Prediction of Patients with COVID-19 Requiring Intensive Care: A Cross-sectional Study Based on Machine-learning Approach from Iran

Golnar Sabetian[1], Aram Azimi[2], Azar Kazemi[3], Benyamin Hoseini[4], Naeimehossadat Asmarian[5], Vahid Khaloo[6], Farid Zand[7], Mansoor Masjedi[8], Reza Shahriarirad[9], Sepehr Shahriarirad[10]

ABSTRACT

Background: Prioritizing the patients requiring intensive care may decrease the fatality of coronavirus disease-2019 (COVID-19).

Aims and objectives: To develop, validate, and compare two models based on machine-learning methods for predicting patients with COVID-19 requiring intensive care.

Materials and methods: In 2021, 506 suspected COVID-19 patients, with clinical presentations along with radiographic findings, were laboratory confirmed and included in the study. The primary end-point was patients with COVID-19 requiring intensive care, defined as actual admission to the intensive care unit (ICU). The data were randomly partitioned into training and testing sets (70% and 30%, respectively) without overlapping. A decision-tree algorithm and multivariate logistic regression were performed to develop the models for predicting the cases based on their first 24 hours data. The predictive performance of the models was compared based on the area under the receiver operating characteristic curve (AUC), sensitivity, and accuracy of the models.

Results: A 10-fold cross-validation decision-tree model predicted cases requiring intensive care with the AUC, accuracy, and sensitivity of 97%, 98%, and 94.74%, respectively. The same values in the machine-learning logistic regression model were 75%, 85.62%, and 55.26%, respectively. Creatinine, smoking, neutrophil/lymphocyte ratio, temperature, respiratory rate, partial thromboplastin time, white blood cell, Glasgow Coma Scale (GCS), dizziness, international normalized ratio, O_2 saturation, C-reactive protein, diastolic blood pressure (DBP), and dry cough were the most important predictors.

Conclusion: In an Iranian population, our decision-based machine-learning method offered an advantage over logistic regression for predicting patients requiring intensive care. This method can support clinicians in decision-making, using patients' early data, particularly in low- and middle-income countries where their resources are as limited as Iran.

Keywords: COVID-19, Intensive care, Iran, Machine-learning, Prediction, Regression.

Indian Journal of Critical Care Medicine (2022): 10.5005/jp-journals-10071-24226

INTRODUCTION

Coronavirus disease-2019 (COVID-19), a widespread and ongoing global acute viral respiratory infection, has been the third coronavirus in the human population in the past two decades.[1] At present, the diagnosis of COVID-19 infection is confirmed by nucleic acid test of reverse transcription-polymerase chain reaction (RT-PCR).[1,2] Chest CT had higher sensitivity for COVID-19 diagnosis (about 98%) than initial RT-PCR with a reported positive rate of 32–60% for throat swab samples.[3,4]

The overwhelming burden of patients with COVID-19 and limited ICU capacities emphasize the importance of effective patient triage and identifying patients at increased risk for severe illness.[5] The early identification of patients who need to be admitted to the ICU is crucial to manage the patients with COVID-19 and reduce morbidity and mortality.[6] The need for ICU treatments and invasive mechanical ventilation were reported up to 33% for COVID-19 patients.[7] This high demand for intensive care is not balanced with the number of beds available and other resources in the ICU[2,8] that may increase the fatality of COVID-19 patients.[2] Prioritizing the patients requiring intensive care may decrease this fatality.[8] Machine-learning approaches can offer valuable recommendations to identify those requiring intensive care.[9,10] Machine-learning is a branch

[1,2,7,8]Shiraz University of Medical Sciences, Anesthesiology and Critical Care Research Center, Shiraz, Iran

[3]Department of Biomedical Informatics, Mashhad University of Medical Sciences, Mashhad, Iran

[4]Mashhad University of Medical Sciences, Pharmaceutical Research Center, Mashhad, Razavi Khorasan Province, Iran

[5,6]Shiraz University of Medical Sciences, Aliasghar Hospital, Shiraz, Iran

[9]Shiraz University of Medical Sciences, Thoracic and Vascular Surgery Research Center, Shiraz, Iran

[10]Shiraz University of Medical Sciences, Student Research Committee, Shiraz, Iran

Corresponding Authors: Aram Azimi, Shiraz University of Medical Sciences, Anesthesiology and Critical Care Research Center, Shiraz, Iran, e-mail: aramazimi@yahoo.com; Azar Kazemi, Department of Biomedical Informatics, Mashhad University of Medical Sciences, Mashhad, Iran, email: azarkazemi.mi@gmail.com.

of artificial intelligence and computer science, focusing on using data and algorithms to imitate how humans learn, gradually improving its accuracy.[11]

Modeling the ICU transfer using available para-clinic data and recorded signs and symptoms during the first 24 hours of hospitalization support the clinician's decision-making for early prediction of severe and critical conditions.[5] There are previous efforts to predict ICU transfer of COVID-19 patients.[11–13] Although these studies' results are beneficial to identify advanced patients' conditions, identifying the need for ICU transfer during the first hours of hospitalization had not shown excellent performance. In the study by Surme et al., demographics, symptoms, and para-clinic features gathered within 24 hours after hospitalization were used to predict ICU admission using multivariate logistic regression.[12] In another study by Cheng et al., machine-learning-based modeling was done on vital signs, laboratory data, and other para-clinic data.[11] Sensitivity was about 70% in both the studies, which meant 30% of patients who might need ICU treatments would not be identified by the models.

The current work emphasizes to develop a well-performed machine-learning model based on the symptom, vital signs, demographic, and available laboratory values during the first 24 hours of hospitalization to prioritize patients requiring intensive care for COVID-19.

METHOD

This cross-sectional study was performed on all hospitalized cases of COVID-19 between March 5 and May 20, 2021, in two centers affiliated with Shiraz University of Medical Sciences, Shiraz, Iran. The data were collected by a trained team of medical students who were not involved in the direct care of patients. To ensure the confidentiality of the patients' data, all the identifiers' data were excluded and patients were named by a researcher-made code.

Patients aged ≥18 years with respiratory symptoms were assessed by RT-PCR test using throat and nose swab specimens and lung imaging incredibly high-resolution CT scans. The images were reported by a specialized radiologist who was blind to the laboratory and clinical diagnosis, management, and outcome. Finally, the patients admitted in the hospital with respiratory complaints were considered COVID-19 infected based on positive RT-PCR testing and were included in the study. Confirmed cases classified as critical [respiratory failure, shock, and multiple organ dysfunction or as severe (dyspnea, respiratory rate ≥30 breaths/min, oxygen saturation ≤93%), and partial pressure of arterial oxygen to fraction of inspired oxygen (PaO_2/FiO_2) ratio <300 mm Hg] were a candidate for transferring to ICU. Owing to capacity limitations, not all critical and severe cases were admitted to the ICU.

The study variables were categorized as demographic data (gender, age), comorbidities, hypertension, cardiovascular disease including ischemic heart disease and congestive heart failure, chronic kidney disease (CKD), diabetes, cancer, chronic obstructive pulmonary disease (COPD), symptoms, vital signs, and laboratory data at the time of admission and outcome (admission to the ICU) (Table 1).

Modeling

We aimed to identify predictors that affect the ICU requirement of the COVID-19 patients. We considered the possibility of the missing data, noises, and outliers that should be detected and treated. On the other hand, more data were related to patients who were

not transferred to ICU (ICU-) and fewer portions to those who were transferred to ICU (ICU+). Therefore, preprocessing should be performed as an essential phase in the study that plays a vital role in meeting the study's objective. All data processes were performed by one of the authors who was not involved in the treatment process.

Preprocessing

The gathered data from 599 patients were integrated and the duplicated records were removed. Data related to 506 confirmed cases were analyzed. We omitted variables that had more than 50% missing values. All continuous, categorical, and ordinal missing variables were imputed by mean, mode, and median values, respectively. Outlier values in continuous variables were replaced with cutoff values. The outlier cutoff value was considered three standard deviations above or below the variable mean. Finally, data were double-checked by clinicians.

The features were reduced to improve the analysis process. Feature selection was made to elect a subset of more informative and relevant predictors. Continuous variables were screened with a threshold of 0.1 coefficient of variation. Threshold of 0.1 is the minimum coefficient of variation, which meant that for each continuous feature, the coefficient of variation is equal or less than 0.1, there is not much variability in values of feature, and therefore the feature would be excluded. Categorical variables were ranked based on the importance value to find out which variables are more related to the outcome. Importance values under 0.9 were also excluded based on the Pearson or *F* statistics (Supplementary File 1 shows the excluded variables in this step).

Machine-learning Analysis

The data were randomly partitioned into training and testing sets (70 and 30% of total data, respectively) without overlapping. An initial tree was pruned to avoid overfitting, help trees deploy well, and help users decide with interpretable results.[14] C5.0 decision-tree algorithm was applied to the dataset to predict the outcome. In this regard, data were split, each sub-data were split again based on different fields, and this process continued until splitting cannot be done further. Variables that are not informative to contribute to making the tree are removed or pruned. Pruning severity, minimum records per child branch, and type of pruning were adjusted, and using k-fold cross-validation several decision trees were modeled and their performances were compared to acquire proper performance, which predicted the clinically significant ones. Machine-learning multivariate logistic regression (MLMLR) was applied as a conventional technique to evaluate the C5.0 in prediction. In this regard, univariate logistic regression was applied to all features to select variables that were significantly associated with ICU transfer. Then, MLMLR model was developed with binomial backward stepwise. Algorithms were applied to the training set to develop the prediction model, then the model was evaluated using the testing set. The machine-learning modeling was made in IBM SPSS Modeler version 18.

Statistical Analysis

Misclassification costs for models were calculated. The coincidence matrix, performance evaluation, AUC, and confidence value were

Table 1: Characteristics of the study confirmed COVID-19 disease patients

Variable	Total	ICU+ N = 100 (19.76)	ICU- N = 406 (80.24)	p value
Age (years)	52.97 ± 18.17	59.07 ± 18.35	51.459 ± 17.827	<0.001*
Gender				
Male	301 (60.080)	67 (67.677)	234 (58.209)	0.844
Female	200 (39.920)	32 (32.323)	168 (41.791)	
O_2 saturation (%)	91.428 ± 6.537	87.468 ± 9.887	92.471 ± 4.822	<0.001*
Temperature (°C)	37.026 ± 0.781	37.231 ± 0.734	36.973 ± 0.785	0.002*
GCS				<0.001*
7	2 (0.517)	2 (3.279)	0	
8	1 (0.258)	1 (1.639)	0	
10	3 (0.775)	3 (4.918)	0	
11	2 (0.517)	2 (3.279)	0	
14	2 (0.517)	0	2 (0.613)	
15	377 (97.416)	53 (86.885)	324 (99.387)	
SBP (mm Hg)	124.577 ± 17.442	122.211 ± 19.761	125.188 ± 16.767	0.106
DBP (mm Hg)	77.909 ± 12.160	71.731 ± 11.633	79.466 ± 11.804	<0.001*
RR	19.247 ± 3.652	20.300 ± 5.887	18.989 ± 2.801	0.260
HR	93.418 ± 16.100	93.677 ± 17.751	93.353 ± 15.685	0.937
Smoking				
Nonsmoker	471 (93.083)	80 (80)	391 (96.305)	<0.001*
Cigarette	13 (2.569)	11 (11)	2 (0.493)	
Opium abuse	22 (4.348)	9 (9)	13 (3.202)	
Comorbid disease (Yes)	221 (43.676)	62 (62)	159 (39.163)	<0.001*
Diabetes	113 (22.332)	34 (34)	79 (19.458)	
Cardiovascular	170 (33.597)	46 (46)	124 (30.542)	
CKD	46 (9.091)	29 (29)	17 (4.187)	
Diarrhea (Yes)	17 (3.360)	4 (4)	13 (3.202)	0.692
Nausea vomiting (Yes)	34 (6.719)	6 (6)	28 (6.897)	0.748
Chest pain (Yes)	23 (4.545)	5 (1)	18 (4.433)	0.808
Sputum (Yes)	55 (10.870)	10 (10)	45 (11.084)	0.755
Headache (Yes)	74 (14.625)	14 (14)	60 (14.778)	0.844
Dizziness (Yes)	20 (3.953)	8 (8)	12 (2.956)	0.039*
Myalgia (Yes)	168 (33.202)	31 (31)	137 (33.744)	0.602
Fatigue (Yes)	129 (25.494)	26 (26)	103 (25.369)	0.897
Dry cough (Yes)	309 (61.069)	53 (53)	256 (63.054)	0.065
Dyspnea	333 (65.810)	71 (71)	262 (64.532)	0.222
Rhinorrhea	5 (0.988)	1 (1)	4 (0.985)	1
Sore throat	30 (5.929)	5 (5)	25 (6.158)	0.661
Chills	86 (16.996)	21 (21)	65 (16.010)	0.234
WBC ($\times10^9$/L)	7.583 ± 7.598	11.499 ± 14.971	6.521 ± 2.792	<0.001*
NLR	5.347 ± 7.6	10.364 ± 13.643	3.944 ± 3.659	<0.001*
INR	1.474 ± 0.571	1.741 ± 0.794	1.388 ± 0.446	<0.001*
PLT ($\times10^9$/L)	235.144 ± 97.626	226.121 ± 117.496	237.573 ± 91.595	0.096
PT (sec)	16.397 ± 3.832	18.464 ± 5.657	15.735 ± 2.724	<0.001*
PTT (sec)	40.476 ± 15.865	46.355 ± 21.053	38.597 ± 13.312	<0.001*
BUN (mg/dL)	18.705 ± 17.942	32.114 ± 30.189	15.149 ± 10.352	<0.001*
Creatinine (mg/dL)	1.524 ± 1.812	2.359 ± 2.254	1.301 ± 1.606	<0.001*
FBS (mg/dL)	125.487 ± 60.126	143.493 ± 67.635	120.951 ± 57.334	<0.001*
HCT (%)	39.180 ± 6.633	35.333 ± 6.884	40.203 ± 6.183	<0.001*
Hb (gm/dL)	13.200 ± 2.184	12.004 ± 2.391	13.517 ± 2.013	<0.001*
Na (mEq/L)	138.557 ± 4.996	136.531 ± 5.949	139.115 ± 4.554	<0.001*

K (mEq/L)	4.148 ± 0.579	4.273 ± 0.751	4.115 ± 0.519	0.245
CRP (mg/L)	30.863 ± 28.139	37.468 ± 34.562	28.909 ± 25.689	0.054
ALT (IU/L)	53.927 ± 221.680	113.030 ± 465.291	37.185 ± 34.505	0.006*
AST (IU/L)	40.766 ± 37.186	66.091 ± 60.619	33.713 ± 22.979	<0.001*
ALP (IU/L)	230.088 ± 113.820	237.661 ± 103.547	227.841 ± 116.832	0.450
Total bilirubin (mg/dL)	1.036 ± 1.188	1.461 ± 2.59	0.925 ± 0.789	0.005*
pH	7.384 ± 0.067	7.364 ± 0.116	7.388 ± 0.052	0.739
PaCO$_2$	41.418 ± 7.785	42.183 ± 11.659	41.263 ± 6.757	0.700
HCO$_3$	24.468 ± 4.060	24.220 ± 6.150	24.519 ± 3.501	0.683

ALT, alanine aminotransferase; ALP, alkaline phosphatase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; CKD, chronic kidney disease; CRP, C-reactive protein; Cr, creatinine; DBP, diastolic blood pressure; FBS, fasting blood sugar; GCS, Glasgow Coma Scale; Hb, hemoglobin; HCO$_3$, bicarbonate; HCT, hematocrit; HR, heart rate; ICU+, patients who needed ICU; ICU-, patients who didn't need ICU; INR, international normalized ratio; K, potassium; LDH, lactate dehydrogenase; Na, sodium; NLR, neutrophil/lymphocyte ratio; PaCO$_2$, partial pressure of carbon dioxide; PCT, procalcitonin; PT, prothrombin time; PTT, partial thromboplastin time; RR, respiratory rate; SBP, systolic blood pressure; WBC, white blood cell; *was considered statistically significant

estimated for each model based on the testing set. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and *F* measure were reported for the best models of C5.0 and were compared with MLMLR. Area under the receiver operating characteristic curve (AUC) comparison was conducted between C5.0 and MLMLR based on DeLong et al. methodology by MedCalc version 14.8.1.[15] Descriptive analysis, univariate logistic regression, and other statistical analyzes were performed by R version 4.0.3. Categorical and continuous characteristics of patients were summarized as frequency and mean ± SD. Pearson Chi-square test or Fisher's exact test and *t*-test or Mann–Whitney *U* tests were used for comparing categorical and continuous characteristics in ICU+ and ICU- groups, respectively. Kruskal–Wallis also was used for comparing an ordinal feature in both the groups. A *p*-value of less than 0.05 was considered statistically significant.

RESULTS

Preprocessing

After deletion of the missing variables, three variables had 50–59.99% missing values, four variables had 60–69.99% missing values, six variables had 70–79.99% missing values, three variables had 80–89.99% missing values, 12 variables had 90–99.99% missing values, and 15 variables had no missing values.

After excluding features with more than 50% missing values, the rest variables were imputed, and then the feature selection process was done before modeling. Among all variables entered into the model, five variables had the range of 59–70% missing values at first and five other variables had the range of 75–90%, but most variables had less than 10% or without missing values.

Patients' Characteristics

The characteristics of patients before preprocessing are shown in Table 1. Study population included 301 (60.08%) male and 200 (39.92%) female with mean ± SD age of 52.97 ± 18.17 years. One hundred (19.76%) patients were transferred to the ICU during the study. Cases in the ICU+ group were older than those in the ICU- group (59.07 ± 18.35 years vs 51.46 ± 17.83 years; *p* <0.001), with more proportion of males to females in the ICU+ group [67(67.68%) males vs 32 (32.32%) females]. Patients who smoked in the ICU+ group were more frequent than the patients who smoked in the ICU- group [20(20%) vs 5(3.695%); *p* <0.001].

A total of 221 (43.68%) patients suffered from one or more comorbid diseases, including cardiovascular diseases (33.597%), diabetes (22.323%), and/or CKD (9.091%). Among COVID-19 patients', cardiovascular diseases were more frequent than other comorbid diseases, but the difference between ICU+ and ICU- groups was more significant in CKD than other diseases. Patients in the ICU+ group had a lower GCS (*p* <0.001), O$_2$ saturation (*p* <0.001), DBP (*p* <0.001), and also higher temperature (*p* = 0.002) compared to the ICU- patients. Dizziness was more common in the ICU+ group (*p* = 0.039). Patients in the ICU+ group had higher white blood cells count (WBC) (*p* <0.001).

The results of univariate logistic regression in Table 2 indicate that age (*p* <0.001), being a current smoker (*p* <0.001), and suffering from comorbidities (*p* <0.001) were significantly associated with ICU transfer. Among vital signs, O$_2$ saturation (*p* <0.001), temperature (*p* = 0.004), GCS (*p* = 0.03), DBP (*p* <0.001), and respiratory rate (*p* = 0.006) were significantly associated with ICU need.

Laboratory markers, such as WBC (*p* <0.001), NLR (*p* <0.001), INR (*p* <0.001), PT (*p* <0.001), PTT (*p* <0.001), BUN (*p* = 0.008), Cr (*p* = 0.001), FBS (*p* <0.001), HCT (*p* <0.001), Hb (*p* <0.001), Na (*p* <0.001), K (*p* = 0.020), CRP (*p* = 0.020), alanine aminotransferase ALT (*p* = 0.006), aspartate aminotransferase AST (*p* <0.001), total bilirubin (*p* = 0.007), and pH (*p* = 0.013) were significantly associated with ICU need (Table 2).

Models' Performance

The 10-fold cross-validation C5.0 model sensitivity was 94.74%, with an accuracy of 98%, and an AUC of 0.969 (with 95% CI: 0.928–0.990), which was the best C5.0 model (Table 3). The sensitivity and accuracy of the MLMLR model were 55.26% and 85.62%, respectively, and AUC was 0.755 (with 95% CI: 0.679–0.820). The results revealed that the C5.0 decision-tree model performed significantly better regarding the discrimination of COVID-19 patients to predict the need for ICU compared with the MLMLR model (difference between Area: 0.215; 95% CI: 0.125–0.304; *p* <0.0001).

Models' Predictors

The results of decision-tree-based model indicated that most important predictive variables were creatinine, smoking, NLR, temperature, RR, PTT, WBC, GCS, dizziness, INR, O$_2$ saturation, CRP, DBP, and dry cough (Table 4). The predictive variables identified in MLMLR were being current cigarette smoking (*p* <0.001); opium

Table 2: Predictive variables of univariate logistic regression for prediction of COVID-19 patients' ICU need

Variable	OR	CI	p value
Age (years)	1.024	1.011–1.037	<0.001[*]
Gender			
Male	0.665	0.414–1.052	0.086
Female			
O_2 saturation (%)	0.899	0.864–0.932	<0.001[*]
Temperature (°C)	1.522	1.149–2.023	0.004[*]
GCS	0.308	0.048–0.587	0.030[*]
SBP (mm Hg)	1.003	0.976–1.003	0.139
DBP (mm Hg)	0.944	0.923–0.964	<0.001[*]
RR	1.088	1.027–1.160	0.006[*]
HR	1.001	0.987–1.015	0.862
Smoking (Nonsmoker)	2.424	3.708–3.708	<0.001[*]
Comorbid diseases (Yes)	3.0945	1.656–6.031	<0.001[*]
Diarrhea (Yes)	1.260	0.349–3.650	0.692
Nausea vomiting (Yes)	0.862	0.315–2.008	0.749
Chest pain (Yes)	1.135	0.367–2.926	0.808
Sputum (Yes)	0.891	0.411–1.769	0.755
Headache (Yes)	0.939	0.484–1.715	0.844
Dizziness (Yes)	2.855	1.091–7.104	0.026[*]
Myalgia (Yes)	0.882	0.545–1.402	0.602
Fatigue (Yes)	1.034	0.619–1.686	0.897
Dry cough (Yes)	0.661	0.425 –1.029	0.066
Dyspnea	1.346	0.842–2.195	0.223
Rhinorrhea	1.015	0.052–6.956	0.989
Sore throat	0.802	0.265–1.988	0.661
Chills	1.396	0.791–2.384	0.236
WBC ($\times 10^9$/L)	1.142	1.080–1.218	<0.001[*]
NLR	1.161	1.097–1.240	<0.001[*]
INR	2.687	1.772–4.333	<0.001[*]
PLT ($\times 10^9$/L)	0.999	0.996–1.001	0.321
PT (sec)	1.192	1.118–1.281	<0.001[*]
PTT (sec)	1.027	1.013–1.043	<0.001[*]
BUN (mg/dL)	1.053	1.037–1.073	<0.001[*]
Cr (mg/dL)	1.322	1.165–1.530	0.001[*]
FBS (mg/dL)	1.005	1.001–1.009	0.008[*]
HCT (%)	0.894	0.861–0.926	<0.001[*]
Hb (gm/dL)	0.725	0.648–0.807	<0.001[*]
Na (mEq/L)	0.889	0.844–0.935	<0.001[*]
K (mEq/L)	1.582	1.075–2.333	0.020[*]
CRP (mg/L)	1.010	1.002–1.019	0.020[*]
ALT (IU/L)	1.009	1.003–1.015	0.006[*]
AST (IU/L)	1.025	1.016–1.036	<0.001[*]
ALP (IU/L)	1.001	0.998–1.003	0.552
Total bilirubin (mg/dL)	1.381	1.119–1.790	0.007[*]
pH	0.008	0.000–0.368	0.013[*]
$PaCO_2$	1.015	0.979–1.049	0.404
HCO_3	0.982	0.915–1.051	0.606

ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; CI, confidence interval; CRP, C-reactive protein; Cr, creatinine; DBP, diastolic blood pressure; FBS, fasting blood sugar; GCS, Glasgow Coma Scale; Hb, hemoglobin; HCO_3, bicarbonate; HCT, hematocrit; HR, heart rate; ICU+, patients who needed ICU; ICU-, patients who didn't need ICU; INR, international normalized ratio; K, potassium; LDH, lactate dehydrogenase; Na, sodium; NLR, neutrophil/lymphocyte ratio; OR, odds ratio; $PaCO_2$, partial pressure of carbon dioxide; PCT, procalcitonin; PT, prothrombin time; PTT, partial thromboplastin time; RR, respiratory rate; SBP, systolic blood pressure; WBC, white blood cell; [*] was considered statistically significant

Table 3: Performance comparison of C5.0 and machine-learning multivariate logistic regression models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	F measure (%)	PPV (%)	NPV (%)
C5.0	98.04	94.74	99.13	0.969	96	97.30	98.28
MLMLR	85.62	55.26	95.65	0.755	65.63	80.77	86.61

AUC, area under the receiver operating characteristic; MLMLR, machine-learning multivariate logistic regression; NPV, negative predictive value; PPV, positive predictive value

Table 4: Predictive variables importance of C5.0 decision-tree model for prediction of COVID-19 patients' ICU requirement

Predictor	Importance
Creatinine (mg/dL)	0.3
Smoking status	0.12
Neutrophil/lymphocyte ratio	0.1
Temperature (°C)	0.1
Respiratory rate	0.05
Partial thromboplastin time (sec)	0.05
White blood cell ($\times 10^9$/L)	0.05
Glasgow Coma Scale	0.05
Dizziness	0.04
International normalized ratio	0.03
O_2 saturation (%)	0.03
C-reactive protein (mg/L)	0.03
Diastolic blood pressure (mm Hg)	0.02
Dry cough	0.01

Table 5: Predictive variables of machine-learning multivariate logistic regression model for prediction of ICU transfer

Predictor	OR	CI	p value
Smoking			<0.001*
Cigarette smoker	70.674	10.628–469.953	<0.001
Opium abuse	4.347	1.115–16.939	0.034
Diastolic blood pressure (mm Hg)	0.935	0.909–0.963	<0.001*
Creatinine (mg/dL)	2.931	1.870–4.593	<0.001*
Fasting blood sugar (mg/dL)	1.008	1.001–1.015	0.023*
Hematocrit (%)	0.951	0.905–0.999	0.048*
Aspartate aminotransferase (IU/L)	1.040	1.016–1.065	0.001*
Neutrophil/lymphocyte ratio	1.249	1.102–1.416	0.001*

CI, confidence interval; OR, odds ratio; *was considered statistically significant

abuse (p=0.034), DBP (p <0.001), Cr (p <0.001), FBS (p=0.023), HCT (p=0.048), AST (p=0.001), and NLR (p=0.043) (Table 5).

DISCUSSION

To our knowledge, this is the first study in Iran to develop a model based on the machine-learning methods to predict patients requiring intensive care for COVID-19. C5.0 develops models in tree structure forms constructed using a heuristic partitioning, and, in this, recursive portioning, division, and conquering are done to show features values in smaller subclasses.[14] The number of patients' features indicated high dimensions of this dataset. In this condition, reducing the error rate, gaining high sensitivity, and overcoming overfitting in unequal conditions are the main challenges in the study. Decision trees work well in high-dimensional datasets to deal with overfitting.[16] In cases where the sensitivity is crucial, C5.0 performs better than other decision trees algorithms.[17] However, C5.0 is a well-performed machine-learning algorithm that has been applied to various real data sets in communicable and non-communicable diseases[14,18,19] to predict the outcome.

The study revealed that the decision-tree-based model identifies the patients who need ICU more accurately than the logistic regression model. The predictive features selected in the machine-learning model can be well summarized as Cr, being a current smoker, NLR, body temperature, respiratory distress or tachypnea, PTT, WBC, GCS, dizziness, INR, O_2 saturation, CRP, DBP, and dry cough that are in line with previous reports.[5,11,20] Among predictors, Cr has the most robust predictive value. Dry cough reported more prevalent in severe conditions[21] may be a sign of progressive lung involvement or heart failure.[22]

We also found that ICU+ patients were significantly older and male patients were more frequent, as emphasized by the others.[23] The majority of smokers had been transferred to ICU. Cigarette smoking could increase the risk and severity of pulmonary viral infections because of a decrease in pulmonary immune function[24] and exaggerated immune-inflammatory responses[25] associated with the severity and survival of COVID-19 infection. Smoking had been reported as the MERS risk factor previously.[26] Patients with underlying health diseases like CKD, diabetes mellitus, and cardiovascular disease, including hypertension or ischemic heart disease, were at higher risk for progression to critical condition and transferred to ICU.

Vital signs and symptoms, which were the most decisional factors for hospitalization, could be significant in ICU transfer prediction, as we found that ICU+ patients had lower O_2 saturation and DBP and higher respiratory rate and temperature at the time of admission. Dizziness was reported as a special nonspecific presentation among COVID-19 patients. High prevalence of dizziness among ICU+ patients recorded in our triage setting may reveal it's importance. Dizziness has not been mentioned in previous studies, but it seems crucial to determine its leading causes.[17] Further studies are needed to prove the importance of dizziness as a main decisional factor in ICU transfer prediction.

In addition, ICU+ patients had a higher WBC count, NLR, INR, PT, PTT, BUN, Cr, and FBS, and lower HCT, Hb, and Na compared to ICU- group patients. These results are matched with previously reported factors for poor outcomes and severe conditions.[4,27,28] COVID-19 infection may progress to severe conditions, such as acute respiratory distress syndrome and multi-organ dysfunction syndrome,[4,29] which are associated with hypercoagulation and disseminated intravascular coagulation.[30]

Patients with COVID-19 may develop kidney damage by direct viral injury, inflammation, or poor clinical outcomes.[31] Hematologic impairment such as increased neutrophils percent, decreased lymphocytes percent, and increased NLR levels

reflecting an enhanced inflammatory process and impaired immune cell function may help predict the severity of clinical outcomes.[32,33] The relationship between pulmonary infections and inappropriate antidiuretic hormone secretion syndrome, resulting in hyponatremia, was previously highlighted.[34] It was an early poor prognostic, predictive factor, probably due to extensive infective lung involvement.

The predictors found in the current study are mostly coherent with previously reported factors associated with the severity of disease.[1,2,4,27] However, we aimed to use machine learning for risk stratification for prediction goals because it has been shown that machine learning has excellent reliability and optimal performance for the assessment of the relationship between data from different domains and outputs.[11,13,19] The model developed in this study was a better-performed classifier to classify ICU- and ICU+ compared to previous studies.[11,12]

The novelty of this study is that it is applied initially at admission time instead of hospital course, supposing that earlier identification of high-risk patients could help in reducing patients' mortality.[35] By accessing the information of 506 confirmed COVID-19 patients, we described some differences between patients who needed and those who did not need ICU transfer. Machine learning could work as a tool for organizing relationships between features and the outcome, especially when the features are extensive in number, nonlinear, and complex.[11] The current study results show that in an appropriate setting, the C5.0 model is a more accurate and functional method than logistic regression for predicting critical conditions in patients with COVID-19.

There are some limitations to the study. First, the data were recorded during the highest surge of the disease, which may affect the recordings because the significant strain was treatment. For example, the patients who needed ICU admission but not transfer were not available. Two clinicians did quality control of data, and preprocessing steps burned time. Second, some important variables such as D-dimer, procalcitonin, PaO_2, PaO_2/FiO_2 ratio, and type of oxygen therapy before transfer to ICU were omitted due to the high missing value rate. In all, the related early data could provide a more comprehensive view of ICU need to improve the understanding of risk factors. Finally, it should be considered that there were limitations for ICU transfer during the overwhelming hospital capacity of the COVID-19 pandemic. The limitations mentioned imposing a lack of generalizability of the findings. However, we still think that the results and methods used will be helpful in low- and middle-income countries where their resources are as limited as Iran.

CONCLUSION

The machine-learning prediction-supporting method may be used for the accurate assessment of the need for ICU admission of COVID-19 patients at the early stage for optimizing patients' triage and allocation of facilities and accurate prediction of severe cases for better management, especially in situations of shortage of medical resources. We encourage the prospective validation of these results in a clinical setting.

ORCID

Golnar Sabetian https://orcid.org/0000-0001-8764-2150
Aram Azimi https://orcid.org/0000-0003-2730-5552
Azar Kazemi https://orcid.org/0000-0003-3197-0491
Benyamin Hoseini https://orcid.org/0000-0002-0355-6181
Naeimehossadat Asmarian https://orcid.org/0000-0001-6091-8890
Vahid khaloo https://orcid.org/0000-0003-1932-0592
Farid Zand https://orcid.org/0000-0003-3489-3372
Mansoor Masjedi https://orcid.org/0000-0001-6175-9289
Reza Shahriarirad https://orcid.org/0000-0001-5454-495X
Sepehr Shahriarirad https://orcid.org/0000-0002-9567-5996

REFERENCES

1. Ghale-Noie ZN, Salmaninejad A, Bergquist R, Mollazadeh S, Hoseini B, Sahebkar A. Genetic aspects and immune responses in Covid-19: Important Organ Involvement. Adv Exp Med Biol 2021;1327:3–22. Epub 2021/07/20. DOI: 10.1007/978-3-030-71697-4_1.
2. Goshayeshi L, Akbari Rad M, Bergquist R, Allahyari A, Hashemzadeh K, Hoseini B. Demographic and clinical characteristics of severe Covid-19 infections: a cross-sectional study from Mashhad University of Medical Sciences, Iran. BMC Infect Dis 2021;21(1):656. DOI: 10.1186/s12879-021-06363-6.
3. Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, et al. Detection of SARS-CoV-2 in different types of clinical specimens. JAMA 2020;323(18):1843–1844. DOI: 10.1001/jama.2020.3786.
4. Goshayeshi L, Milani N, Bergquist R, Sadrzadeh SM, Rajabzadeh F, Hoseini B. Covid-19 presented only with gastrointestinal symptoms: a case report of a 14-year-old patient. Govaresh 2021;25(4):300–304.
5. Heo J, Han D, Kim HJ, Kim D, Lee YK, Lim D, et al. Prediction of patients requiring intensive care for COVID-19: development and validation of an integer-based score using data from Centers for Disease Control and Prevention of South Korea. J Intensive Care 2021;9(1):16. DOI: 10.1186/s40560-021-00527-x.
6. Khoshrounejad F, Hamednia M, Mehrjerd A, Pichaghsaz S, Jamalirad H, Sargolzaei M, et al. Telehealth-based services during the COVID-19 Pandemic: A systematic review of features and challenges. Front Public Health 2021;9:711762. Epub 2021/08/06. DOI: 10.3389/fpubh.2021.711762.
7. Goyal P, Choi JJ, Pinheiro LC, Schenck EJ, Chen R, Jabri A, et al. Clinical characteristics of Covid-19 in New York City. N Engl J Med 2020;382(24):2372–2374. DOI: 10.1056/NEJMc2010419.
8. Rahmatinejad Z, Rahmatinejad F, Sezavar M, Tohidinezhad F, Abu-Hanna A, Eslami S. Internal validation and evaluation of the predictive performance of models based on the PRISM-3 (Pediatric Risk of Mortality) and PIM-3 (Pediatric Index of Mortality) scoring systems for predicting mortality in Pediatric Intensive Care Units (PICUs). BMC Pediatr 2022;22(1):199. Epub 2022/04/14. DOI: 10.1186/s12887-022-03228-y.
9. Rahmatinejad Z, Tohidinezhad F, Rahmatinejad F, Eslami S, Pourmand A, Abu-Hanna A, et al. Internal validation and comparison of the prognostic performance of models based on six emergency scoring systems to predict in-hospital mortality in the emergency department. BMC Emerg Med 2021;21(1):68. Epub 2021/06/12. DOI: 10.1186/s12873-021-00459-7.
10. Ahouz F, Golabpour A. Predicting the incidence of COVID-19 using data mining. BMC Public Health 2021;21(1):1087. DOI: 10.1186/s12889-021-11058-3.
11. Cheng FY, Joshi H, Tandon P, Freeman R, Reich DL, Mazumdar M, et al. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. J Clin Med 2020;9(6):1668. DOI: 10.3390/jcm9061668.
12. Surme S, Buyukyazgan A, Bayramlar OF, Cinar AK, Copur B, Zerdali E, et al. Predictors of intensive care unit admission or death in patients with coronavirus disease 2019 Pneumonia in Istanbul, Turkey. Jpn J Infect Dis 2021;74(5):458–464. DOI: 10.7883/yoken.JJID.2020.1065.
13. Cai W, Liu T, Xue X, Luo G, Wang X, Shen Y, et al. CT Quantification and machine-learning models for assessment of disease severity and

prognosis of COVID-19 patients. Acad Radiol 2020;27(12):1665–1678. DOI: 10.1016/j.acra.2020.09.004.

14. Saba T, Abunadi I, Shahzad MN, Khan AR. Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types. Microsc Res Tech 2021;84(7):1462–1474. DOI: 10.1002/jemt.23702.

15. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845. PMID: 3203132.

16. Roumani YF, May JH, Strum DP, Vargas LG. Classifying highly imbalanced ICU data. Health Care Manag Sci 2013;16(2):119–128. DOI: 10.1007/s10729-012-9216-9.

17. Saniasiaya J, Kulasegarah J. Dizziness and COVID-19. Ear Nose Throat J 2021;100(1):29–30. DOI: 10.1177/0145561320959573.

18. Peng J, Chen C, Zhou M, Xie X, Zhou Y, Luo CH. A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators. Sci Rep 2020;10(1):3118. DOI: 10.1038/s41598-020-60042-1.

19. Kazemi A, Kazemi K, Sami A, Sharifian R. Identifying factors that affect patient survival after orthotopic liver transplant using machine-learning techniques. Exp Clin Transplant 2019;17(6):775–783. DOI: 10.6002/ect.2018.0170.

20. Allenbach Y, Saadoun D, Maalouf G, Vieira M, Hellio A, Boddaert J, et al. Development of a multivariate prediction model of intensive care unit transfer or death: a French prospective cohort study of hospitalized COVID-19 patients. PLoS One 2020;15(10):e0240711. DOI: 10.1371/journal.pone.0240711.

21. Li J, Chen Z, Nie Y, Ma Y, Guo Q, Dai X. Identification of symptoms prognostic of COVID-19 severity: multivariate data analysis of a case series in Henan province. J Med Internet Res 2020;22(6):e19636. DOI: 10.2196/19636.

22. Muzzarelli S, Leibundgut G, Maeder MT, Rickli H, Handschin R, Gutmann M, et al. Predictors of early readmission or death in elderly patients with heart failure. Am Heart J 2010;160(2):308–314. DOI: 10.1016/j.ahj.2010.05.007.

23. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020;395(10229):1054–1062.

24. Bauer CMT, Morissette MC, Stampfli MR. The influence of cigarette smoking on viral infections: translating bench science to impact COPD pathogenesis and acute exacerbations of COPD clinically. Chest 2013;143(1):196–206. DOI: 10.1378/chest.12-0930.

25. McElvaney OJ, McEvoy NL, McElvaney OF, Carroll TP, Murphy MP, Dunlea DM, et al. Characterization of the Inflammatory response to severe COVID-19 illness. Am J Respir Crit Care Med 2020;202(6):812–821. DOI: 10.1164/rccm.202005-1583OC.

26. Park JE, Jung S, Kim A, Park JE. MERS transmission and risk factors: a systematic review. BMC Public Health 2018;18(1):574. DOI: 10.1186/s12889-018-5484-8.

27. Ou M, Zhu J, Ji P, Li H, Zhong Z, Li B, et al. Risk factors of severe cases with COVID-19: a meta-analysis. Epidemiol Infect 2020;148:e175. DOI: 10.1017/S095026882000179X.

28. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis 2020;20(6):669–677. DOI: 10.1016/S1473-3099(20)30243-7.

29. Wu T, Zuo Z, Kang S, Jiang L, Luo X, Xia Z, et al. Multi-organ dysfunction in patients with COVID-19: a systematic review and meta-analysis. Aging Dis 2020;11(4):874–894. DOI: 10.14336/AD.2020.0520.

30. Chang JC. Acute respiratory distress syndrome as an organ phenotype of vascular microthrombotic disease: based on hemostatic theory and endothelial molecular pathogenesis. Clin Appl Thromb Hemost 2019;25:1–20. DOI: 10.1177/1076029619887437.

31. Zheng X, Yang H, Li X, Li H, Xu L, Yu Q, et al. Prevalence of kidney injury and associations with critical illness and death in patients with COVID-19. Clin J Am Soc Nephrol 2020;15(11):1549–1556. DOI: 10.2215/CJN.04780420.

32. Lagunas-Rangel FA. Neutrophil-to-lymphocyte ratio and lymphocyte-to-C-reactive protein ratio in patients with severe coronavirus disease 2019 (COVID-19): a meta-analysis. J Med Virol 2020;92(10):1733–1734. DOI: 10.1002/jmv.25819.

33. Zhao Q, Meng M, Kumar R, Wu Y, Huang J, Deng Y, et al. Lymphopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: a systemic review and meta-analysis. Int J Infect Dis 2020;96:131–135. DOI: 10.1016/j.ijid.2020.04.086.

34. Aggarwal S, Garcia-Telles N, Aggarwal G, Lavie C, Lippi G, Henry BM. Clinical features, laboratory characteristics, and outcomes of patients hospitalized with coronavirus disease 2019 (COVID-19): early report from the United States. Diagnosis (Berl) 2020;7(2):91–96. DOI: 10.1515/dx-2020-0046.

35. Rahmatinejad Z, Hoseini B, Rahmatinejad F, Abu-Hanna A, Bergquist R, Pourmand A, et al. Internal validation of the predictive performance of models based on three ED and ICU scoring systems to predict inhospital mortality for intensive care patients referred from the Emergency Department. BioMed Research International 2022;2022:3964063. DOI: 10.1155/2022/3964063.