Original Article

Access this article online

Quick Response Code:

Website:
www.jehp.net

DOI:
10.4103/jehp.jehp_1424_20

Performance evaluation of selected machine learning algorithms for COVID-19 prediction using routine clinical data: With versus Without CT scan features

Mostafa Shanbehzadeh[1], Hadi Kazemi-Arpanahi[2,3]*, Azam Orooji[4], Sara Mobarak[5], Saeed Jelvay[6]

[1]Assistant Professor of Health Information Management, Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran, [2]Assistant Professor of Health Information Management, Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran, [3]Assistant Professor of Health Information Management, Student Research Committee, Abadan University of Medical Sciences, Abadan, Iran, [4]Assistant Professor of Medical Informatics, School of Medicine, North Khorasan University of Medical Science, North Khorasan, Iran, [5]Assistant Professor of Infectious Diseases, School of Medicine, Abadan University of Medical Sciences, Abadan, Iran, [6]MSc of Health Information Technology, Department of Student Research Committee, Abadan University of Medical Sciences, Abadan, Iran

Address for correspondence:
Dr. Hadi Kazemi-Arpanahi, Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran.
E-mail: h.kazemi@abadanums.ac.ir

Received: 24-10-2020
Accepted: 19-11-2020
Published: 31-08-2021

Abstract:

BACKGROUND: Given coronavirus disease (COVID-19's) unknown nature, diagnosis, and treatment is very complex up to the present time. Thus, it is essential to have a framework for an early prediction of the disease. In this regard, machines learning (ML) could be crucial to extract concealed patterns from mining of huge raw datasets then it establishes high-quality predictive models. At this juncture, we aimed to apply different ML techniques to develop clinical predictive models and select the best performance of them.

MATERIALS AND METHODS: The dataset of Ayatollah Talleghani hospital, COVID-19 focal center affiliated to Abadan University of Medical Sciences have been taken into consideration. The dataset used in this study consists of 501 case records with two classes (COVID-19 and non COVID-19) and 32 columns for the diagnostic features. ML algorithms such as Naïve Bayesian, Bayesian Net, random forest (RF), multilayer perceptron, K-star, C4.5, and support vector machine were developed. Then, the recital of selected ML models was assessed by the comparison of some performance indices such as accuracy, sensitivity, specificity, precision, F-score, and receiver operating characteristic (ROC).

RESULTS: The experimental results indicate that RF algorithm with the accuracy of 92.42%, specificity of 75.70%, precision of 92.30%, sensitivity of 92.40%, F-measure of 92.00%, and ROC of 97.15% has the best capability for COVID-19 diagnosis and screening.

CONCLUSION: The empirical results reveal that RF model yielded higher performance as compared to other six classification models. It is promising to the implementation of RF model in the health-care settings to increase the accuracy and speed of disease diagnosis for primary prevention, screening, surveillance, and early treatment.

Keywords:

Artificial intelligence, computed tomography scan, coronavirus, COVID-19, data mining, machine learning, random forest

Introduction

Emerging and new pathogens are major threats for global public health. This is especially true for virus-induced diseases that are extremely contagious and have asymptomatic infectivity periods.[1-4] Since December 2019 a new strand of coronavirus named SARS-CoV2, which causes novel coronavirus disease (COVID-19) was detected in Wuhan District, China. It is thought that COVID-19 outbreak has animal origins that slipped from animal species into the human population.[5-7]

How to cite this article: Shanbehzadeh M, Kazemi-Arpanahi H, Orooji A, Mobarak S, Jelvay S. Performance evaluation of selected machine learning algorithms for COVID-19 prediction using routine clinical data: With versus without CT scan features. J Edu Health Promot 2021;10:285.

The COVID-19 outbreak still continues to spreading aggressively world-wide. Complex and highly contagious nature of COVID-19 had led the World Health Organization (WHO) to declare this outbreak a public health emergency.[8-10] In spite of taken drastic preventive measures and implement entire lockdown policies by many governments, COVID-19 now becomes a notable pandemic in global scale, which made tremendous impact on the health and safety of people all over the world, as well affecting their health status and causing a significant number of deaths. There are also other induced critical situations as indirect impacts of this pandemic, such as psychological distress, economic crises and might lead to serious challenges and threats in many societies.[11-14] The exponential daily increasing number of infected cases, and high rate mortality particularly in susceptible populations such as elderly, pregnant women, and people with underlying comorbidities such as low immune functions, cardiopulmonary diseases, cancer, infectious diseases, hypertension, and diabetes make it necessary to seek early detection and isolation positive cases as rapidly and accurately as possible for containing the transmission of the virus especially for asymptomatic cases in early stages.[15-20]

Medical diagnosis is intrinsic and intricate task that demands principally being accomplished precisely and proficiently.[21,22] In context of COVID-19, Owing to its unfamiliar many aspects of and the similarity of its primary symptoms to other respiratory infectious diseases, this makes it challenging for early differential diagnosis.[23,24] So far real-time polymerase chain reaction (RT-PCR) tests are complicated in operation and can take up to 2 day or even longer to get the results. As well, due to the low virus loads in early infected COVID-19 patients, RT-PCR tests show false negative results in a number of cases.[25,26] Besides RT-PCR, computed tomography (CT) scan has become a valuable method to assist in the diagnosis and management of symptomatic suspected COVID-19 cases. Nevertheless, it findings are normal or with minor radiological signs in some patients at early stages of disease and the lesions are usually small and their appearances are quite similar with that of other pneumonia. In addition, these techniques are time-consuming and increase the infection risk of the clinicians and usually prohibited for all suspected cases due to the limitation of resources.[27-29]

As a good alternative, artificial intelligence (AI) may be the unique preparation to take up this challenge.[30,31] AI can enable the machine to learn from past experience, adjust to new inputs, and simulating human intelligence tasks.[32,33] Machine learning (ML) is a subset of AI that uses computer algorithms seeking for hidden and previously unknown patterns from large sets of data.[34] A major focus of ML in health care is to automatically produce models and correlations from raw data and leverage this extracted useful information to make clinical decisions.[35] ML based predictive model is of great importance to prediction and prognosis of COVID-19 infection.[36] These are the important factors while the healthcare industry resources are limited to fighting against disease pandemic. Albahri *et al*. showed that the use of ML technologies to provide predictive models can be significantly helpful in a timely, effective, and economical diagnosis of the disease.[37] Sharma suggested the use of the ML algorithms compared to traditional methods is good choices for early stage disease screening. It is extremely important to identify the disease at early asymptomatic phases of COVID-19 and promptly confinement the infected cases.[38] The real-time and reliable diagnosis of COVID-19 through the use of computational ML algorithms is the foundation for the prompt management of the patients.[36,39] It could discriminate COVID-19 patients from other similar conditions with a better accuracy than other common and traditional approaches.[40,41] Therefore, in this COVID-19 big data era, we aimed to construct predicting diagnostic models using of selected ML algorithms such as Naïve Bayesian (NB), Bayesian Net (BN), random forest (RF), multilayer perceptron (MLP), K-star, C4.5, and support vector machine (SVM) on COVID-19 dataset. Then, the recital of selected ML was assessed on six diverse classification performance indices in presence and absence of chest CT-Scan features. In addition, the receiver operating characteristic (ROC) curve is also used for performance measurement. Finally, a data-driven predictive analytics model using the best performance was developed.

Materials and Methods

Dataset description and preprocessing

The dataset applied in this research has been obtained from Ayatollah Talleghani hospital, focal point center for COVID-19 care and treatment in South-West region of Khuzestan province, Iran. This dataset consists of 537 records/entries with two classes (COVID-19 and non COVID-19) and 32 columns for the features. After quantitative analysis of medical records, 36 incomplete case records which had a lot of missing data (more than 70%) were excluded from analysis. Finally, the 501 records were remained with 398 confirmed as positive cases (presence) and 103 healthy persons (absence). This study was reviewed and approved by the Intuitional Ethical Committee board of Abadan University of Medical Sciences (IR. ABADANUMS. REC.1399.064).

Feature selection

In this process, the most important parameters have been determined using a combination of extensive systematic literature review (SLR) coupled with

an expert consensus by the present research team at three epidemiological, clinical, and para-clinical categories.[42] Then a questionnaire was developed in three aforementioned categories, six data classes with 54 parameters. The content validity of the questionnaire was assessed by an expert panel including two infectious specialists and two virologists. In addition, test-retest (at 10-day interval) was done to evaluate the reliability of the questionnaire. The experts were asked to review the initial list of parameters to score each item according to their importance in developing COVID-19 based on a 5-point Likert scale, ranging from 1 to 5, where 1 indicated "not important" and 5 indicated "highly important". Only the parameters with average score of 3.75 (70%) and higher were allowed into the study (predictors features). Finally, the proposed clinical features were validated using Delphi survey by a group of multidisciplinary expert team [Table 1].

Model building

The predictive classifier models were developed for accurately diagnosis COVID-19 patients. The ML algorithms such as NB, BN, RF, MLP, K-star, C4.5, and SVM were used to developed prediction models. We considered these seven models due to their following characteristics.

Support vector machine

The SVM method was first introduced by vapnik on the basis of statistical learning theory. It was mainly formed for twofold classification. Yet, it can be successfully expanded for multi class problems. The aim of SVM is to find class-separating hyperplane in a multidimensional space that splits the feature space into two distinct groups.[43-45]

C4.5

C4.5 is an important decision tree algorithm. Its capabilities such as missing values accounting, decision tree pruning by determining confidence factors, extracting rule and considering continuous attribute value range, make the C4.5 algorithm a better choice than other tree algorithms. This algorithm uses divide and conquer strategies for decision tree making based on independent and dependent variables. In each node of the tree, the splitting function is done by attribute that can predict samples in each class more precisely. Initially, the C4.5 rule sets are made by unpruned tree and each path from root node to leaf transformed to a prototype rule that is associated to leaf node label.[46-48]

Table 1: Demographic characteristics of Delphi participants

Variables	Frequency (%)
Specialty	
Infectious disease	5 (41.66)
Radiologist	3 (25)
Virologist	2 (16.66)
Epidemiologist	2 (16.66)
Gender	
Male	8 (66.66)
Female	4 (33.33)
Work experience	
<10	6 (50)
10-20	4 (33.33)
20-30	2 (16.66)
Total	12 (100)

Random forest

This algorithm is applied for datasets with a large dimension. It uses additional layers of randomness than other decision tree algorithms. In contrary to other algorithms that the node splitting process are done by the best all variable split, in RF this process done by the subset of this predictors randomly. The diversity of trees is important in RF performance.[49,50]

Bayesian Net

BN is a probabilistic graphical models and knowledge representation technique for uncertainty management and decision making that formulates a set of random variables and their conditional dependencies within an annotated directed acyclic graph. These graphical structures are used to shows the variables that each variable occurs independently. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables.[51,52]

Multilayer perceptron

MLP is a feed forward ANN model for prediction of the class label of tuples that maps input data onto a set of appropriate outputs in three input, output and processing (hidden) layers. Each layer contains a group of neurons that are generally associated with all the neurons of the other layers in a directed graph. Except for the input nodes, every node is a neuron (or processing element) with a nonlinear activation.[53-55]

K-star

K-star is an instance based classifier that is the class of test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance based learners in that it uses an entropy based distance function.[56,57]

Naïve Bayesian

NB is a probability-based model that works on the basis of Bayes theorem and assumes independence between variables to perform a classification task. It is greatly leveraged in to classify and label the objects or points. In

addition, owing to its simplicity, this algorithm is very suitable for large datasets and produce highly accurate models.[58]

Model assessment

At first the performance of each model was evaluated for dataset with and without CT-Scan features. The evaluation of diagnostic models was done using of 10-fold cross validation according accuracy, specificity, precision, sensitivity, F measure, and ROC criteria [Table 2].

All models were implemented using R3.2.3 software (R Foundation for Statistical Computing). Then, the best model for the data set in absence of CT-scan results considered for developing the decision support model. Finally, the best model was used as the basis of decision support system, then the system tags and graphical user interface (GUI) was developed by Waikato Environment for Knowledge Analysis (Weka) version 3.8 developed at the university of Waikato, New Zealand and Visual-Studio 2018 (Microsoft) application software, respectively.

Results

Identification of clinical features

Overall, three data categories, six data classes and 54 data items (proposed for COVID-19 reporting) were identified from comprehensive SLR.[42] After a Delphi study with a group of multidisciplinary experts, the total number of disease criteria and predictors for basic, clinical, and para-clinical data categories were 12, 15, and 5, respectively [Table 3]. The results of this stage (sum 32 predictor) were used to design the ML data collection form.

Data collection

Data was gathered using a data collection form in six sociodemographic, life style, exposure, sign and symptoms, co-existing conditions and CT-scan classes. It consisted of 501 case records and 32 diagnosis criteria [ML features displayed in Table 4] stored in the patient's medical record (COVID-19 patients: 398; healthy person: 103). It should be noted that data collection was done in two stages, 1-with CT-Scan (whole of data), and 2-without CT-Scan features (whole of data except CT findings).

Developing and evaluation of models

In this step, seven ML algorithms including MLP, C4.5, RF, SVM, NB, BN and K-star, were designed for developing COVID-19 diagnostic model. Then the performance of each developed ML model was evaluated on the dataset (with and without CT-Scan findings) and the results are also presented. Figures 1 and 2 depicted the performance metrics for both datasets.

Model performance evaluation is a fundamental part of building effective ML model. The evaluation can be carried out with a set of performance indices, most of whom are derived from confusion matrix.

As shown in Figure 1, the best predicting model for COVID-19 using clinical and CT-scan features is RF algorithm, however based on F-measure, the Bayesian methods yielded better performance. In Figure 2 also according accuracy, precision and sensitivity, the best performance for predicting COVID-19 only using clinical features (without CT results) was RF, but according F-measure the NB yield better performance, it caused that NB in terms of ROC placed after K-star with slight difference. But since the purpose of making this predictive model is screening the referred people to non-specialist medical centers, the best models will be selected based on sensitivity. Because in a model that has high sensitivity, the positive class is well recognized. Accordingly, the best method for both datasets (with and without CT features) is RF.

In Figure 3, the increase in specificity in all models and the decrease in sensitivity in both Bayesian models show that the elimination of risk factors related to CT features led to better detection of negative samples but had little effect on positive samples. However, in the RF, the value of all performance criteria increased by an average of 3.02. In nonspecialist medical centers, the diagnosis is usually made only on the basis of clinical examination and patient history. As a result, RF is the best model for screening people in these centers. Eventually, the RF tags was designed by Weka application software version 3.8 [Figure 4] and it GUI developed by Visual Studio 2018 application software [Figure 5].

Table 2: The calculation of performance metric

Performance criteria	Calculation
Accuracy	
Precision	
Sensitivity/recall	
Specificity	
F-measure	

FP=False positive, TP=True positive, FN=False negative, TN=True negative

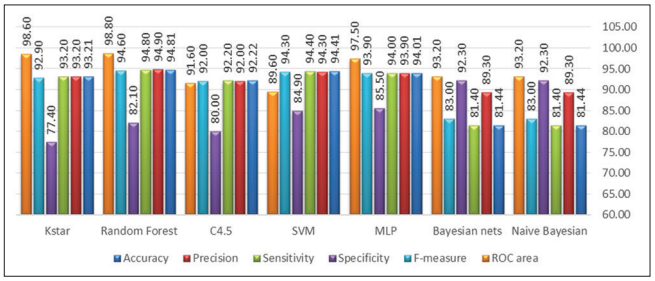

Figure 1: Evaluating ML algorithms in the presence of CT-scan data

Table 3: Evaluation of machines learning criteria and features

Data classes/items	Mean (%)	Final decision
Basic data		
Sociodemographic		
Age	4.26 (85.2)	Accept
Sex	3.90 (78)	Accept
Living situation	4.47 (89.48)	Accept
Occupation status	4.53 (90.53)	Accept
Nationality/race	3.37 (67.36)	Refuse
BMI	4.21 (84.21)	Accept
Life style		
Drug addiction	3.95 (78.95)	Accept
Alcohol consumption	3.90 (78)	Accept
Living in epidemic area	3.35 (67)	Refuse
Recent travelling	4.26 (85.20)	Accept
Exercise	2.47 (49.47)	Refuse
Exposure data		
Exposure history (yes, no, unknown)	4.26 (85.2)	Accept
Exposure frequency	4.05 (81)	Accept
Contact with suspicious person	4.36 (87.36)	Accept
Transmission mode (person-person, contaminated surface, other)	3.78 (75.79)	Accept
Susceptible population	2.42 (48.42)	Refuse
Clinical data		
Clinical manifestations (sign and symptoms)		
Fever	4.53 (90.53)	Accept
Dry cough	4.37 (87.37)	Accept
Sputum/expectoration	4.05 (81)	Accept
Dyspnea	4.26 (85.2)	Accept
Myalgia or fatigue	3.58 (71.57)	Refuse
Headache	2. 47 (49.47)	Refuse
Sore throat	3.84 (76.84)	Accept
Dizziness	2.42 (48.42)	Refuse
Rhinorrhea	3.53 (70.53)	Accept
Chest pain	3.78 (75.79)	accept
Pharyngeal congestion	2.47 (49.47)	Refuse
Chill	3.35 (67)	Refuse
Night sweat	3.84 (76.84)	Accept
Abdominal pain	2.37 (47.37)	Refuse
Diarrhea	2.15 (43.16)	Refuse
Anorexia	2.73 (54.73)	Refuse
Vomiting and nausea	3.84 (76.84)	Accept
Respiratory rate (per min)	4.16 (83.16)	Accept
Heart rate (beats/per min)	3.42 (68.42)	Refuse
Blood pressure (mmHg)	3.53 (70.53)	Refuse
Number sign or symptom (asymptomatic)	4.32 (86.32)	Accept
Co-existing conditions (comorbidities)		
Cardiovascular	2.95 (58.95)	Refuse
Cerebrovascular	1.84 (36.84)	Refuse
Diabetes	2.47 (49.47)	Refuse
Malignant tumors	2.47 (49.47)	Refuse
Upper respiratory diseases	4.84 (96.84)	Accept
Other chronic co-morbidities	3.79 (75.8)	Accept
Long-term use of immunosuppressive	3.84 (76.84)	Accept
Pregnancy	3.79 (75.8)	Accept
Disease complication	3.35 (67)	Refuse
Disease severity (mild, moderate, severe, critical)	3.32 (66.32)	Refuse
Disease status (active, inactive, recovered)	2.1 (42)	Refuse

Contd...

Table 3: Contd...

Data classes/items	Mean (%)	Final decision
Para clinical data		
CT-scan features		
Pattern of the lesion (GGS, consolidation, both)	3.84 (76.84)	Accept
Lesion distribution (unilateral, bilateral)	4.32 (86.32)	Accept
Lesion morphology (patchy, spherical, both)	3.26 (65.26)	Refuse
Lesion staging (early, progression, severe)	3.84 (76.84)	Accept
Lesion location (peripheral, central, both)	4.05 (81.05)	Accept
Involved lobe (right, left, both)	4.05 (81.05)	Accept

BMI=Body mass index, GGS=Gorlin-Goltz syndrome, CT=Computed tomography

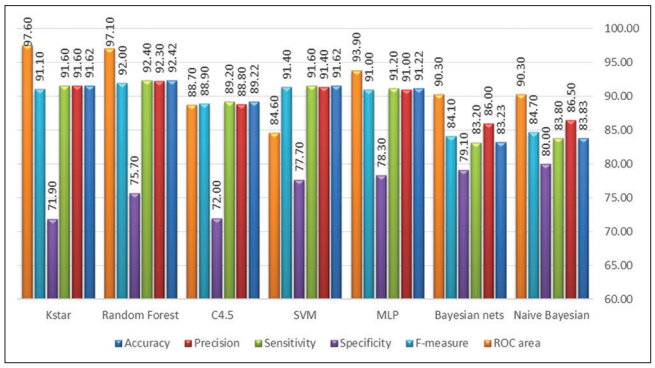

Figure 2: Evaluating ML algorithms in the absence of CT-scan data

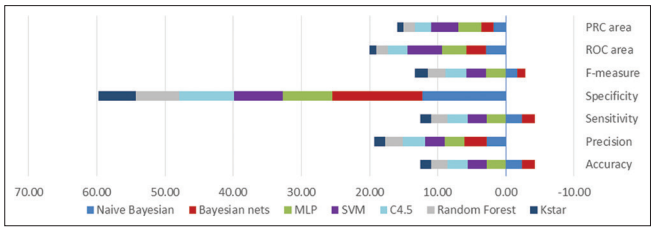

Figure 3: The difference of ML performance metrics for COVID-19 Prediction

Discussion

Deadly complications, long incubation period, difficulties of testing, vague characteristics, and uncertainty of nature, high-transmission power, and differential diagnosis with other respiratory diseases make COVID-19 a very critical public health issue that has captured the attention of the worldwide public.[59] In this regard, for appropriate preparedness against to ongoing global pandemic, WHO, and scientific community across the world have are suggesting for various innovative technologies and IT-based solutions to monitoring of infected patients, find best clinical trials, accurate patient screening and diagnosis, control the spread of virus, and tracking of infected patients.[60] Timely and accurate diagnosis of COVID-19, provide a better plan for health policymakers and clinician in order to mitigating disease outbreak and improved patient survival probability.[61]

Although multiple studies have shown that traditional statistical modeling can offer exact models, today AI-based solutions (computational techniques) could play a pivotal role to mining high-quality models and relationships.[34,62] Nowadays, like prior pandemics, health-care industries, and clinicians worldwide employed various novel technologies such as AI (intelligent diagnoses based ML or deep learning (DL) algorithms) to fight against COVID-19 and address the challenges during this severe universal life-threatening disease.[63] Developing diagnostic prediction models for pandemic diseases such as COVID-19 is very imperative in determining their likely new cases at early stage.[64,65]

The purpose of the current study was to apply ML algorithms to devise a prediction model to identify of likely infected cases by providing an accurate and reliable tool to help medical decision makers and triage COVID-19 patients more effectively and accurately. Predictive models based ML models for COVID-19 diagnosis can greatly contribute to recognize high risk groups, early detection of disease, and adoption of effective treatment plans. Using of ML for mining of large amounts of dataset is essential for optimal prevention, screening, care, treatment and tracing of COVID-19. This led to reducing uncertainty and ambiguity by offering evidence-base medicine for risk analysis, prediction, and treatment.[66]

Given that ML techniques can efficiently extract and exploring of hidden patterns in large datasets and identifying feature correlations, hence this article is focused on ML techniques to dealing with preparedness for COVID-19. This study aimed to the evaluation of seven ML algorithms on COVID-19 dataset in the presence and absence of CT-Scan findings. In addition, we performed feature selection to determine the most important feature set for this task by using an extensive literature review alongside expert consensus approach. These methods were validated using 10-fold cross validation and evaluated in terms of various performance measurements.

Several researches have been published focused on applying and evaluating of ML techniques in COVID-19 early prognosis, screening, risk assessment, and trend

Table 4: The important diagnostic criteria

Categories	Variable	Values	Frequency (%)
Basic data	Age	Infant<12 months	5 (1)
		Child (1-5 year)	8 (1.6)
		Adolescent (5-17 year)	21 (4.2)
		Young (18-34)	218 (43.5)
		Middle age (34-65 year)	186 (37.1)
		Old (>65 years old)	63 (12.5)
	Gender	Female	225 (44.9)
		Male	276 (55.1)
	Type of occupation	High risk	50 (10)
		Middle-risk	172 (34.3)
		Low risk	132 (26.3)
		Unemployed	147 (29.3)
Epidemiological	Statues of geographic region	Low risk	0 (0)
		Middle risk	19 (3.8)
		High risk	482 (96.2)
	High risk contact	No	191 (38.1)
		Yes	310 (61.9)
	Type of high risk contact	Contact with patients	229 (45.7)
		Contact with contaminated surfaces	59 (11.8)
		Contact with contaminated foods/water	7 (1.4)
		Contact with contaminated air	8 (1.6)
		Other	7 (1.4)
		No contact	191 (38.1)
	Contact frequency	Usually	202 (40.3)
		Rarely	108 (21.6)
		No contact	191 (38.1)
	BMI	<18.5	22 (4.4)
		18.5-24.9	237 (47.3)
		25-29.2	216 (43.1)
		>30	26 (5.2)
Clinical findings	Chest pain	No	399 (79.7)
		Yes	102 (20.3)
	Angina	No	356 (71.1)
		Yes	145 (28.9)
	Runny nose	No	391 (78)
		Yes	110 (22)
	Headache	No	360 (71.9)
		Yes	141 (28.1)
	Sputum	No	448 (89.4)
		Yes	53 (10.6)
	Night sweats	No	371 (74.1)
		Yes	130 (25.9)
	Diarrhea	No	279 (55.7)
		Yes	222 (44.3)
	Loss of taste or smell	No	355 (70.9)
		No	146 (29.1)
	Nausea and vomiting	No	421 (84)
		Yes	80 (16)
	Underlying chronic diseases	No	392 (78.2)
		Yes	109 (21.9)
	History of upper respiratory tract infections	No	372 (74.3)
		Yes	129 (25.7)
	History of ARDS	No	410 (81.8)
		Yes	91 (18.2)

Contd...

Table 4: Contd...

Categories	Variable	Values	Frequency (%)
	Recent travel	No	346 (69.1)
		Yes	155 (30.9)
	Immunosuppressive drugs	No	389 (77.6)
		Yes	112 (22.4)
	Addiction	No	405 (80.8)
		Yes	96 (19.2)
	Alcohol abuse	No	469 (93.6)
		Yes	32 (6.4)
	Pregnancy	No	483 (96.4)
		Yes	18 (3.6)
	SPO2	>95	371 (74.1)
		85-95	125 (24.9)
		<85	5 (1)
CT-Scan features	Pulmonary lesion	No	255 (50.9)
		Yes	246 (49.1)
	Diffusion status	One-sided	158 (31.5)
		Two-sided	88 (17.6)
		Non	255 (50.9)
	Appearance of the lesion	GGO	159 (31.7)
		Consolidation	71 (14.2)
		Hybrid	16 (3.2)
		Non	255 (50.9)
	Lesion position	Centered	160 (31.9)
		Distributed	63 (12.5)
		Hybrid	23 (4.6)
		Non	255 (50.9)

CT=Computed tomography, BMI=Body mass index, GGO=Ground-glass opacity, ARDS=Acute respiratory distress syndrome, SPO2=Oxygen saturation

```
RandomTree
==========

Fever = No
|   Nausea-vomiting = No
|   |   Lesion-Situation = No
|   |   |   Dyspnea = No
|   |   |   |   History-ARDS = No
|   |   |   |   |   History-Immuno_drug = No
|   |   |   |   |   |   Cough = No
|   |   |   |   |   |   |   Diarrhea = No : 0 (6/0)
|   |   |   |   |   |   |   Diarrhea = Yes
|   |   |   |   |   |   |   |   Rhinorhea = No : 0 (2/0)
|   |   |   |   |   |   |   |   Rhinorhea = Yes : 1 (3/0)
|   |   |   |   |   |   Cough = Yes
|   |   |   |   |   |   |   History-of-contact = No
|   |   |   |   |   |   |   |   Rhinorhea = No
|   |   |   |   |   |   |   |   |   Diarrhea = No : 1 (3/0)
|   |   |   |   |   |   |   |   |   Diarrhea = Yes : 0 (3/1)
|   |   |   |   |   |   |   |   Rhinorhea = Yes
|   |   |   |   |   |   |   |   |   SPO2 = >95
|   |   |   |   |   |   |   |   |   |   Diarrhea = No : 0 (8/2)
|   |   |   |   |   |   |   |   |   |   Diarrhea = Yes : 0 (6/0)
|   |   |   |   |   |   |   |   |   SPO2 = 85-95 : 0 (1/0)
|   |   |   |   |   |   |   |   |   SPO2 = <85 : 1 (0/0)
|   |   |   |   |   |   |   History-of-contact = Yes : 0 (4/0)
|   |   |   |   |   History-Immuno_drug = Yes : 0 (9/0)
|   |   |   |   History-ARDS = Yes
|   |   |   |   |   SPO2 = >95 : 0 (4/0)
|   |   |   |   |   SPO2 = 85-95
|   |   |   |   |   |   Cough = No
|   |   |   |   |   |   |   Rhinorhea = No : 0 (1/0)
|   |   |   |   |   |   |   Rhinorhea = Yes : 1 (2/0)
|   |   |   |   |   |   Cough = Yes
|   |   |   |   |   |   |   Rhinorhea = No
|   |   |   |   |   |   |   |   Diarrhea = No : 1 (2/0)
|   |   |   |   |   |   |   |   Diarrhea = Yes : 0 (2/0)
|   |   |   |   |   |   |   Rhinorhea = Yes : 0 (3/0)
|   |   |   |   |   SPO2 = <85 : 1 (0/0)
|   |   |   Dyspnea = Yes : 0 (13/0)
|   |   Lesion-Situation = Central : 1 (4/0)
|   |   Lesion-Situation = Spreaded : 1 (0/0)
|   |   Lesion-Situation = Compound : 1 (0/0)
|   Nausea-vomiting = Yes : 0 (32/0)
Fever = Yes
```

Figure 4: RF schematic diagram

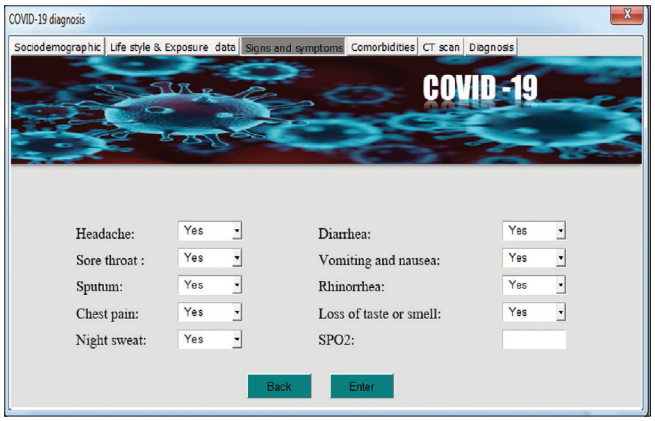

Figure 5: Clinical decision support system user interface by Visual Studio 2018

estimation. Narin *et al*., study showed deploying RF algorithm for COVID-19 detection was more efficient than Convolutional Neural Network (CNN) based on ResNet50, InceptionV3, and Inception-ResNetV2 models.[67] Yasar *et al*. compared the selected ML techniques including SVM, k-NN, and CNN classification success in COVID-19 detection from CT lung images. Finally, the most successful results are obtained using CNN technique.[68] Rodriguez *et al*. concluded a better fit between the observed data and those obtained by the computational ANN model than two Gompertz

and Logistic techniques.[14] Moftakhar *et al*. showed that ARIMA prediction model was more accurate than ANN, SVM and RF techniques.[69] Alakus *et al*., assessed the performance of selected DL algorithms including ANN, CNN, LSTM, RNN, CNNLSTM, and CNNRNN for COVID-19 diagnosis and the best meaningful results observed from LSTM model.[70] It is proven using DL reached better accuracy, and AUC scores than ML classifiers.[68,71] However in this study, we did not use of DL approaches. We developed seven different ML application models, and the empirical results reveal that RF algorithm yielded higher performance as compared to other six ML techniques. The experimental results indicate that our suggested model (RF) distinguishing between patients and healthy cases at an accuracy of 86.66%, F1-score of 91.89%, precision of 86.75%, recall of 99.42%, and AUC of 62.50%.

The capabilities of selected ML algorithms in COVID-19 diagnosis assessed based on 32 clinical variables and then were compared according existence or absence of chest CT-Scan features. For that end, the data were standardized and used as inputs for the ML algorithms. Later, classification was performed and the performances of the models were measured. The best accuracy, recall and AUC values were obtained with RF model of 92.3%, 93.68%, and 90.00%, respectively. By comparing the results of model performance assessment, it could be found that after conduct feature selection, the performance of all seven algorithms was improved and better than the baseline model. Finally, after adding CT-scan data, the performance of suggested ML algorithm did not make much difference anyway.

This study has several limitations that need to be addressed. First, the major limitation is the size of the dataset. 501 patient's information from only one medical center was considered as sample size although external validation of variables was conducted by multi-disciplinary of medical and public health experts. Second, some laboratory variables could not be measured for some patients. In addition to these, the data were imbalanced or uneven, thus we balanced the data by omitting some fields. The performance of these models can be enhanced with a larger dataset. Further studies need to be carried out with more optimized clinical parameters from bigger and multicenter databases. Third, we did not compare the performance of algorithms other than the seven algorithms that used in this study. It requires further investigations that exploiting different ML or even DL algorithms on COVID-19 dataset. In future research, the scope of application of the model should be expanded by incorporating more comprehensive training data. This study introduced the ML techniques as an effective and practical alternative to routine para-clinical measures for identifying COVID-19 positive patients. This is in particular beneficial in those developing countries like Iran which is exposed to heavy sanctions and suffering from shortages of essential healthcare resources.

Conclusion

In this paper, the efficiency of several ML classifying algorithms was analyzed and compared in predicting of COVID-19 by using easily available clinical features in presence and absence of CT-scan features. All features were validated by medical and public health experts. It has been observed that RF model performed best on classification accuracy better than the other six ML algorithms. This study may assist future researchers, policy makers and clinicians in choosing the optimal predictive models for taken evidence-based decisions to better dispense their resources and plan ways to overall prevent or at least decrease its outbreak. The comparison results of diagnostic models' performance in this study were satisfactory to some extent, and we believed further investigations are needed to validate our model to predict COVID-19 in various, larger, multi-central, and qualitative dataset.

Acknowledgments

This article is extracted from a research project supported by Abadan University of Medical Sciences (IR. ABADANUMS. REC.1399.064). We also thank the Research Deputy of Abadan University of Medical Sciences for financially supporting this project. We also would like to thank all experts who participated in this study and played a role in the validation of the developed models.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Rao AS, Vazquez JA. Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone–based survey when cities and towns are under quarantine. Infect Control Hosp Epidemiol 2020;41:826-30.
2. Chan JF, Yuan S, Kok KH, To KK, Chu H, Yang J, *et al*. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. Lancet 2020;395:514-23.
3. Bikdeli B, Talasaz AH, Rashidi F, Sharif-Kashani B, Farrokhpour M, Bakhshandeh H, *et al*. Intermediate versus standard-dose prophylactic anticoagulation and statin therapy versus placebo in critically-ill patients with COVID-19: Rationale and design of the INSPIRATION/INSPIRATION-S studies. Thromb Res 2020;196:382-94.
4. Panahi S, Ashrafi-Rizi H, Panahi M. Exposure to coronavirus (COVID-19) using narrative and simulated experience approaches: A commentary. J Educ Health Promot 2020;9:135.

5. Peeri NC, Shrestha N, Rahman MS, Zaki R, Tan Z, Bibi S, *et al*. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: What lessons have we learned? Int J Epidemiol 2020;49:717-26.
6. Mackenzie JS, Smith DW. COVID-19-A novel zoonotic disease: A review of the disease, the virus, and public health measures. Asia Pac J Public Health 2020;32:145-53.
7. Yoo HS, Yoo D. COVID-19 and veterinarians for one health, zoonotic-and reverse-zoonotic transmissions. J Vet Sci 2020;21:e51.
8. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R. Features, evaluation and treatment coronavirus (COVID-19). In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020.
9. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, *et al*. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Int J Surg 2020;76:71-6.
10. Shanbehzadeh M, Kazemi-Arpanahi H, Mazhab-Jafari K, Haghiri H. Coronavirus disease 2019 (COVID-19) surveillance system: Development of COVID-19 minimum data set and interoperable reporting framework. J Educ Health Promot. 2020 Aug 31;9:203.
11. Jacobsen KH. Will COVID-19 generate global preparedness? Lancet 2020;395:1013-4.
12. Wang P, Zheng X, Li J, Zhu B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos Solitons Fractals 2020;139:110058.
13. El Zowalaty ME, Järhult JD. From SARS to COVID-19: A previously unknown SARS-related coronavirus (SARS-CoV-2) of pandemic potential infecting humans – Call for a one health approach. One Health 2020;9:100124.
14. Torrealba-Rodriguez O, Conde-Gutiérrez RA, Hernández-Javier AL. Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. Chaos Solitons Fractals 2020;138:109946.
15. Liu Y, Wang Z, Ren J, Tian Y, Zhou M, Zhou T, *et al*. A COVID-19 risk assessment decision support system for general practitioners: Design and development study. J Med Internet Res 2020;22:e19786.
16. ALOM, Md Zahangir, *et al*. COVID_MTNet: COVID-19 Detection with Multi-Task Deep Learning Approaches. arXiv preprint arXiv: 2004.03747, 2020.
17. Bansal A, Padappayil RP, Garg C, Singal A, Gupta M, Klein A. Utility of artificial intelligence amidst the COVID 19 pandemic: A review. J Med Syst 2020;44:156.
18. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. Int J Antimicrob Agents 2020;55:105924.
19. Hussain A, Bhowmik B, do Vale Moreira NC. COVID-19 and diabetes: Knowledge in progress. Diabetes Res Clin Pract 2020;162:108142.
20. Moujaess E, Kourie HR, Ghosn M. Cancer patients and research during COVID-19 pandemic: A systematic review of current evidence. Crit Rev Oncol Hematol 2020;150:102972.
21. Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. Health Informatics J 2020;26:264-86.
22. Gunčar G, Kukar M, Notar M, Brvar M, Černelč P, Notar M, *et al*. An application of machine learning to haematological diagnosis. Sci Rep 2018;8:1-12.
23. Bryce C, Ring P, Ashby S, Wardman JK. Resilience in the face of uncertainty: Early lessons from the COVID-19 pandemic. J Risk Res 2020; 23 (7): 140-8. [Doi: https://doi.org/10.1080/13669877.2020 0.1756379].
24. Chater N. Facing up to the uncertainties of COVID-19. Nat Hum Behav 2020;4:439.
25. Mei X, Lee HC, Diao KY, Huang M, Lin B, Liu C, *et al*. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat Med. 2020 Aug; 26 (8):1224-1228. [Doi: 10.1038/ s41591-020-0931-3].
26. Wang Y, Song W, Zhao Z, Chen P, Liu J, Li C. The impacts of viral inactivating methods on quantitative RT-PCR for COVID-19. Virus Res 2020;285:197988.
27. Brogna B, Bignardi E, Salvatore P, Alberigo M, Brogna C, Megliola A, *et al*. Unusual presentations of COVID-19 pneumonia on CT scans with spontaneous pneumomediastinum and loculated pneumothorax: A report of two cases and a review of the literature. Heart Lung 2020; 49:864-68. [Doi: 10.1016/j. hrtlng. 2020.06.005].
28. Chen J, Peng S, Zhang B, Liu Z, Liu L, Zhang W. An uncommon manifestation of COVID-19 pneumonia on CT scan with small cavities in the lungs: A case report. Medicine (Baltimore) 2020;99:e21240.
29. Gündüz Y, Öztürk MH, Tomak Y. The usual course of thorax CT findings of COVID-19 infection and when to perform control thorax CT scan. Turk J Med Sci 2020;50:684-6.
30. Hassanien AE, Salama A, Darwsih A. Artificial intelligence approach to predict the COVID-19 patient's recovery. No 3223 Easy Chair; 2020.
31. Jin C, Chen W, Cao Y, Xu Z, Tan Z, Zhang X, *et al*. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 2020;11:5088.
32. Mei X, Lee HC, Diao KY, Huang M, Lin B, Liu C, *et al*. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat Med 2020;26:1224-8.
33. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr 2020;14:337-9.
34. Wong ZSY, Zhou J, Zhang Q. Artificial Intelligence for infectious disease big data analytics. Infect Dis Health 2019;24:44-8.
35. Wu CC, Yeh WC, Hsu WD, Islam MM, Nguyen PAA, Poly TN, *et al*. Prediction of fatty liver disease using machine learning algorithms. Comput Methods Programs Biomed 2019;170:23-9.
36. Vinod DN, Prabaharan SRS. Data science and the role of artificial intelligence in achieving the fast diagnosis of COVID-19. Chaos Solitons Fractals 2020;140:110182.
37. Albahri AS, Hamid RA, Alwan JK, Al-Qays ZT, Zaidan AA, Zaidan BB, *et al*. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): A systematic review. J Med Syst 2020;44:122.
38. Sharma S. Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: A study on 200 patients. Environ Sci Pollut Res Int 2020;27:37155-63.
39. Mantas J. Setting up an Easy-to-use machine learning pipeline for medical decision support: A case study for COVID-19 diagnosis based on deep learning with CT scans. Importance Health Inform Public Health Pandemic 2020;272:13.
40. Li WT, Ma J, Shende N, Castaneda G, Chakladar J, Tsai JC, *et al*. Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis. BMC Med Inform Decis Mak 2020;20:247.
41. Agbehadji IE, Awuzie BO, Ngowi AB, Millham RC. Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. Int J Environ Res Public Health 2020;17: 5330.
42. Shanbehzadeh M, Kazemi-Arpanahi H. Development of minimal basic data set to report COVID-19. Med J Islam Repub Iran 2020;34:754-63.
43. Chapman BP, Weiss A, Duberstein PR. Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. Psychol Methods 2016;21:603-20.

44. Chao CM, Yu YW, Cheng BW, Kuo YL. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J Med Syst 2014;38:106.
45. Wang GL, Li YF, Bi DX. Support vector networks in adaptive friction compensation. IEEE Trans Neural Netw 2007;18:1209-19.
46. Abdar M, Kalhori SR, Sutikno T, Subroto IM, Arji G. Comparing performance of data mining algorithms in prediction heart diseases. Int J Electr Comput Eng (2088-8708) 2015;5:1569-76. [Doi: 10.11591/ijece. v5i6.pp1569-1576].
47. Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. Int J Comput Appl 2014;98:41. [Doi: 10.1186/s12911-019-0790-3].
48. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, *et al*. Top 10 algorithms in data mining. Knowl Inf Syst 2008;14:1-37.
49. Ozçift A. Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Comput Biol Med 2011;41:265-71.
50. Liaw A, Wiener M. Classification and regression by random forest. R News. 2002;2:18-22.
51. Amirkhani H, Rahmati M, Lucas PJF, Hommersom A. Exploiting experts' knowledge for structure learning of Bayesian networks. IEEE Trans Pattern Anal Mach Intell 2017;39:2154-70.
52. Zhang S, Tjortjis C, Zeng X, Qiao H, Buchan I, Keane J. Comparing data mining methods with logistic regression in childhood obesity prediction. Inf Syst Front 2009;11:449-60.
53. Zhang Q, Deng D, Dai W, Li J, Jin X. Optimization of culture conditions for differentiation of melon based on artificial neural network and genetic algorithm. Sci Rep 2020;10:3524.
54. Cho YB, Farrokhkish M, Norrlinger B, Heaton R, Jaffray D, Islam M. An artificial neural network to model response of a radiotherapy beam monitoring system. Med Phys 2020;47:1983-94.
55. Baitharu TR, Pani SK. Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. Procedia Comput Sci 2016;85:862-70.
56. Maliha SK, Islam T, Ghosh SK, Ahmed H, Mollick MR, Ema RR. Prediction of cancer using logistic regression, K-Star and J48 algorithm. 2019 4th International Conference on Electrical Information and Communication Technology, EICT 2019; 2019.
57. Wiharto W, Kusnanto H, Herianto H. Intelligence system for diagnosis level of coronary heart disease with K-star algorithm. Healthc Inform Res 2016;22:30-8.
58. Han J, Pei J, Kamber M. Data mining: Concepts and techniques. Amsterdam, Netherlands: Elsevier; 2011.
59. Saba AI, Elsheikh AH. Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. Process Saf Environ Prot 2020;141:1-8.
60. Chretien JP, George D, Shaman J, Chitale RA, McKenzie FE. Influenza forecasting in human populations: A scoping review. PLoS One 2014;9:e94130.
61. Afshar S, Afshar S, Warden E, Manochehri H, Saidijam M. Application of Artificial Neural Network in miRNA Biomarker Selection and Precise Diagnosis of Colorectal Cancer Iran Biomed J 2019;23:175-83.
62. Lawson AB. Statistical methods in spatial epidemiology. Hoboken, New Jersey: John Wiley & Sons; 2013.
63. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals 2020;139:110059.
64. Oliveira BA, Oliveira LC, Sabino EC, Okay TS. SARS-CoV-2 and the COVID-19 disease: A mini review on diagnostic methods. Rev Inst Med Trop Sao Paulo 2020;62:e44.
65. Mahmood A, Gajula C, Gajula P. Clinical and diagnostic criteria of COVID 19; a study of 4659 patients evaluating diagnostic testing and establishing an algorithm. J Med Surg Sci 2020;2:2.
66. Rehm GB, Woo SH, Chen XL, Kuhn BT, Cortes-Puch I, Anderson NR, *et al*. Leveraging IoTs and Machine learning for patient diagnosis and ventilation management in the intensive care unit. IEEE Pervasive Comput 2020;19:68-78.
67. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. arXiv preprint arXiv2020;3:200310849.
68. Yasar H, Ceylan M. A novel comparative study for detection of Covid-19 on CT lung images using texture analysis, machine learning, and deep learning methods. Multimed Tools Appl 2020;79:1-25.
69. Moftakhar L, Mozhgan S, Safe MS. Exponentially increasing trend of infected patients with COVID-19 in Iran: A comparison of neural network and ARIMA forecasting models. Iran J Public Health 2020;49:92-100.
70. Alakus TB, Turkoglu I. Comparison of deep learning approaches to predict COVID-19 infection. Chaos Solitons Fractals 2020;140:110120. [Doi: 10.1371/journal.pone. 0236621].
71. Sedik A, Iliyasu AM, El-Rahiem A, Abdel Samea ME, Abdel-Raheem A, Hammad M, *et al*. Deploying machine and deep learning models for efficient data-augmented detection of covid-19 infections. Viruses 2020;12:769.