used in decision trees are Gini impurity and entropy:

\[
E(A) = \sum_{i=1}^{T} \frac{P_i + n_i}{P + n} * I(P_i + n_i)
\]

(2.2)

\[
\text{Gain}(A) = I(P,n) - E(A)
\]

(2.3)

\[
\log_2(x) = \frac{\log_{10}(x)}{\log_{10} 2}
\]

(2.4)

The above equations help us to calculate the total gain and entropy from the COVID-19 datasets and help to design and exact prediction of COVID-19 patient. According to analysis, only three possibilities are there, named COVID-19, normal cold, and flu (Table 2.5 and Fig. 2.9).

(b) CT image or CXR report collection and identification: Fig. 2.10 shows two images, one representing the healthy lungs and the other one representing infected images. We are applying a deep learning-based technique for detecting COVID-19 on CT scan and lung X-rays using MATLAB.

All the images were compared by the technique given in the flowchart and which will be well written in MATLAB (Fig. 2.11).

Codes for change in image or image comparison of healthy and infected lungs are well written in Python/MATLAB and simulate the results in WEKA, offer a list of suggestions to the doctors, and assist them in the diagnosis process, predicting results of infection (Fig. 2.12).
All the images are compared by the following code, which is well written in MATLAB. This will help us to classify the two different images and help in identifying the difference between the two images. For differentiation, we use X-ray images of a coronavirus-infected person and found that every day the infection is increasing with a growth rate of 4.75%. We are dealing with a limited dataset, and to achieve better and effective result, we use the same images for different comparison algorithm. Every CXR image of the lungs has some difficulties in the exact calculation of results, and these problems may be summarized as ABCDEF [15, 20].

| S.no | Disease     | Sex | Age | TH | Cough | Fever (°F) | Previous history | Remark 1 | Remark 2     |
|------|-------------|-----|-----|----|-------|------------|-----------------|----------|--------------|
| 1    | COVID-19    | M   | 0.6 | N  | Y     | 97.9       | N               | 1        | Infected     |
| 2    | Influenza   | M   | 1   | N  | Y     | 99.9       | N               | 0        | Flu          |
| 3    | Influenza   | M   | 5   | N  | Y     | 99.9       | N               | 0        | Flu          |
| 4    | Influenza   | M   | 10  | N  | Y     | 98.9       | N               | 1        | Flu          |
| 5    | COVID-19    | F   | 15  | Y  | Y     | 100.9      | N               | 1        | Infected     |
| 6    | Cold        | F   | 20  | Y  | Y     | 100.5      | N               | 0        | Normal cold  |
| 7    | COVID-19    | M   | 25  | Y  | Y     | 101        | N               | 1        | Infected     |
| 8    | COVID-19    | M   | 27  | Y  | Y     | 102        | N               | 1        | Infected     |
| 9    | COVID-19    | F   | 30  | N  | Y     | 100        | Y               | 1        | Infected     |
| 10   | COVID-19    | F   | 35  | N  | Y     | 101.7      | N               | 1        | Infected     |
| 11   | Cold        | M   | 40  | Y  | Y     | 99.9       | Y               | 0        | Normal cold  |
| 12   | COVID-19    | M   | 45  | N  | Y     | 99.9       | N               | 1        | Infected     |
| 13   | Cold        | F   | 50  | Y  | Y     | 102        | N               | 0        | Normal cold  |
| 14   | COVID-19    | F   | 55  | Y  | Y     | 101        | N               | 1        | Infected     |
| 15   | COVID-19    | F   | 60  | N  | Y     | 101        | Y               | 1        | Infected     |
| 16   | Cold        | M   | 65  | N  | Y     | 101        | Y               | 0        | Normal cold  |
| 17   | Cold        | M   | 7   | N  | Y     | 97.9       | Y               | 0        | Normal cold  |
| 18   | COVID-19    | M   | 75  | N  | Y     | 97.9       | Y               | 1        | Infected     |
| 19   | Cold        | F   | 80  | N  | Y     | 99.9       | Y               | 0        | Normal cold  |
| 20   | Influenza   | F   | 85  | N  | Y     | 101.9      | N               | 0        | Flu          |

Fig. 2.9 Decision tree based on attribute selection
• Airways
• Breast shadows (females)
• Bones
• Cardiac outline
• Costophrenic angles
• Diaphragm
• Edges
• Extrathoracic tissues
• Fields (lung parenchyma)
• Failure

Due to all these difficulties, we cannot rely only on CXR images [21]. We need some more accurate and authentic results including CXR images, which will help us in accurate prediction of COVID-19 stage. However, this is not the only problem in testing COVID-19. In every city, basic scanning facilities like CT scan machine and chest X-ray machine are not available especially in rural areas. To save the citizens of countries, governments are performing COVID-19 tests on people.

(c) Main laboratory report and final identification for quarantine: CXR result analysis has some limitations which were described earlier. Testing is also an essential process for a suitable response to the COVID-19 illness [8]. OPD diagnosis allows COVID-19-infected people to know that they are infected or not (first level). The changes in CT scan and lung X-ray shows the second level of diagnosis. People do not know that they suffering from COVID-19 or not. If no, one might stay at home, and others might go for quarantine stage for 14 days. For more accurate results, we require few more laboratory test results of patients infected with COVID-19 on final admission to hospital. However, this is an expensive and time-consuming process (Tables 2.6, 2.7, and 2.8).

Fig. 2.10 Clinical image of healthy lungs vs infected lungs
Fig. 2.11 Flowchart for comparison of two images
Results

On the basis of above data, we design decision tree for the same. The max_depth of the COVID-19 decision tree is 3. If max_depth = 0, split the nodes. Maximum height indicates over-fitting, and the minimum value indicates under-fitting (Tables 2.9 and 2.10).

The algorithm for calculating accuracy is as follows:

![Fig. 2.12 Comparisons of day-wise CXR reports](image)

**Table 2.6** First-level clinical data

|                  | Mild/no ICU (N = 42) | Severe/ICU (N = 9) |
|------------------|-----------------------|--------------------|
| First-level data (from Table 2.3) |                       |                    |
| COVID-19 RT-PCR test |                       |                    |
| SARS-CoV-2 N1    | ±                     | ±                  |
| SARS-CoV-2 N2    | ±                     | ±                  |
| SARS-CoV-2 N3    | ±                     | ±                  |

2.5 Results

On the basis of above data, we design decision tree for the same. The max_depth of the COVID-19 decision tree is 3. If max_depth = 0, split the nodes. Maximum height indicates over-fitting, and the minimum value indicates under-fitting (Tables 2.9 and 2.10).

The algorithm for calculating accuracy is as follows:
# Train Decision Tree Classifier

Step 1: `clf = clf.fit(X_train, y_train)`

# Predict the response for test dataset

Step 2: `y_pred = clf.predict(X_test)`

We have collected medical records relevant to COVID-19 of approximately 5000 patients (source of data is the Internet), and we try to predict its stage.

*is the count (150 to 400).
# Model Accuracy

Step 3: print "ACCURACY:\n\nCorrectly classified instances 20 100%
Incorrectly classified instances 0 0
Kappa statistic 1
Mean absolute error 0

Table 2.8 Diagnosis result of COVID-19 patients

| S. no | Factor          | R1     | R2     | R3     | R4     | R5     |
|-------|-----------------|--------|--------|--------|--------|--------|
| 1     | DCXR            | 1.67%  | 1.47%  | 1.42%  | 1.22%  | 1.21%  |
| 2     | Cough           | +      | +      | ±      | ±      | ±      |
| 3     | Fever           | +      | +      | ±      | ±      | ±      |
| 4     | Chest pain      | +      | +      | −      | −      | −      |
| 5     | Glucose PP      | +      | +      |        |        |        |
| 6     | Glucose fasting | +      | +      |        |        |        |
| 7     | Qualitative PCR | +      | +      |        |        |        |
| 8     | SARS-CoV-2 N1   | +      | +      | −      | −      | −      |
| 9     | SARS-CoV-2 N2   | +      | +      | −      | −      | −      |
| 10    | SARS-CoV-2 N3   | +      | ±      | +      | −      | −      |
| 11    | RNase P         | ±      | ±      | ±      | +      | −      |

| COVID-19 infected | COVID-19 infected | Presumptive positive | SARS-CoV-2 not detected | Invalid results |

DCXR Difference between two consecutive CXR
+ Infected above the range
− Absence
± Value increments or decrement based on recovery

Table 2.9 Detailed accuracy by class of Table 2.7

| S. no | TP rate | FP rate | Precision | Recall | F-measure | MCC | ROC area | PRC area | Class      |
|-------|---------|---------|-----------|--------|-----------|-----|----------|----------|------------|
| 1     | 1.0     | 0.0     | 1.00      | 1.00   | 1.00      | 1.00| 1.00     | 1.00     | Infected   |
| 2     | 1.0     | 0.0     | 1.00      | 1.00   | 1.00      | 1.00| 1.00     | 1.00     | Flu        |
| 3     | 1.0     | 0.0     | 1.00      | 1.00   | 1.00      | 1.00| 1.00     | 1.00     | Normal cold|
|       | 1.0     | 0.0     | 1.00      | 1.00   | 1.00      | 1.00| 1.00     | 1.00     |            |
Table 2.10  Confusion matrix of Table 2.5

|     | A    | B    | C    | Classified as |
|-----|------|------|------|---------------|
| 10  | 0    | 0    |      | Infected      |
| 0   | 4    | 0    |      | Flu           |
| 0   | 0    | 6    |      | Normal cold   |

Accuracy level is 77.53%, which is better for achieving approximation results (Figs. 2.13 and 2.14).

After 100 folds, we found the following difficulties in decision tree method:

1. COVID-19 dataset is changing every minute, and all the predictions are failing, so a small change in data causes a change in complete decision tree.
2. This method involves higher time to train the model.
3. This algorithm is quite complex and expensive, and decision trees are NP-complete (Fig. 2.15).

The estimation of COVID-19 patient can be identified by dividing the death rate by the number of infected patients [22, 23]. The mortality rate is an “estimate of the
portion of a population that dies during a specified period”; case fatality rate is calculated by our approach by the following formula:

\[
CFR(\%) = \frac{\text{No of Death from Disease}}{\text{No of Diagnosis cases of Disease}} \times 100 \quad (2.5)
\]

CFR is sometimes different from original data, but crude mortality rate is also called the crude death rate and is accurate, as shown in Fig. 2.16.

Case fatality rate is sometimes called infection fatality rate [22] and calculated as:

\[
IFR(\%) = \frac{\text{No of Death from Disease}}{\text{No of Diagnosis cases of Disease}} \times 100 \quad (2.6)
\]

2.6 Conclusion

In this paper, we applied data mining and machine learning techniques to identify the symptoms of COVID-19. Accurate diagnosis of COVID-19 origin can be done significantly for improving influenza surveillance and treatment. Our approach is simple, accurate, and cost-effective. Our paper is based on the analysis of randomly selected data samples available on the Internet and CXR report of patients. Decision tree method helps us in diagnosing the symptoms and helps in the exact calculation of CFR and IFR rate for COVID-19.
Fig. 2.15  Simple $K$-means clustering

Fig. 2.16  Crude death rate globally (last modified on April 7)
References

1. Attaluri, P.K.: Applying Machine Learning Techniques to Classify H1N1 Viral Strains Occurring in 2009 Flu Pandemic (2009)
2. Schoeman, D., Fielding, B.C.: Coronavirus envelope protein: current knowledge. Virol. J. 16, 69 (2019)
3. WHO. Novel Coronavirus—China. https://www.who.int/docs/defaultsource/coronavirus/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=f6b4d49b1_2. Accessed 11 Feb 2020
4. https://www.who.int/docs/default-source/coronavirus/who-china-joint-mission-on-covid-19-final-report.pdf (Accessed 1 Feb 2020)
5. WHO. Novel coronavirus—China (2020). https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/ (Accessed 1 Feb 2020)
6. Virological.org. Novel 2019 coronavirus genome (2020). http://virological.org/t/novel-2019-coronavirus-genome319 (Accessed 1 Feb 2020)
7. Naudé, W.: Artificial intelligence against COVID-19: an early review. https://towardsdatascience.com (2020)
8. Bhopalwala, B.: A potential machine learning approach that can help stop COVID-19. https://towardsdatascience.com (2020)
9. Kumar, H., Solanki, A. K., & Singh, K. K. (2020). Progressive Machine Learning Approach with WebAstro for Web Usage Mining. Procedia Computer Science, 167, 1400–1410
10. https://www.ncbi.nlm.nih.gov/books/NBK554776/figure/article-52171.image.f3/?report=objectonly (Accessed 1 Feb 2020)
11. Agrawal, G.L., Gupta, H.: Optimization of C4. 5 decision tree algorithm for data mining application. Int. J. Emerg. Technol. Adv. Eng. 3(3), 341–345 (2013)
12. Roobaert, D., Karakoulas, G., & Chawla, N. V. (2006). Information gain, correlation and support vector machines. In Feature extraction (pp. 463–470). Springer, Berlin, Heidelberg.
13. LabCorp COVID-19 RT-PCR test EUA summary. https://www.fda.gov/media/136151/download (Accessed 1 Feb 2020)
14. Haque, M., Sartelli, M., McKimm, J., & Bakar, M. A. (2018). Health care-associated infections—an overview. Infection and drug resistance, 11, 2321. https://doi.org/10.2147/IDR.S177247. Published online 2018 Nov 15. PMCID: PMC6245375
15. https://radiopaedia.org/articles/chest-x-ray-ABCDE-summary (Accessed 1 Feb 2020)
16. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Niu, P. (2020). A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine. https://doi.org/10.1056/NEJMoa2001017
17. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cheng, Z. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5
18. Maingi, N. N., Lukandu, I. A., & Mwau, M. Comparative Analysis of the C4. 5 and ID3 Decision Tree Algorithms for Disease Symptom Classification and Diagnosis. Int. Res. J. Adv. Eng. Sci. 4(2), 188–194 (2019) ISSN (Online): 2455–9024
19. Shaukat, K., Masood, N., Mehreen, S., Azmeen, U.: Dengue fever prediction: a data mining problem. J. Data Mining Genomics Proteomics. 6, 181 (2015). https://doi.org/10.4172/2153-0602.1000181
20. Mettler, F.A., Huda, W., Yoshizumi, T.T., Mahesh, M.: Effective doses in radiology and diagnostic nuclear medicine: a catalog. Radiology. 248, 254–263 (2008)
21. Yang, R., Li, X., Liu, H., Zhen, Y., Zhang, X., Xiong, Q., ... & Zeng, W. (2020). Chest CT severity score: an imaging tool for assessing severe COVID-19. Radiology: Cardiothoracic Imaging, 2(2), e200047.
22. Spychalski, P., Blażyńska-Spychalska, A., & Kobiela, J. (2020). Estimating case fatality rates of COVID-19. The Lancet. Infectious Diseases. 20(7), 774–775 (2020). https://doi.org/10.1016/S1473-3099(20)30246-2
23. Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., & Favre, G. (2020). Real estimates of mortality following COVID-19 infection. The Lancet infectious diseases. Lancet Infect. Dis. 20(7), 773 (2020). https://doi.org/10.1016/S1473-3099(20)30195-X