Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

A numerical approach to maximize the number of testing of COVID-19 using conditional cluster sampling method

Naurin Zoha [a], Sourav Kumar Ghosh [b,*], Mohammad Arif-Ul-Islam [c], Tusher Ghosh [d]

[a] *Department of Industrial & Production Engineering, Bangladesh University of Engineering & Technology (BUET), Dhaka, 1000, Bangladesh*
[b] *Department of Industrial & Production Engineering, Bangladesh University of Textiles (BUTEX), Tejgaon, Dhaka, 1208, Bangladesh*
[c] *Department of Applied Chemistry & Chemical Engineering, Noakhali Science & Technology University (NSTU), Noakhali, 3802, Bangladesh*
[d] *Department of Marketing, University of Rajshahi, Rajshahi, 6205, Bangladesh*

ARTICLE INFO

Keywords:
Random forest classifier
Conditional cluster sampling
Monte Carlo simulation
COVID-19
Pool testing

ABSTRACT

The COVID-19 pandemic is the defining health crisis of the world in 2020 and the world economy is affected as well. Bangladesh is also one of the impacted countries, which needs to conduct sufficient tests to identify patients and accordingly adopt measures to limit the massive outbreak of this viral infection. But due to economic drawbacks and also unavailability of testing equipment, Bangladesh is lagging critically behind in test numbers. This study shows a pool testing method named Conditional Cluster Sampling (CCS) that utilizes soft computing and data analysis techniques to reduce the expense of total testing equipment. The proposed method also demonstrates its effectiveness compared to the traditional individual testing method. Firstly, according to patients' symptoms and severity of their conditions, they are classified into four classes- Minor, Moderate, Major, Critical. After that Random Forest Classifier (RFC) is used to predict the class. Then random sampling is done from each class according to CCS. Finally, using Monte Carlo Simulation (MCS) for 100 cycles, the effectiveness of CCS is demonstrated for different probability levels of infection. It is shown that the CCS method can save up to 22% of the test kits that can save a huge amount of money as well as testing time.

1. Introduction

The worldwide COVID-19 pandemic is now difficulty to humanity which is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Bangladesh is one of the many countries to be affected by COVID-19. The first positive patient was identified in Bangladesh on March 7, 2020. The transmission was minimal through the month of March, but exhibited a steep increase since April 2020 and up to June 24, 2020, Bangladesh is placed at 17th position considering the total number of infected patients [1]. The exponential increase in the number of sample testing could help beat the transmission, but as of June 24, 2020, Bangladesh stands in the 147th position in the number of tests per million populations in the world which is the bottom-most position in the South Asia region. Not only in Bangladesh, even in the most developed countries, reverse transcription-polymerase chain reaction (RT-PCR) testing, which involves swab testing for the virus' genetic material and is currently the standard test, is severely constrained. This is due to shortages in key supplies, such as reagents, and a limit to the number of tests that can be performed per day using existing equipment [2].

Germany and India have already adopted pool testing methods to enhance their number of tests with the expense of a limited number of test kits [3]. This study shows a Conditional Cluster Sampling (CCS) method where patients are tested in pools, instead of individual testing using a numerical method that implements both machine learning and statistical data and upon obtaining results of each pool, a decision is taken whether to continue the testing iteration or to terminate it.

The prime motives of this study are to:

1. Classify the total number of patients on the basis of severity of their conditions
2. Apply CCS to decrease the expense of test kits.

The entire paper is organized as follows- Section 2 describes the related recent literary works. In section 3, the basic methodologies along with the soft computing methods are discussed. At first, data collected

[*] Corresponding author.
E-mail addresses: naurin.zoha@gmail.com (N. Zoha), sourav@butex.edu.bd (S.K. Ghosh), arif.rahad@gmail.com (M. Arif-Ul-Islam), tusherghosh46@gmail.com (T. Ghosh).

https://doi.org/10.1016/j.imu.2021.100532
Received 4 August 2020; Received in revised form 18 January 2021; Accepted 5 February 2021
Available online 17 February 2021

from a population of potential patients for their symptom details, according to those details they are classified into four classes- Minor, Moderate, Major, and Critical taking the opinions of specialist frontline doctors. Then using a machine learning algorithm, Random Forest Classifier (RFC), each patient is classified into one of these classes and after implementing the CCS method, the patients are tested. Afterward, using Monte Carlo Simulation (MCS) techniques for different probability levels of infection, the efficacy of the method is demonstrated. Section 4 covers the results and discussions of the methodologies, including data description as well as the comparative analysis of the traditional method and proposed method. Finally, in section 5, conclusions with limitations and assumptions on the paper and its future directions are provided.

2. Literature review

Underdeveloped countries in Africa encounter greater limitations to testing resources, leaving them ill-equipped to react to the pandemic [4]. Rapid detection tests (RDT) using kits based on antibody detections are less reliable than the PCR-based tests. So, the rapid microfluidic RT-PCR method can be replaced with that to ensure accuracy which is a very sensitive issue regarding this virus spread [5].

2.1. *Symptoms of COVID-19*

Fever, high body temperature, cough, fatigue, headache, hemoptysis, diarrhea, and dyspnoea are the major symptoms of COVID-19 since the early days of the outbreak in China [6]. Fever, cough, shortness of breath, myalgia, haemoptysis, sputum production, sore throat, rhinorrhoea, chest pain, and diarrhea are found to be the major symptoms of this disease in the literature review [7].

2.2. *Predicting the COVID-19 outbreak*

A modified stacked autoencoder for modeling the transmission dynamics was proposed to predict the confirmed cases of COVID-19 in China [8]. The robust Weibull model based on iterative weighing was used to predict the number of active cases of COVID-19 in countries worldwide [9]. The COVID-19 outbreak was predicted by different mathematical evolutionary algorithms and two distinct Machine Learning (ML) techniques. Among ML techniques, artificial neural network (ANN) outperformed adaptive neuro-fuzzy inference system (ANFIS) [10]. 9 different ML algorithms were employed to estimate the new cases of COVID-19 outbreak in 10 densely populated countries worldwide to find the best-fitted model for each country [11]. The autoregressive integrated moving average (ARIMA) and least square support vector machine (LS-SVM) models were employed to predict the confirmed cases of COVID-19 in the five countries of the world. Both models showed good results. However, the accuracy of the LS-SVM model is better than the ARIMA model [12]. Support vector regression model was proposed to forecast the death and active cases of COVID-19 in India for the period of 1st March to April 30, 2020 [13]. Country-based prediction models for the COVID-19 pandemic are proposed and fathomed by multi-gene genetic programming (MGGP) [14]. A deterministic mathematical model based on susceptible, infectious, exposed, and recovered (SEIR) persons is developed to predict the COVID-19 outbreak. This model considers the effect of lockdown to estimate the number of affected people in Saudi Arabia [15].

2.3. *Predicting the COVID-19 patient condition*

A review on group testing is discussed, and it is found that group testing can reduce the constraints in the available testing methods for SARS-CoV-2 [2]. The probability to be a positive victim of COVID-19 was predicted based on the neural network. The cluster sampling was consisted based on that prediction and it is posited that 73% of tests can be mitigated [16]. It is proposed that 30 samples per pool can ameliorate test capacity with existing test kits and identify positive samples with sufficient adequate diagnostic accuracy [17]. A single positive sample can be determined in pools of up to 32 samples, with 90% accuracy. With certain cycle amplification, the sampling size may be increased up to 64 samples with a minimum error rate [18]. The optimum pool size was calculated based on the prevalence conditions of positive tests. If the pool is positive, all samples will be tested individually while for negative tests, the pool was unaffected [19]. It is found that the pool testing method depends on the infection rate. If the infection rate is high, the pool size will be small. It is proposed that for 30.78% positive tests, the optimal pool size should be 3. On the contrary, the pool size is considered to be 25 for a 0.18% infection rate [20]. ANN is used to predict the condition of recovered and death cases of COVID-19 patients in South Korea based on seven major variables such as country, infection reason, sex, group, confirmation date, birth year, and region. It is discerned that infection reason and region are the most significant variables for predicting the status of recovered and dead victims, respectively [21]. A detailed review on ML and deep learning models for the classification of coronavirus images such as x-ray and CT scans are represented and it is posited that convolutional neural network (CNN) could be a useful technique to identify early-stage detection, distinguishing, and extraction of essential features automatically [22]. A random forest algorithm was used to predict the condition of affected victims of the COVID-19 based on geographical, travel, health, and demographic data [23].

3. Methodology

This research focuses on conditional cluster sampling (CCS) for COVID-19 patients based on the health condition of patients. Basically, this work is divided into four major steps as follows:

1. Collecting the patient database.
2. Applying Random Forest Classifier (RFC) to classify each patient's condition.
3. Implementing the CCS method based on the condition of the patient.
4. Applying Monte Carlo Simulation (MCS) at different levels of probability for several cycles to estimate the total number of tests.

3.1. *Step-1: collecting patient database*

The data was collected for patients across different age groups. In

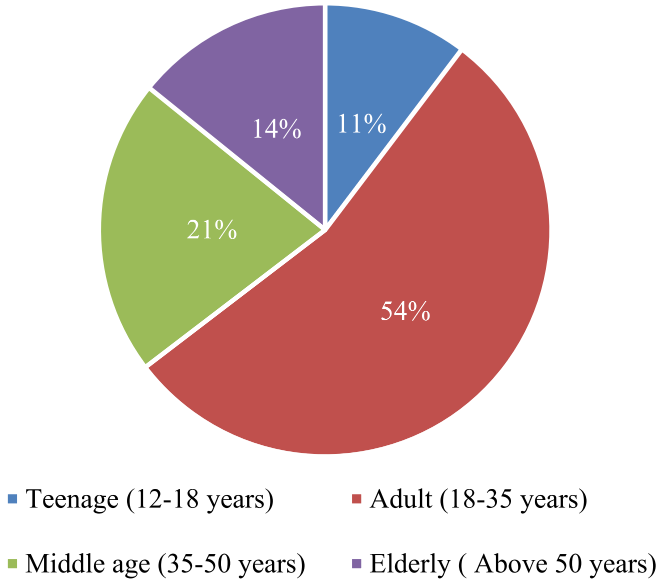

Fig. 1. Patients' age group.

Fig. 1 the different age groups are shown, where the majority of the patients are adults, and the least number of patients are from the teenage category.

In Fig. 2, the gender divisions of the patients are shown. Its seen that most percentages of the patients were males, and the least percentage were transgender.

In Fig. 3 the locations of different patients are displayed. All the patients were from Dhaka city, only from different zones. Depending on COVID-19 spread, the entire Dhaka city is divided into 3 zones- Red, Yellow, and Green [24]. The figure shows the percentage of patients belonging to each of these regions.

In Fig. 4, the numbers of patients are demonstrated against each of the listed symptoms under consideration. The most common symptoms among patients are found to be fever and headache.

Symptoms of individual patients are collected over the survey. This information is sent to frontline doctors directly involved in the treatment of COVID-19 patients to analyze their conditions and depending on the doctors' report, the database is completed.

3.2. Step 2: applying Random Forest Classifier (RFC) to classify each patient's condition

Random Forest (RF) is a supervised machine learning algorithm that is mainly used for classification applications also used for both classification and prediction; however, it is mainly applied for classification applications. Forest means trees and the more trees the more robust the forest is. In the random forest classification method, this model creates different decision trees based on data samples and when new data points are inserted for its class prediction, each decision tree gives one prediction, and finally, the best solution is selected by voting. For an input vector *(x)*, each decision tree will give a vote. Then, $C_{rf}^{B} = \mathit{majorityvote}\{C_b(x)\}_1^B$ where $C_b(x)$ is the prediction of class on b^{th} random-forest tree and C_{rf}^{B} is the final prediction using the majority vote [25]. The main concept behind this model is simple but a powerful one. The reason for this wonderful effect is that the models protect each other from their errors. The choice of attribute selection and pruning methods are necessary for the design of decision trees. There are many attribute selection methods but the most frequently used attribute selection measures in decision tree induction are the gain ration criterion [26] and the Gini Index [27]. RFC uses the Gini Index method for its attributes' selection which measures the impurity of an attribute with respect to its classes. For a given training set P, selecting a sample case randomly and to predict its class as C_i, the Gini index can be written as-

$$\sum\sum_{j\neq i}(f(C_i,\ T)\ /\ |T|)(f(C_j,\ T)\ /\ |T|)$$

Here, $(f(C_i,\ T)\ /|T|)$ is the probability that a selected case belongs to the class C_i.

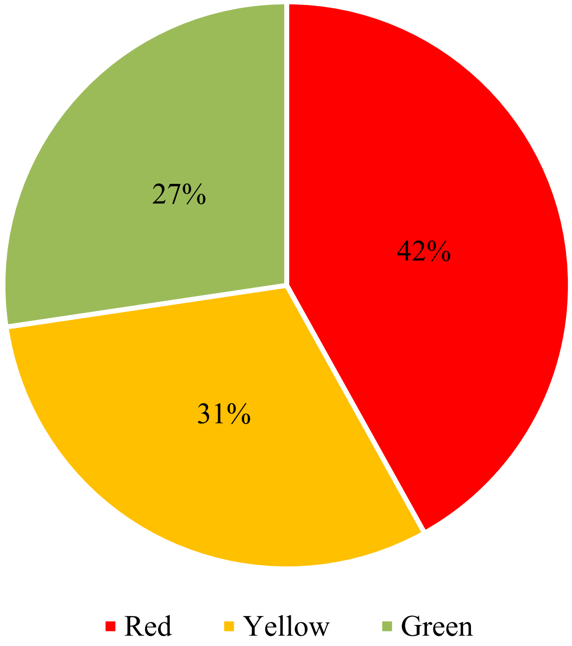

Fig. 3. Patients' location.

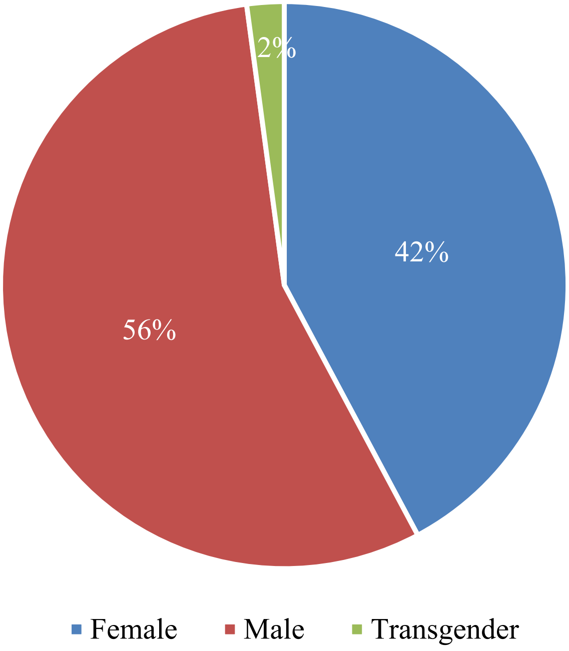

Fig. 2. Patients' gender divisions.

For generating a prediction model through RFC, basically, two parameters are rudimentarily required-the number of classification trees and the predicting variables that reside in each node to spread out the trees. The selected features are expanded for each node and this way, N decision trees are grown where N is a user-defined value about the number of trees to be grown. When new data points are introduced, these are passed down to all those trees and then it chooses its class by maximum votes out of N votes.

For this research, input data with various features and an output attribute with different levels are split into two datasets: training dataset and testing dataset. Then bootstrap aggregating and attribute bagging are developed to form a randomly selected decision tree by minimizing the misclassification rate. Finally, the testing dataset is examined to predict the class. 90% of data is used as training data [28] while the rest of the data is assigned as testing data to classify the patient's condition.

The patients' conditions are basically divided into four prime classes- Minor, Moderate, Major, and Critical that are named class-1, class-2, class-3, and class-4, respectively. Using the training dataset, the RFC module is trained and then after applying RFC, the classification of the test dataset is done based on the symptoms of the patients. Fig. 5 shows the yielded patient categorizations.

3.3. Step 3: implementing the CCS method based on the condition of the patient

Conditional Cluster Sampling is a technique to stratify the cluster sample based on the condition of the patient. For the better accuracy of the test, the maximum cluster size chosen is 64 [18]. The sample size is inversely proportional to the severity that means more severe cases are clustered into small sample sizes. The main reason behind this is that the

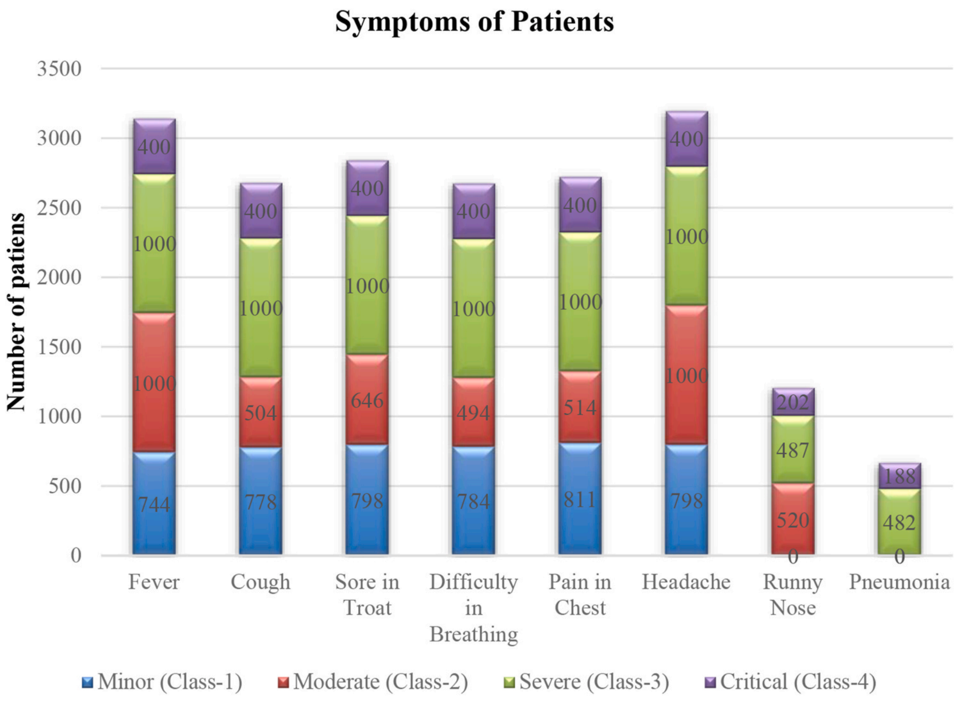

Fig. 4. Patients' symptoms.

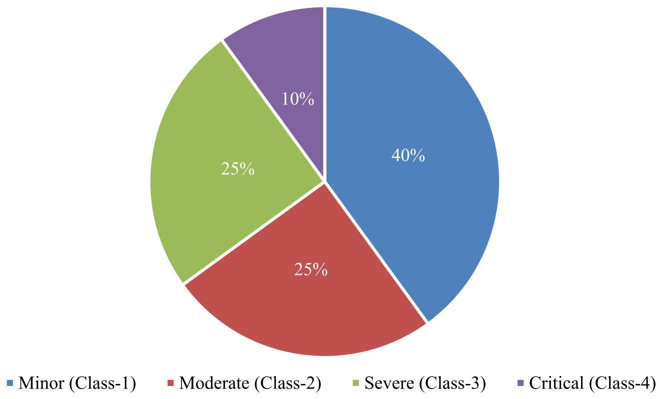

Fig. 5. Patient type.

probability of a critical patient to be positive for COVID-19 is higher than other patients. On the contrary, the possibilities of a minor patient to be affected are less than other patients. A larger cluster size can truncate the testing cost and time. The clusters are classified into 4 sizes such as 64, 32, 16, and 8. Sample sizes of the cluster are chosen for ease of division. For instance, 64 can be divided into two 32 groups, then 32 can also be divided into two 16 groups, and so on. The sample size of 16 was assigned for the most critical patients while the rest of the patients are classified into 64. Afterward, based on the severity of the patients' health condition and their designated class, the CCS testing method is implemented on the test data.

3.4. *Step 4: applying Monte Carlo Simulation (MCS) at different levels of probability for several cycles to estimate the total number of tests*

In Monte Carlo simulation, random samples from each statistical distribution are used as the values of the input variables. For each set of input parameters, we get a set of output parameters. We collect such output values from a number of simulation runs for different classes. Then variations in output are analyzed to make decisions about the final simulation results [29]. Finally, MCS was implemented to estimate the total number of tests for separate classes, and results were compared with the traditional individual testing method. This provided an estimate of the total percentage of test kits saved with CCS.

3.5. *Process flowchart of the research work*

Fig. 6 depicts the process flowchart of conditional cluster sampling (CCS). At first, the feedback from the designated patient group is collected through different media such as the Google response form, over the phone, and via Email. Then the symptoms are identified, and the doctor's opinion about the patient's condition is recorded, to create a database containing the symptom levels and patient condition. The data is used to train the RFC model and the tested data is employed to classify the patient's condition based on a variety of symptom levels. There are four levels of patients based on severity such as class- 1, 2, 3 & 4. Here, the severity level is taken to be increasing with higher classes; for instance, the class-4 patients are considered more severe than any other class of patients. Accordingly, clustering is done among different classes. For example, if the patient is from class-2, the cluster size is 64. After testing the cluster sample, if the test result is negative, all the patients are free from COVID-19, and further testing is not needed. If the result is otherwise, it is divided into two clusters. The sample size is then 32 and these two clusters are formed and tested again. If the result is positive for both clusters, again each cluster is divided into two sub-clusters and the cycle continues until the results come negative or the sample size reduces down to 8. If the test results are positive again when the sample size is 8, then all the samples are examined individually. Thus, this condition-based cluster sampling method continues. The effectiveness of this CCS is evaluated using MCS at different probability levels. The probability level refers to the infected rate of various classes of patients. The probability to generate random data is assumed based on historical data. Thus, the total number of required tests using CCS is evaluated. Here the MCS is iterated for 100 cycles for more accuracy. Finally, results are compared to the traditional testing method to check the efficacy of the proposed method.

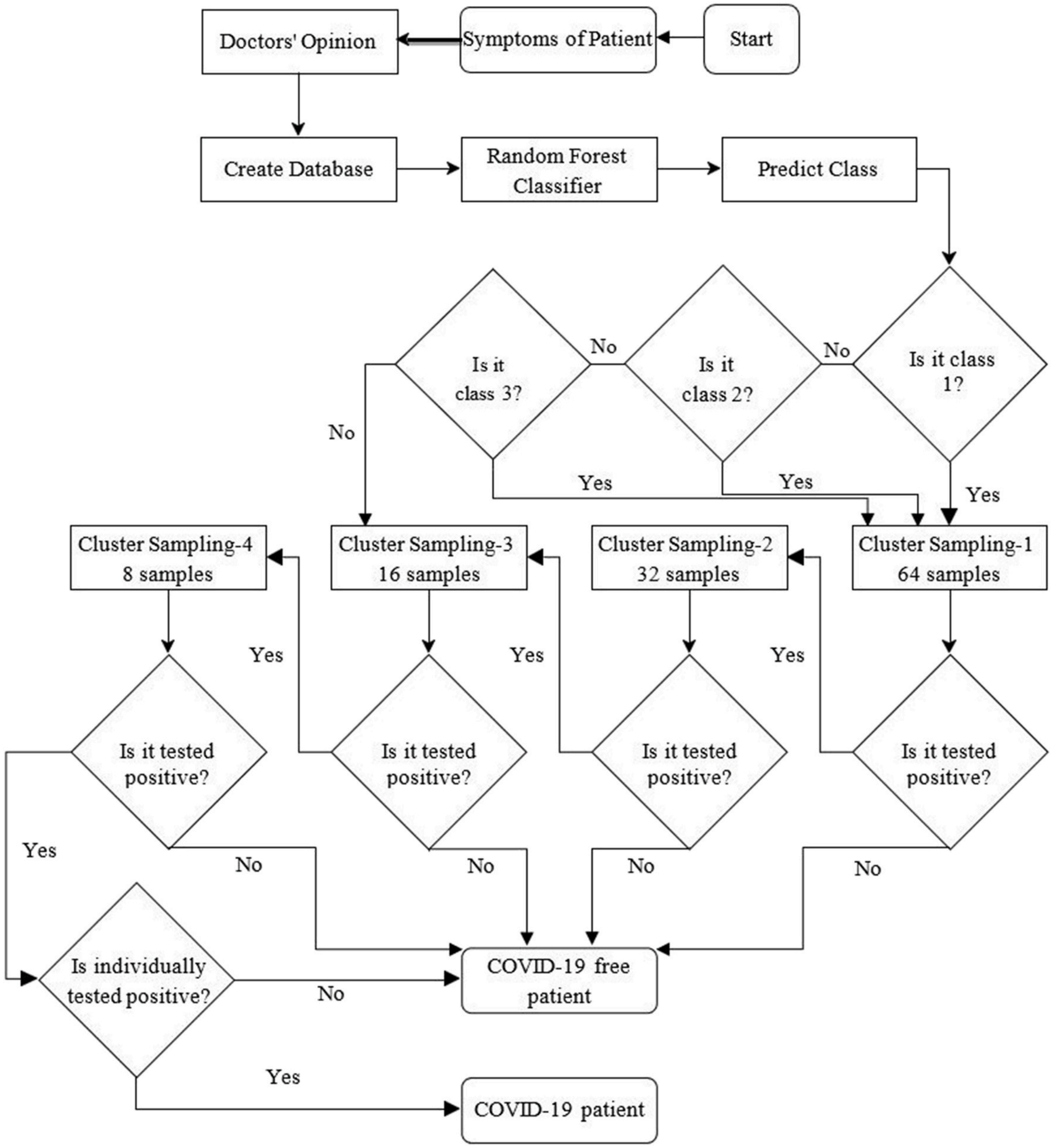

Fig. 6. Process flowchart of conditional cluster sampling (CCS).

4. Results

The key symptoms of COVID-19 found in the literature are classified into eight types with different levels that are shown in Table 1.

A survey was conducted on 4000 people from different areas of Bangladesh to get their symptoms. Then expert doctors' reports on suspected victims are collected. From the reports collected, the severity of the victims' condition is classified into four types that are delineated below:

i. Minor (class-1)
ii. Moderate (class-2)
iii. Major (class-3)
iv. Critical (class-4)

Thereafter, Random Forest Classifier (RFC) is used to predict patient

Table 1
Symptoms of COVID-19.

	Symptom-1	Symptom-2	Symptom-3	Symptom-4	Symptom-5	Symptom-6	Symptom-7	Symptom-8
	Fever	Cough	Sore in throat	Difficulty in breathing	Pain in chest	Runny nose	Headache	Pneumonia
Level	No 0–2 days 2–4 days More than 4 days	No Wet Dry	No Mild Moderate Severe	No Mild Moderate Severe	No Mild Moderate Severe	Yes No	No Mild Moderate Severe	Yes No

condition, where 90% of the data are used as training and the remaining as testing in R studio (version 3.6.3). We have tuned the model by coding, and the best model was found with ntree = 100 (number of trees in the random forest) and mtry = 2 (no of attributes selected randomly for each tree during attribute bagging in a random forest). The accuracy of the model is 96%, which indicates that the training dataset is well constructed. A random tree and the confusion matrix are depicted below in Fig. 7 and Table 2, respectively.

How outcomes are dependent on each of the symptoms has been shown in Fig. 7. For any random patient, if symptom-1 (fever) sustain more than or equal to 2 days, it will check whether it sustained more than or equal to 5 days. If so, the patient is considered to be a critical patient. If else, it will further check the level of symptom-4 (difficulty in breathing). If there is no breathing problem, it will be taken as a moderate patient. If else, it will again examine the level of symptom-5 (pain in chest). If the level is minor or moderate, it will look over symptom-2 (cough). If there is no cough, it will be a moderate patient else if else it will examine symptom-3 (sore in the throat). If there is a mild or moderate sore throat, the patient is minor else it will again check this symptom. If sore throat is severe it will be a critical patient otherwise it will be a moderate patient.

The predicted data is utilized to apply in CCS using R studio (version 3.6.3) to find out the total number of tests needed. Up until June 24, 2020, the COVID-19 positive cases in Bangladesh are 18% against the total number of tests performed [1]. In this study, the results are depicted in two ranges of probability levels for a patient being tested positive. For the first case, the maximum probability of a patient testing positive is assumed to be 25% (Table 3) and in the second case, the maximum probability is assumed to be 20% (Table 4). The probability range of patients to be considered COVID-19 positive are distributed for each class and in Tables 3 and 4, the first 2 columns show the number of patients in each severity category. Then after running the CCS method for 100 MCS cycles for the associated probability levels of testing positive, the average number of tests required for each category of patient is shown in column 4. In column 5, the percentages of tests saved are shown. In the last column, the number of times fewer tests were required in 100 cycles of MCS is shown. It represents out of 100 cycles of simulation how many times CCS required fewer tests than the traditional test which will give us the effectiveness of the CCS method.

From Table 3, it is seen that for the minor category, the probability of a patients' testing positive is assumed to be 10% or 0.10. Under this probability, the CCS method is applied for 100 MCS cycles, and for a total of 107 patients in this minor category, on average 88 tests are required per 100 cycles, which yields a 17.75% savings in terms of total test kits. Also, from the simulation, it is seen that among the 100 MCS cycles, in 98 cycles the CCS method expensed less number of test kits compared to the conventional individual testing method. Hence, the last

Table 2
Confusion matrix of patients' condition.

		Actual Patients' condition			
		Minor	Moderate	Major	Critical
Predicted patients' condition by RFC	Minor	101	0	6	0
	Moderate	3	137	0	0
	Major	1	0	99	1
	Critical	0	0	0	51

Table 3
The result of CCS at a higher probability level.

Patient condition	No of patients	Probability of being positive	Average Number of tests needed in pool testing in 100 cycles of MCS	Percentage of tests saved compared to traditional test	No of times fewer tests required in 100 cycles of MCS
Minor	107	0.10	88	17.75%	98
Moderate	140	0.15	125	10.71%	71
Major	101	0.20	90	10.89%	65
Critical	51	0.25	45	11.76%	35
Total	**399**		**348**	**12.78%**	

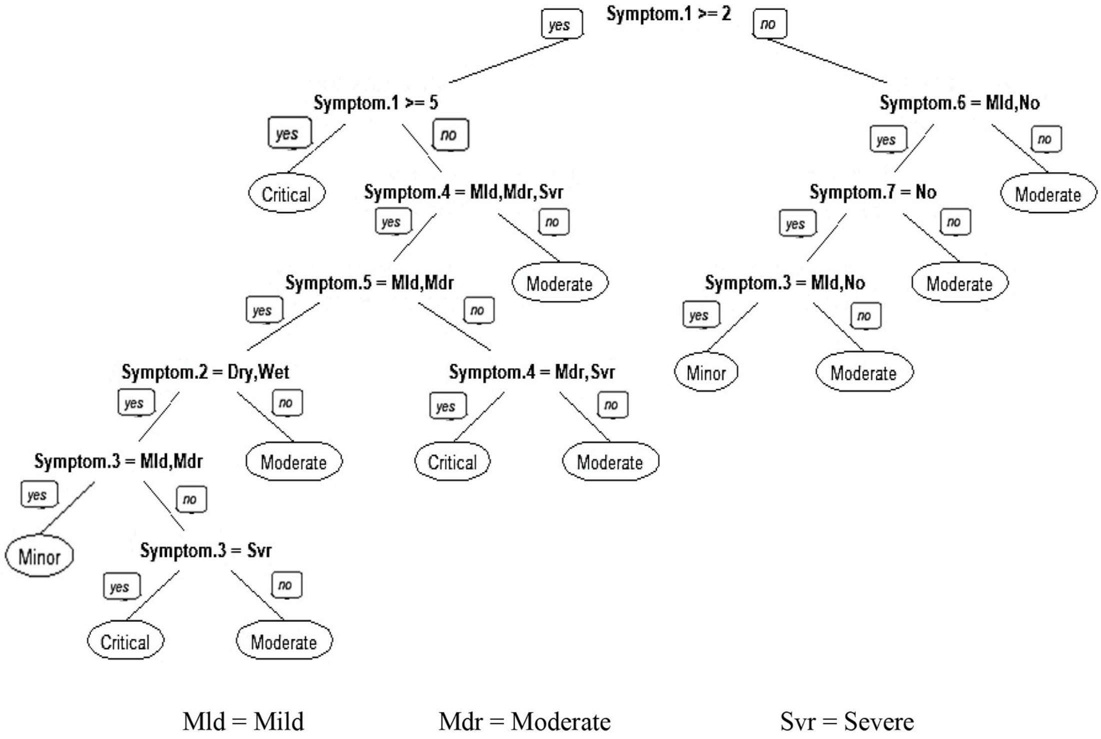

Fig. 7. A random tree from Random Forest predictor.

Table 4
The result of CCS at a lower probability level.

Patient condition	No of patients	Probability of being positive	Average Number of tests needed in pool testing in 100 cycles of MCS	Percentage of tests saved compared to traditional test	No of times fewer tests required in 100 cycles of MCS
Minor	107	0.05	68	36.45%	100
Moderate	140	0.10	112	20%	96
Major	101	0.15	85	15.84%	83
Critical	51	0.20	43	15.69%	60
Total	**399**		**308**	**22.81%**	

column of Table 3 shows the effectiveness of CCS for a higher probability level. Combining the results from all the categories, for 399 patients it takes only 348 tests on average for 100 MCS cycles, which results in a saving of 12.78% test kits for a maximum 25% probability of testing COVID-19 positive which prove that CCS outperforms the traditional testing method.

On the contrary, at a lower probability level (maximum 20%), the performance of CCS is found to be enhanced greatly in all aspects. The total percentage of reduction in the test kit is 22.81%, where the average number of tests needed in CCS for 100 cycles of MCS is only 308 against 399 total patients. In the last column where the effectiveness of CCS is shown, it is observed that in every cycle for a minor symptomatic patient, CCS outperformed the traditional testing method and for a critical patient, in 60 cycles CCS performed better. The results at a lower probability level along with the adjacent probability level are shown in Table 4.

The comparative analysis is depicted is Fig. 8 where the graph is showing the comparison between the test numbers needed to conduct the traditional individual testing method, CCS method under higher infection probability, and CCS under lower infection probability respectively under four patient classes. For minor classes, in the traditional testing method, 107 tests are needed for 107 patients. Under higher probability level 88 tests are needed whereas for lower probability level 68 tests are needed to test all 107 patients.

In Fig. 9, the comparative analysis is done between two probability levels regarding the percentage of test kits saved with a change in the severity class of a patient. In minor classes, almost 19% of test kits are less required in lower probability level of infection than higher probability level of infection in comparison with the traditional individual testing method.

In Fig. 10, the graphs show the number of times CCS performed better in all four severity levels of the patients, and also a comparison can be derived from this graph between the performance of the CCS method itself, respectively, under lower and higher probability levels of infection. This graph basically shows the effectiveness of the CCS method for different infection levels. The Y-axis data is plotted from column 6 of the table no. 3 and 4 where data is derived for 100 MCS cycles and the X-axis is plotted the severity of patients.

5. Discussion

The compiled results from Table 3 and Table 4 show that in all the cases, CCS takes up less test kits compared to the individual testing method. From both probability levels analysis, it is evident that CCS performs far better in patients with minor symptoms, but its performance degrades as the severity of the patient's condition increases. The main reason for this is that the probabilities of infection are taken to be less in these classes due to less severity of the physical condition. Also, among the less severe class, pool size starts from 64, which gives more windows to salvage the benefits of clustering. On the other hand, for critical patients, only 35 times out of 100 cycles (Critical class of patients in Table 3) and 60 times out of 100 cycles (Critical class of patients in Table 4), better results are obtained compared to the traditional individual testing method. This can be the result of a higher probability of infection in this class as well as the initial pool size which is 16. For critical patients, upon getting positive results in the 1st cluster test, individual tests are conducted among this class, like the traditional method. Thus, poorer results are obtained with CCS as patients' conditions tend to get more critical.

From Fig. 8, we can observe that in every simulation step, CCS takes up fewer test numbers than the traditional testing method. But it is also observed that with the advance in the severity level, the saving of test kits reduces as the sample size reduces. Again from Figs. 9 and 10, it is observed that the savings of test kits decrease as the probability of infection increases. So, it can be deduced that CCS saves comparatively fewer test kits in a scenario where there are higher infection rates. Also, in Fig. 10, the negative slope of both curves shows that as the severity class of patients' advances, the test kit saving decreases per 100 MCS cycles for both cases which means that CCS tends to be more like an individual testing method with an increase in the patients' severity. Also, the CCS method mainly relies on symptoms to classify the patients. So the method is not effective for asymptomatic patients.

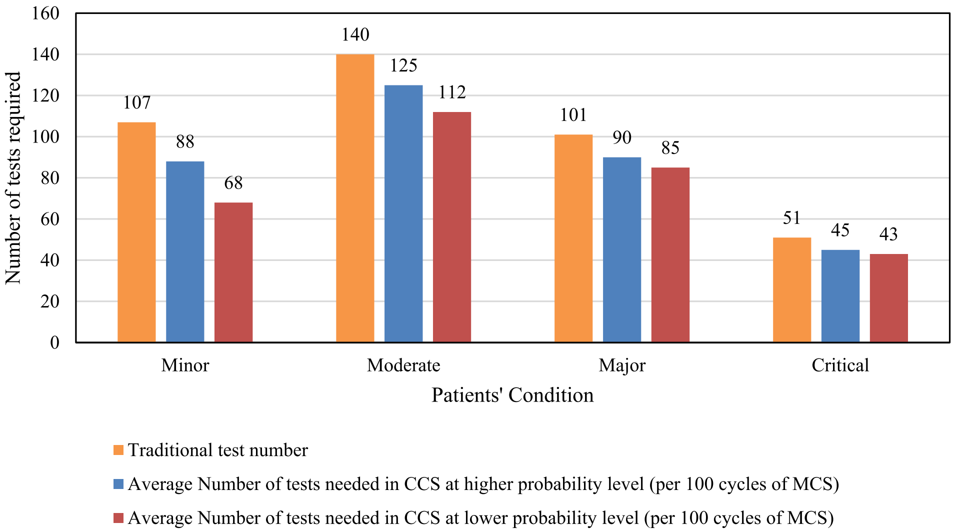

Fig. 8. Comparison of test numbers for traditional method and CCS (at both probability levels).

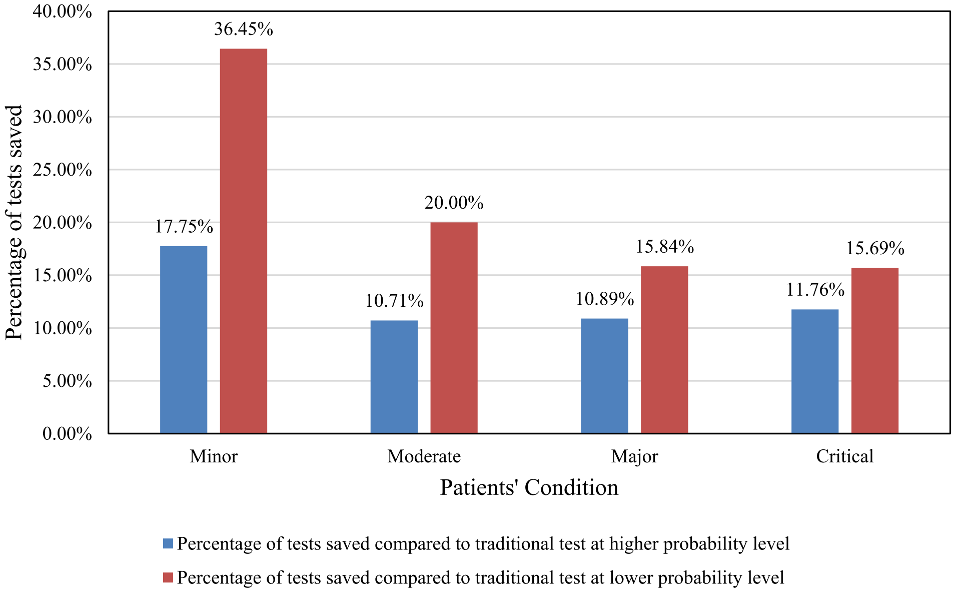

Fig. 9. Percentage of tests saved compared to traditional test at two probability levels.

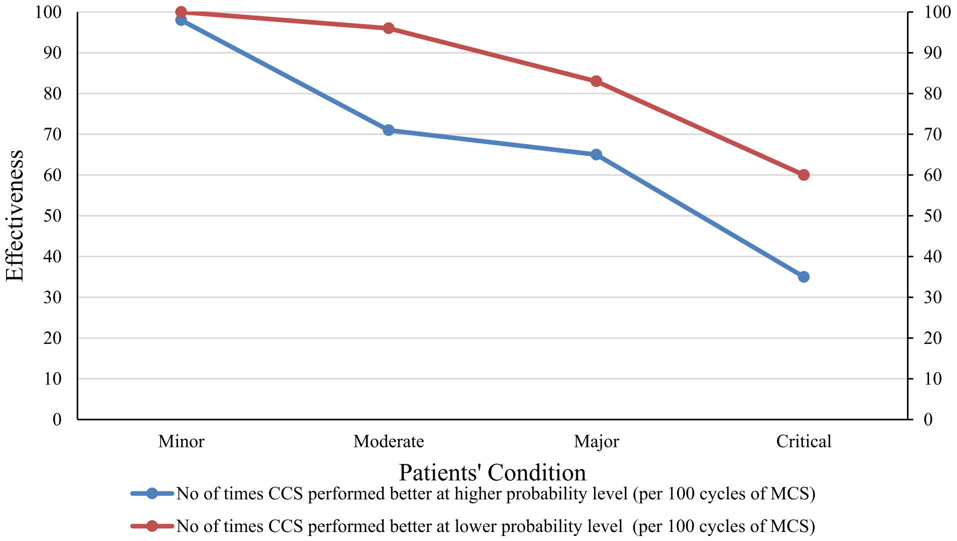

Fig. 10. Percentage of tests saved compared to traditional test at two probability levels.

6. Conclusion

Bangladesh is undergoing community transmission in the spreading COVID-19 and to address this, the initial focus has been on case identification. The case identification is currently very low due to a shortage of testing kits. This study suggests a means to mitigate this issue by utilizing conditional cluster sampling. This study incorporates a numerical method, probabilistic sampling, and health science to arrive at a systematic cluster specimen testing method, which is the CCS method. The accuracy of RFC to predict a patients' class is 96%. The CCS method is repeated for 100 cycles according to MCS, whuch resulted in a saving of 12% test kits for higher probabilities of positive cases detection and 22% for lower probabilities of positive cases detection of the test kits. This will save both time and money for rapidly obtaining test reports.

The CCS method is beneficial in terms of mass specimen testing. However, this study has some limitations-

1. The probability ranges are selected based on current statistics and infection patterns. The probability is contingent upon different infection patterns and situations.
2. The test data set is only 399 patients. Testing on a higher population will most likely derive a more accurate scenario.
3. The model does not consider asymptomatic patients.
4. Due to computational simplifications, 100 cycles of simulation is conducted in MCS. An increase in the number of cycles is likely to deliver a more precise result.

This study can also be explored using other intricate tools in the future. This research utilizes RFC to classify the test data which can also be done using Deep Learning or a Deep Neural Network algorithm to add more dimensions.

Declaration of competing interest

None Declared.

Acknowledgment

We would like to express our gratitude to all the doctors who give their valuable opinions on patient's reports. I would also like to extend my thanks to all others who have encouraged and have provided support directly or indirectly, throughout this work.

References

[1] "Bangladesh Coronavirus: 122,660 Cases and 1,582 Deaths - Worldometer." https://www.worldometers.info/coronavirus/country/bangladesh/ (accessed Jun. 24, 2020).

[2] Pouwels KB, Roope LSJ, Barnett A, Hunter DJ, Nolan TM, Clarke PM. Group testing for SARS-CoV-2: forward to the past? PharmacoEconomics - Open 2020;4(2): 207–10. https://doi.org/10.1007/s41669-020-00217-8.

[3] Dhawan D. Covid-19 update: how pool testing can enhance speed, scale. New Delhi: Hindustan Times; Apr. 05, 2020. New Delhi.

[4] Gilbert M, et al. Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. Lancet 2020;395(10227):871–7. https://doi.org/10.1016/S0140-6736(20)30411-6.

[5] Djalante R, et al. Review and analysis of current responses to COVID-19 in Indonesia: period of january to March 2020. Progress in Disaster Science 2020;6: 100091. https://doi.org/10.1016/j.pdisas.2020.100091.

[6] Huang C, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

[7] Harapan H, et al. Coronavirus disease 2019 (COVID-19): a literature review. Journal of Infection and Public Health 2020;13(5):667–73. https://doi.org/10.1016/j.jiph.2020.03.019.

[8] Hu Z, Ge Q, Li S, Jin L, Xiong M. Artificial intelligence forecasting of covid-19 in China. arXiv:2002.07112. arXiv preprint; 2020. p. 1–20 [Online]. Available, http://arxiv.org/abs/2002.07112.

[9] Tuli S, Tuli S, Tuli R, Gill SS. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet of Things 2020; 11:100222. https://doi.org/10.1016/j.iot.2020.100222.

[10] Ardabili SF, et al. COVID-19 outbreak prediction with machine learning. Algorithms 2020;13(10):249. https://doi.org/10.3390/a13100249.

[11] Khakharia A, et al. Outbreak prediction of COVID-19 for dense and populated countries using machine learning. Annals of Data Science 2020. https://doi.org/10.1007/s40745-020-00314-9. 0123456789.

[12] Singh S, Parmar KS, Makkhan SJS, Kaur J, Peshoria S, Kumar J. Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries. Chaos, Solit Fractals 2020;139. https://doi.org/10.1016/j.chaos.2020.110086.

[13] Parbat D, Chakraborty M. A python based support vector regression model for prediction of COVID19 cases in India. Chaos, Solit Fractals 2020;138:109942. https://doi.org/10.1016/j.chaos.2020.109942.

[14] Niazkar M, Niazkar HR. COVID-19 outbreak: application of multi-gene genetic programming to country-based prediction models. Electronic Journal of General Medicine 2020;17(5). https://doi.org/10.29333/ejgm/8232.

[15] Alrashed S, Min-Allah N, Saxena A, Ali I, Mehmood R. Impact of lockdowns on the spread of COVID-19 in Saudi Arabia. Informatics in Medicine Unlocked 2020;20: 100420. https://doi.org/10.1016/j.imu.2020.100420.

[16] Cohen T, et al. A combination of 'pooling' with a prediction model can reduce by 73% the number of COVID-19 (Corona-virus) tests. arXiv preprint; 2020. https://doi.org/10.11693/hyhz20181000233. arXiv:2005.03453.

[17] Lohse S, et al. Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. Lancet Infect Dis 2020;3099(20). https://doi.org/10.1016/S1473-3099(20)30362-5.

[18] Yelin I, et al. Evaluation of COVID-19 RT-qPCR test in multi-sample pools. medRxiv; 2020. https://doi.org/10.1101/2020.03.26.20039438.

[19] Aragón-Caqueo D, Fernández-Salinas J, Laroze D. Optimization of group size in pool testing strategy for SARS-CoV-2: a simple mathematical model. J Med Virol 2020. https://doi.org/10.1002/jmv.25929. April.

[20] Jia LIU, Yi C, Kefan XIE, Xiaohong C. Is pool testing method of COVID-19 employed in Germany and India effective?. 2020. p. 1–10.

[21] Al-Najjar H, Al-Rousan N. A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea. Eur Rev Med Pharmacol Sci 2020; 24(6):3400–3. https://doi.org/10.26355/eurrev_202003_20709.

[22] Waleed Salehi A, Baglat P, Gupta G. Review on machine and deep learning models for the detection and prediction of coronavirus. Mater Today: Proceedings 2020; xxxx. https://doi.org/10.1016/j.matpr.2020.06.245.

[23] Iwendi C, et al. COVID-19 patient health prediction using boosted random forest algorithm. Frontiers in Public Health 2020;8:1–9. https://doi.org/10.3389/fpubh.2020.00357. July.

[24] Coronavirus: Bangladesh divided into red, yellow, green zones. https://en.somoynews.tv/8611/news/Coronavirus-Bangladesh-divided-into-red-yellow-green-zones. accessed Jun. 24, 2020.

[25] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. https://doi.org/10.1201/9780367816377-11.

[26] Salzberg S. Book review: C4. 5: by j. ross quinlan. inc., 1993. programs for machine learning morgan kaufmann publishers. Mach Learn 1994;16:235–40. https://doi.org/10.1016/S0019-9958(64)90259-1.

[27] Liu C, White M, Newell G. Measuring the accuracy of species distribution models: a review. In: Proceedings 18th world IMACs/MODSIM congress. Cairns, Australia; 2009. p. 4241–7. July.

[28] Arfiani A, Rustam Z. Ovarian cancer data classification using bagging and random forest. AIP Conference Proceedings 2019;2168. https://doi.org/10.1063/1.5132473. November.

[29] Raychaudhuri S. Introduction to Monte Carlo simulation. In: Winter simulation conference. IEEE; 2008. p. 91–100.