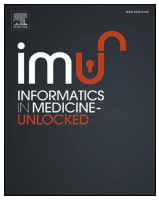
Informatics in Medicine Unlocked 30 (2022) 100919

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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



Using decision tree algorithms for estimating ICU admission of COVID-19 patients

Mostafa Shanbehzadeh [a], Raoof Nopour [b], Hadi Kazemi-Arpanahi [c,d,*]

[a] Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran
[b] Department of Health Information Management, Student Research Committee, School of Health Management and Information Sciences Branch, Iran University of Medical Sciences, Tehran, Iran
[c] Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran
[d] Department of Student Research Committee, Abadan University of Medical Sciences, Iran

ARTICLE INFO

Keywords:
COVID-19
Coronavirus
Machine learning
Intensive care unit
Decision tree

ABSTRACT

Introduction: Coronavirus disease 2019 (COVID-19) outbreak has overwhelmed many healthcare systems worldwide and put them at the edge of collapsing. As intensive care unit (ICU) capacities are limited, deciding on the proper allocation of required resources is crucial. This study aimed to develop and compare models for early predicting ICU admission in COVID-19 patients at the point of hospital admission.

Materials and methods: Using a single-center registry, we studied the records of 512 COVID-19 patients. First, the most important variables were identified using Chi-square test (at $p < 0.01$) and logistic regression (with odds ratio at $P < 0.05$). Second, we trained seven decision tree (DT) algorithms (decision stump (DS), Hoeffding tree (HT), LMT, J-48, random forest (RF), random tree (RT) and REP-Tree) using the selected variables. Finally, the models' performance was evaluated. Furthermore, we used an external dataset to validate the prediction models.

Results: Using the Chi-square test, 20 important variables were identified. Then, 12 variables were selected for model construction using logistic regression. Comparing the DT methods demonstrated that J-48 (F-score of 0.816 and AUC of 0.845) had the best performance. Also, the J-48 (F-score = 80.9% and AUC = 0.822) gained the best performance in generalizability using the external dataset.

Conclusions: The study results demonstrated that DT algorithms can be used to predict ICU admission requirements in COVID-19 patients based on the first time of admission data. Implementing such models has the potential to inform clinicians and managers to adopt the best policy and get prepare during the COVID-19 time-sensitive and resource-constrained situation.

1. Introduction

Coronavirus disease 2019 (COVID-19) is a life-threatening infection caused due to a recently originating zoonotic virus, named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The COVID-19 symptoms range from asymptomatic to mild or moderate symptoms such as fever, cough, shortness of breath, fatigue and other baseline clinical manifestations that start in the first week after infection [2,3]. Later, critical complications may develop in some patients including dyspnea, severe pneumonia and organ dysfunctions that need patients to be admitted to intensive care units (ICUs) [4]. Approximately 20% of COVID-19 patients must be hospitalized and almost 20–30% of in-hospital COVID-19 patients need to enter the ICU for urgent care [5]. In Iran, the ICU admission rate is estimated at 32% of hospitalized patients and the ICU death rate is about 39% [6]. Currently, the ICU resources are limited; generally, more than 50% of its beds are occupied under normal conditions [7].

The pandemic situation poses a great hazard to worldwide health and welfare. Despite all the preventive and lockdown measures to slow the spreading and contain the virus, the global healthcare systems have been stunned with high demands for hospital ICU resources such as personal protective equipment (PPE), ICU beds and medical ventilators [8]. To manage these scarce resources in the best possible way and enable an effective and efficient sharing, prognosis models for individual disease courses and outcomes are essential [9,10]. Healthcare providers can use predictive models to prioritize patients at increased risk of

* Corresponding author. Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran.
E-mail address: H.kazemi@abadanums.ac.ir (H. Kazemi-Arpanahi).

https://doi.org/10.1016/j.imu.2022.100919
Received 22 January 2022; Received in revised form 25 February 2022; Accepted 15 March 2022
Available online 18 March 2022

clinical deteriorating and public health authorities can use them to inform target public health interventions [11,12].

Several studies have been pursued to detect factors contributing to poor outcomes resulting from COVID-19 [13–16]. Some studies have revealed that machine learning (ML) can be applied to construct effective predictive models for critical and fatal courses in COVID-19 patients [17–19]. ML classifiers comprise supervised and unsupervised techniques; we employed supervised ones in our study. In these methods, a part of the data is used as a training section to develop models and the remaining data is for testing the developed models [20]. To predict disease progression, patient condition deterioration, need for ICU hospitalization and intubation risk, previous studies have employed multiple supervised ML models, including artificial neural networks (ANNs), DT, support vector machine (SVM), random forest (RF) and Naive Bayes (NB) [15,21].

ML helps analyze a large dimensional dataset automatically and reveals significant hidden relationships or patterns. ML-based approaches can increase sensitivity and specificity by training data on COVID-19 patients [11]. However, the likelihood of some methods, including DT algorithms, has not yet been addressed in enhancing the prediction capabilities of COVID-19 poor outcomes. It is also required to find techniques for producing precise predictions [22]. In this study, to address these issues, we retrospectively analyzed the data of COVID-19 patients easily available at the time of admission to the hospital. We studied the most affecting clinical features for ICU admission. Furthermore, we developed and compared various DT algorithms to distinguish COVID-19 patients with high likelihood for ICU admission from those without.

2. Material and methods

2.1. Study design and participants

This study retrospectively reviewed a COVID-19 hospital-based registry database from Ayatollah Talleghani Hospital (COVID-19 referral center), Abadan city, Southwest of Khuzestan Province, Iran, from February 9, 2020, to December 20, 2020. During the study period, 7214 suspected cases with COVID-19 were referred to Ayatollah Talleghani Hospital's ambulatory and emergency departments (EDs), of whom 2253 cases were introduced as positive RT-PCR COVID-19, 2472 as negative and 2489 as unknown. After applying the inclusion/exclusion criteria, 512 hospitalized record cases were entered into the study (311 and 201 records belonged to ICU and non-ICU admitted, respectively) Fig. 1.

2.2. Study features

The included cases were defined based on 53 features in five categories including patient's basic information such as age (year), sex (men/women), height (centimeters), weight (Kg) and blood group (five features), clinical features such as cough (Have/Haven't), nausea (Have/Haven't), headache (Have/Haven't), gastrointestinal (GI) manifestation (Have/Haven't), chill (Have/Haven't), loss of taste (Have/Haven't) and smell (Have/Haven't), rhinorrhea (Have/Haven't), sore throat (Have/Haven't), contusion (Have/Haven't), fever (Have/Haven't), muscular pain (Have/Haven't), vomiting and dyspnea (Have/Haven't), history of personal diseases such as cardiac disease (Have/Haven't), smoking (Yes/No), pneumonia (Have/Haven't), hypertension (Have/Haven't), alcohol addiction (Have/Haven't), diabetes (Have/Haven't) and other underlining diseases (Have/Haven't), laboratory

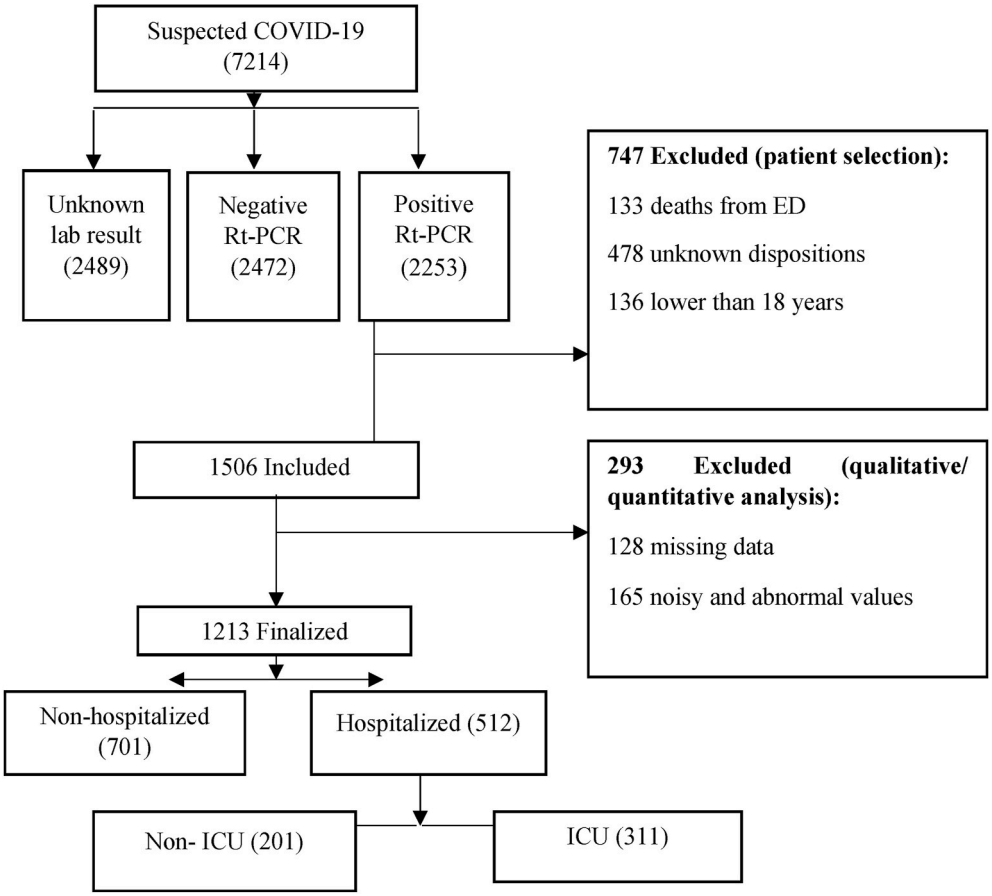


Fig. 1. Flow chart describing patient selection.

information such as red-cell count, hematocrit, hemoglobin, absolute lymphocyte count, blood calcium, blood potassium, absolute neutrophil count, alanine aminotransferase (ALT), magnesium, activated partial, prothrombin time, alkaline phosphatase, platelet count, hypertensive troponin, creatinine, white cell count, aspartate aminotransferase (ASP), blood glucose, total bilirubin, erythrocyte sedimentation rate (ESR), c-reactive protein, albumin, thromboplastin time, lactate dehydrogenase (LDH), blood phosphorus, blood sodium and blood urea nitrogen (BUN), remedies such as oxygen therapy (Have/Haven't), length of hospitalization (day) and an attribute serving as an output variable (ICU admission (Yes, No)). In Table 1, more details about the laboratory variables are represented.

2.3. Preprocessing

First, the incomplete case records with many missing values (more than 70%) were excluded from the analysis. Also, the remaining missing cells were credited with the mean and 9999 values of each variable for quantitative and qualitative fields, respectively. In addition, noisy and abnormal values, errors, duplicates and meaningless data were checked by two health information management experts (M: SH and H: KA) collaborating with two infectious diseases specialists and one hematologist. For different interpretations about data preprocessing, we contacted the corresponding physicians.

2.4. Feature selection

The feature selection process is a beneficial statistical method for determining the most important variables highly correlated with the dependent (output) variable, especially in large-scale databases [23]. Benefits of this statistical process include preventing from overfitting the data mining algorithms, better classifying the dataset samples in terms of performance, investigating the fewer variables for work simplification and better clustering the samples in databases without classes [24]. In this study, the independence test of Chi-square (Equation (1)) was utilized for weighting the features based on their importance in predicting ICU hospitalization among COVID-19 patients. In Equation (1), O_i and E_i are the observed and expected variables existing for the variables, respectively. $P < 0.01$ was regarded as the significant level in this respect. Also, logistic regression was utilized for determining the variables with the high odd ratio at $p < 0.05$ before the model construction.

$$\chi 2 = \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

2.5. Model development and evaluation

In this section, first, a set of the best variables for predicting ICU hospitalization was selected using independence test of Chi-square. Then, logistic regression analysis was performed to calculate odd ratio with specific Wald at $P < 0.05$. Afterwards, seven DT algorithms, including the decision stump (DS), Hoeffding tree (HT), LMT, J-48, random forest (RF), random tree (RT) and REP-Tree, were trained for developing the prediction models for predicting ICU hospitalization. Finally, the DT predictivity capabilities were compared to select the most performing algorithms ones. The 10 fold cross-validation was utilized in this respect. The performance criteria were positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, accuracy and F-score (Equation 2 through 7, respectively).

We obtained all the performance criteria using the confusion matrix, including the true positive (TP), false positive (FP), false negative (FN) and true negative (TN). The TP and TN are ICU and non-ICU admitted cases that are correctly classified by the model. Also, FN and FP are the cases incorrectly classified by the model.

$$PPV = TP/(TP + FP) \quad (2)$$

Table 1
The characteristics of laboratory variables.

NO	Variable (Units)	Ranges	Description
1	Blood creatinine (mg/dL)[1]	Reference: 0.7–1.3 (men), 0.6–1.1 (women) Low: <0.7 (men), <0.6 (women), High: >1.3 (men), >1.1 (women)	The creatinine rate in the blood
2	Red cell count (mc/mL)[2]	Reference: 4.35–5.65 (men), 3.92–5.13 (women) Low: <4.35 (men), <3.92 (women) High: > 5.65 (men) > 5.13 (women)	The red cells count in plasma
3	Hematocrit (L/L)[3]	Reference: 0.40–0.54 (men), 0.37–0.47 (women) Low: <40 (men) <0.37 (women) High: >0.54 (men) > 0.47 (women)	The proportion of the red cells count to the plasma cells count
4	Hemoglobin rate (g/dL)[4]	Reference: 14.0–17.5 (men) 12.3–15.3 (women) Low: <14.0 (men) < 12.3 (women) High: >17.5 (men) > 15.3 (women)	The protein rate in red blood cells that carries iron
5	Platelet count (Cells/μL)[5]	Reference: 150,000–400,000. Low: <150000 High: >400000	Number of platelet cells count in the plasma
6	Absolute lymphocyte count (10^3 Cells/μL)[5]	Reference: 1–4.8 Low:< 1 High:> 4.8	The absolute number of lymphocyte cells in the blood that can be acquired by multiplying the number of white cells and lymphocyte percentage
7	Absolute neutrophil count (10^3Cells/μL)[5]	Reference: 2.5–6 Low <2.5 High:> 6	The absolute number of neutrophil cells in the blood that can be acquired by multiplying the number of white cells and neutrophil percentage
8	Blood calcium (mg/dL)[1]	Reference: 8.6–10.3 Low: <8.6 High: >10.3	The calcium rate in the blood
9	Blood sodium (mEq/L)[6]	Reference: 135-145 Low:< 135 High:> 145	The sodium rate in the blood
10	Blood magnesium (mEq/L)[6]	Reference: 1.3–2.1 Low<1.3 High: >2.1	The magnesium rate in the blood
11	Blood phosphor (mg/dL)[1]	Reference: 3.4–4.5 Low: <3.4 High:> 4.5	The phosphor rate in the blood
12	Blood potassium (mEq/L)[6]	Reference: 3.5–5.2 Low: <3.5 High>5.2	The potassium rate in the blood
13	Blood urea nitrogen (mg/dL)[1]	Reference: 6-24 Low: <6 High:> 24	Amount of urea nitrogen found in blood
14	Total bilirubin (mg/dL)[1]	Reference: 1.2 Low:< 1.2 High:> 1.2	Amount of bilirubin in the blood
15	Aspartate aminotransferase (units/L)[7]	Reference: 8-33 Low: <8 High:> 33	The amount of aspartate aminotransferase enzymes in the blood
16	Alanine aminotransferase (units/L)[7]	Reference: 29–33 (men) 19–25 (women) Low: <29 (men) < 19 (women) High: >33 (men) > 25 (men)	The amount of alanine aminotransferase enzymes in the blood

(continued on next page)

Table 1 (*continued*)

NO	Variable (Units)	Ranges	Description
17	Serum albumin (g/dL)[8]	Reference: 3.4–5.4. Low: <3.4 High:> 5.4	albumin amount which are in vertebrate blood
18	Blood glucose (mg/dL)[1]	Reference: <140 Diabetes: >200 Prediabetes: 140-199	The glucose rate in the blood
19	Lactate dehydrogenase (Units/L)[7]	Reference: 140 -280 Low: <140 High: >280	Amounts of lactic acid dehydrogenase in the blood
20	Activated partial thromboplastin time (s)[9]	Reference: 30-40 Fast:<30 Slow: >40	Measures the time that the clot is formed in a blood specimen
21	Prothrombin time (s)[9]	Reference: 11–13.5. Fast: <11 Slow: >13.5	Measures the time that the liquid portion of blood are clotted
22	Alkaline phosphatase (Units/L)[7]	Reference: 44-147 Low: <44 High:> 147	The amount of Alkaline phosphatase enzymes in the blood
23	C-reactive protein (mg/L)[10]	Reference: <10 High: ≥10.	The amount of this protein in the blood and increases in inflammation conditions
24	Erythrocyte sedimentation rate (mm/hr)[11]	Reference: 0–22 (men), 0–29 (women) Abnormal: >22 (men), >29 (women)	Measure the quantity at which red-type blood cells subsist at the end of a test tube containing a blood specimen
25	White cell count (Cells/mL)[12]	Reference: 4500–11,000 Low:<4500 High:> 11000	The white-type cells count in the plasma
26	Hypersensitive troponin (ng/L)[13]	Normal: =<14 Abnormal: >14	This test can be used for heart attack and insufficiency, in other words the >14 in bloodstream indicates heart attack

1- Milligram per deciliter. 2- Million cells per microliter. 3- Number of red cells per liter per number of cells per liter. 4- Grams per deciliter. 5- Number of cells per microliter. 6- Miliequivalents per liter. 7- Units per liter. 8- Grams per deciliter. 9-Seconds. 10- Milligrams per liter. 11- Millimeters per hour. 12- Cell per microliter. 13- Nanograms per liter.

$$NPV = TN/(TN + FN) \quad (3)$$

$$Sensitivity = TP/(TP + FN) \quad (4)$$

$$Specificity = TN/(TN + FP) \quad (5)$$

$$Accuracy = TP + TN/(TP + FN + TN + FP) \quad (6)$$

$$F - Score = TP\ /\ (TP + 1\ /\ 2(FP + FN) \quad (7)$$

Moreover, the area under the ROC curve (AUC) of seven DT algorithms was compared in terms of their ability to classify the samples. In the next step, the best DT algorithm for predicting ICU hospitalization among COVID-19 patients was obtained by comparing their performance measured using the mentioned evaluation criteria. Finally, the best performing algorithm was described and the most weighted clinical rules were extracted.

2.6. Ethical consideration

Ethical Committee Board of Abadan University of Medical Sciences (ethics code: IR. ABADANUMS.REC.1400.110) approved the study. To protect the privacy and confidentiality of the patients, we concealed the unique identification information of all the patients in the process of data collection and presentation.

3. Results

3.1. Characteristics of participants

After applying the inclusion/exclusion criteria, in total, 512 patients met the eligibility criteria. Of these, 388 (75.78%) were male and 124 (24.22%) were female with the median age of 57.25 (interquartile 18–100, mean +/- SD = 57.25+/- 17.606). Also, 311 (60.75%) were ICU admitted and 201 (39.25%) were non-ICU admitted.

3.2. Features and their importance

After using the independence test of Chi-square, 20 variables had significant relationship with output class (ICU hospitalization) at $P < 0.01$, as shown in Table 2.

Given the information in Table 1, the length of hospitalization (x^2 =28.71), loss of smell (x^2 =13.372), history of other underlining diseases (x^2 =23.277) and cardiac disease (x^2 =12.491), blood pressure (x^2 =13.281), activated partial thromboplastin time (x^2 =117.458), age (x^2 =35.292) and pleural fluid (x^2 =30583) had a good relationship with ICU hospitalization possibility at $P < 0.001$. Thus, they were considered as the most important determinants to predict ICU hospitalization. Results of determining the odds ratio of 20 important variables in predicting ICU hospitalization among COVID-19 patients are demonstrated in Table 3.

Based on the information provided by Table 2, a set of 12 variables such as length of hospitalization (ORs = 2.022) 95% ORs CI = [1.225,

Table 2
The most important variable at $P < 0.01$ using Chi-squared test.

No.	Variable name	Variable type	Frequency or mean ± SD	x^2	P (level)
1	Length of hospitalization	Numeric	5.03 ± 2.188	28.71	<0.001
2	Contusion	Nominal	Have (180) Haven't (302)	7.97	<0.01
3	Oxygen therapy	Nominal	Have (437) Haven't [45]	7.99	<0.01
4	Dyspnea	Nominal	Have (442) Haven't [40]	7.023	<0.01
5	Loss of taste	Nominal	Have (124) Haven't (358)	8.722	<0.01
6	Loss of smell	Nominal	Have (137) Haven't (345)	13.372	<0.001
7	Runny nose	Nominal	Have (202) Haven't (280)	10.239	<0.01
8	Other underline diseases	Nominal	Have (339) Haven't (143)	23.277	<0.001
9	Cardiac diseases	Nominal	Have (157) Haven't (325)	12.491	<0.001
10	Blood pressure	Nominal	Have (189) Haven't (293)	13.281	<0.001
11	Diabetes	Nominal	Have (124) Haven't (358)	10.026	<0.01
12	White cell count	Numeric	9684 ± 1241	196.616	<0.01
13	Absolute lymphocyte count	Numeric	21.702 ± 12.01	83.41	<0.01
14	Absolute neutrophil count	Numeric	76.71 ± 12.765	97.661	<0.01
15	Blood sodium	Numeric	138.27 ± 3.44	40.667	<0.01
16	Blood glucose	Numeric	148.4 ± 96.946	12.884	<0.01
17	Activated partial thromboplastin time	Numeric	35.453 ± 9.25	117.458	<0.001
18	Hypertensive troponin	Nominal	Abnormal [38] Normal (444)	14.588	<0.01
19	Age	Numeric	57.25 +/- 17.606	35.292	<0.001
20	Pleural fluid	Nominal	Have (275) Haven't (78)	30.583	<0.001

Table 3
The most important determinant in predicting ICU hospitalization using odds ratio.

No	Variable	Wald	df	P-$_{value}$	Odds ratio	95% Confidence interval for odds ratio	
						Lower	Upper
1	Oxygen therapy	4.007	1	0.031	1.375	1.055	2.545
2	Dyspnea	3.830	1	0.036	1.335	2.032	4.523
3	Loss of taste	4.565	1	0.033	1.489	1.254	1.943
4	Loss of smell	4.726	1	0.030	1.474	1.242	1.929
5	Runny nose	3.473	1	0.042	1.570	1.315	2.030
6	Other underline disease	2.690	1	0.010	1.499	1.002	1.945
7	Cardiac disease	3.137	1	0.028	2.671	1.323	3.396
8	Blood pressure	0.179	1	0.673	0.853	0.408	1.784
9	Diabetes	3.356	1	0.031	2.776	1.437	3.285
10	White-cell count	0.000	1	0.092	1.000	1.000	1.000
11	Absolute lymphocyte count	0.075	1	0.784	0.987	0.899	1.084
12	Absolute neutrophil count	0.878	1	0.349	1.042	0.956	1.135
13	Sodium	0.816	1	0.366	1.039	0.956	1.129
14	Glucose	0.885	1	0.347	1.002	0.998	1.007
15	Activated partial thromboplastin time	4.072	1	0.017	3.004	1.977	5.031
16	Hypersensitive troponin	5.741	1	0.117	0.016	0.001	0.471
17	Age	6.380	1	0.012	3.565	2.227	5.708
18	Pleural fluid	2.285	1	0.025	1.222	0.89	2.999
19	Length of hospitalization	3.101	1	0.019	2.022	1.225	3.166
20	Contusion	2.277	1	0.131	0.622	0.336	1.152

3.166], age (ORs = 3.565) 95% ORs CI = [2.227, 5.708], activated partial thromboplastin time (ORs = 3.004) 95% ORs CI = [1.977, 5.031], diabetes (ORs = 2.776) 95% ORs CI = [1.437, 3.285], cardiac disease (ORs = 2.671) 95% ORs CI = [1.323, 3.396], other underlining diseases (ORs = 1.499) 95% ORs CI = [1.002, 1.945], runny nose (ORs = 1.570) 95% ORs CI = [1.315, 2.030], loss of smell (ORs = 1.474) 95% ORs CI = [1.242, 1.929], loss of taste (ORs = 1.489) 95% ORs CI = [1.254, 1.943], oxygen therapy (ORs = 1.375) 95% ORs CI = [1.055, 2.545], dyspnea (ORs = 1.335) 95% ORs CI = [2.032, 4.523] and pleural fluid (ORs = 1.222) 95% ORs CI = [0.89, 2.999] had the higher odds ratio than other variables with higher specific Wald along with DF = 1 at $P < 0.05$ in predicting ICU hospitalization among the COVID-19 patients. Therefore, they were used for building DT models.

3.3. Predictive performance of the models

The results of classifying the sample in the selected DT algorithms with specific characteristics are shown below:

DS: Batch size = 100, Number of decimal places = 2, TP = 278, FP = 137, FN = 13 and TN = 54.

HT: Batch size = 100, Grace period = 200, Hoffding tie threshold = 0.05, Split confidence = 1.0E-7, Split criterion = Info gain, TP = 254, FP = 117, FN = 37 and TN = 74.

J-48: Batch size = 100, Confidence factor = 0.25, Minimal object number = 2, Number of seed = 1, Fold number = 3, TP-269, FP = 117, FN = 22 and TN = 126.

LMT: Batch size = 15, Minimum instances in leaves = 15, Number of boosting iterations = -1, Number of decimal places = 2, TP = 254, FP = 89, FN = 37 and TN = 102.

RF: Bach size and bag size = 100, Number of decimal places = 2, Max depth = 0, Number of iterations = 100, Number of seed = 1, TP = 233, FP = 73, FN = 58 and TN = 118.

RT: Batch size = 100, Number of decimal places = 2, Minimum variance property = 0.001, Number of seed = , TP = 211, FP = 95, FN = 80 and TN = 96.

REP-Tree: Batch size = 100, Minimum variance property = 0.001, Number of folds = 3, Number of seed = 1, TP = 257, FP = 88, FN = 34 and TN = 103.

Based on the information provided, the DS and J-48 tree algorithms with TP = 278 and TN = 126 acquired the best performance in classifying the ICU hospitalized versus the non-hospitalized cases, respectively. Some of the DT algorithm performance criteria are depicted in Fig. 2.

Based on Fig. 1, DS had lower specificity (specificity = 0.28) than other algorithms, meaning this algorithm's lowest capacity in classifying the negative cases (non-hospitalized COVID-19 patients). On the contrary, the sensitivity of this algorithm (sensitivity = 0.96) was higher than others, which demonstrated its better ability in classifying positive cases (hospitalized COVID-19 patients) in this research. In general, the J-48 algorithm based on the PPV, NPV and accuracy obtained better performance than others.

Considering F-score as a criterion related to the strength of algorithms in classifying both positive and negative cases demonstrated that the J-48 algorithm with the F-score of 0.816 had better ability than the others in this respect. Also, the LMT (F-score = 0.729), RF (F-score = 0.726) and REP-tree (F-score = 0.737) had desirable performance in this regard. The ROCs of all DT algorithms are depicted in Fig. 3. The vertical and horizontal vertices showed sensitivity and 1-specificity, respectively.

Based on comparing the AUC of the selected DT algorithms, it is determined that the J-48 algorithm with the AUC of 0.845 had more area under the ROC curve than other algorithms. The ROC diagram of this algorithm was closer to sensitivity or TP and, simultaneously, farther than 1-specificity or FP, which demonstrated the better performance of this algorithm in classifying ICU and non-ICU COVID-19 hospitalized patients. Generally, the results of comparing different DT algorithms for predicting ICU hospitalization among COVID-19 patients using various evaluation criteria demonstrated that the J-48 algorithm with PPV = 0.805, NPV = 0.85, sensitivity = 0.924, specificity = 0.659, accuracy = 0.819, F-score = 0.816 and AUC = 0.845 had the higher performance than other DT algorithms in classifying the ICU and non-ICU cases. The important characteristics for building the tree are mentioned below with more details.

3.4. Important characteristics for constructing the J-48 algorithm with the highest performance

Batch size = 100, Binary split = haven't, Collapse tree = true, Confidence factor = 0.25, Minimum number of objects = 2, Number of decimal places = 2, Number of folds = 3, Reduced error pruning = have and Number of seed = 1.

In Fig. 4, we brought the pruned J-48 algorithm with confidence factor = 0.25 for classifying the samples of ICU and non-ICU COVID-19 patients. According to the drawn tree with SIZE = 31 and number of leaves = 16, we found five important points of the tree's leave. Most of the dataset samples were classified and the clinical rules for predicting ICU hospitalization among the COVID-19 patients were extracted. We considered the activated partial thromboplastin time as the tree's root. They existed in the tree's 1st, 3rd, 11th, 14th and 16th leaves with 64, 46, 189, 57 and 66 samples, respectively. Now, we interpreted two numbers of these five extracted rules belonging to these five important leaves with more classified samples.

Rule 1: IF (Activated partial thromboplastin time >31 && Age>65, && activated partial thromboplastin time≤41 && Diabetes = No && Loss of smell = No && Pleural fluid = Yes THEN ICU = 0. (189/25).

Rule 2: IF (Activated partial thromboplastin time >31 && Age>65, && activated partial thromboplastin time≤41 && Diabetes = Yes && Loss of taste = No THEN ICU = 0. (66/20).

In rule 1 based on the J-48, we can interpret that, if a COVID-19 patient has an activated partial thromboplastin time between 31 and 41 with age and without loss of smell and history of having diabetes and

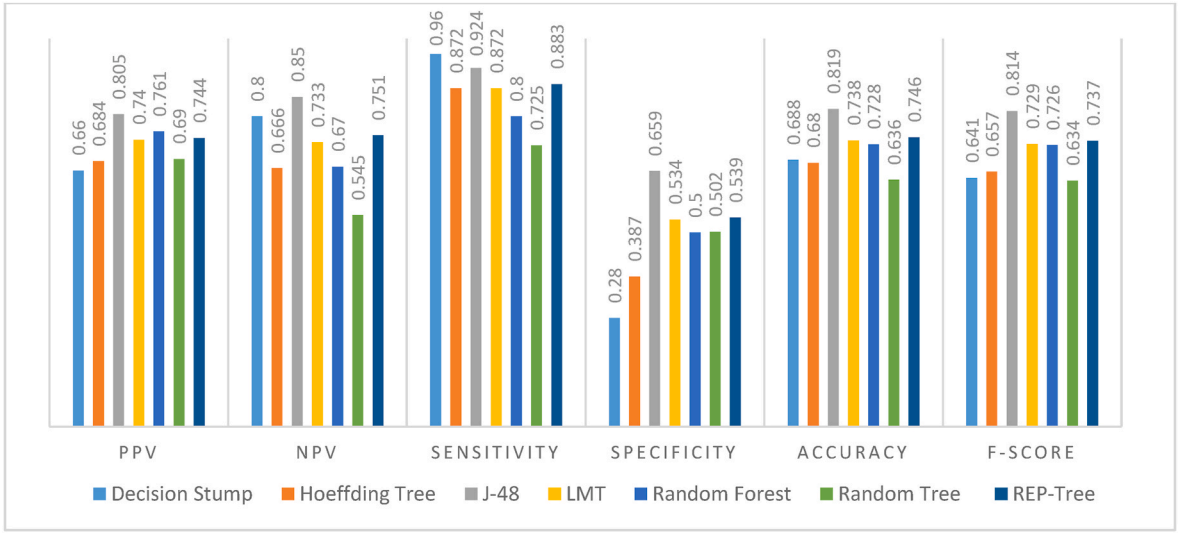


Fig. 2. Different evaluation criteria of decision tree algorithms.

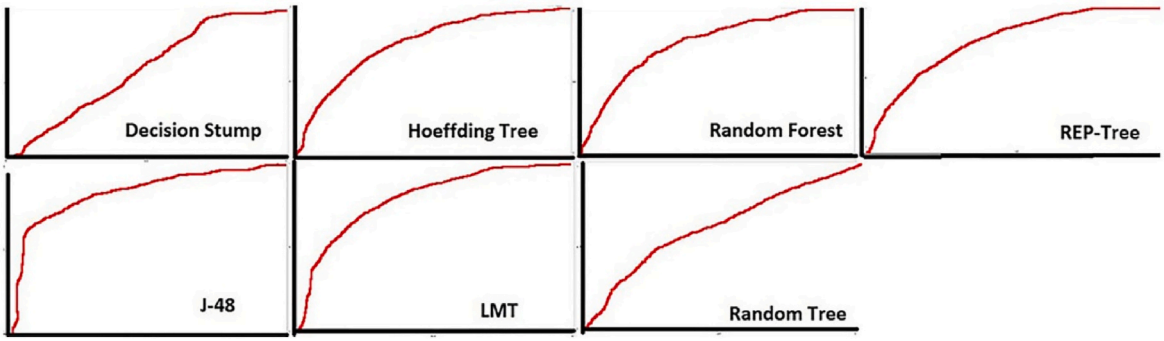


Fig. 3. AUC of different decision tree algorithms.

pleural fluid with the probability of 86%, the person will not be admitted in ICU. In rule 2, if a COVID-19 patient has an activated partial thromboplastin time between 31 and 41 with higher age and history of diabetes and without loss of taste with the probability of 69%, the COVID-19 patients will not enter the ICU.

The results of external cohort validation of the predictive model using the confusion matrix are shown in Table 4.

As shown in Fig. 5, the J-48 decision tree algorithm with AUC = 0.822 gained acceptable performance in predicting the ICU admission using the test dataset in an external environment. The performance was near the internal test results, which used cross-validation (AUC = 0.845). Generally, the J-48 decision tree with F-score = 80.9% and AUC = 0.822 had the common performance, especially in classifying the ICU-admitted cases.

4. Discussion

With the COVID-19 outbreak, the global health system faces challenges from the overwhelming workload of health staff to decreased resources such as ICU beds and ventilators. The shortage in ICU resources and increasing number of patients will force health policy-makers and managers to rely on scientific and specified programs to deal with limited hospital resources. Predicting which patients are at high risk for progression and poor outcomes can guide physicians in selecting appropriate treatment and allocating scarce specialized and vital equipment toward critically ill patients [25]. ML prediction models create remarkable opportunities to identify the most involved factors and best decisions about each situation. This study aimed to develop prediction models for estimating ICU hospitalization among COVID-19 patients based on data that are easily obtained at the first time of admission. For this purpose, seven DT methods, including DS, HT, LMT, J-48, RF, RT and REP-tree, were trained using 512 de-identified case records of COVID-19 in-hospital patients. For this purpose, we used Abadan COVID-19 registry, including 201 samples of non-ICU admitted and 311 ICU admitted patients.

4.1. Features of interest

This single-center retrospective study, first, determines and ranks contributing predictors affecting ICU admission. Selecting reliable and clinically relevant predictors related to COVID-19 patients could help improve the accuracy of prediction models. In addition, the selection of significant variables in predictive models can provide insight into forecasters and their acceptable relations to the pathophysiology of clinical decline in COVID-19 patients [26]. We identified 20 important factors for predicting the needing ICU care for COVID-19 hospitalized patients based on the independence test of Chi-square. Logistic regression was used to determine the variables with the high odds ratio. Accordingly, in our study, old age, length of hospitalization, activated partial thromboplastin time, diabetics, cardiac diseases, runny nose, loss of smell, loss of taste, oxygen therapy, dyspnea and pleural fluid had a high odds ratio with specific Wald at $p < 0.05$. So, they were selected as the most contributing factor in predicting COVID-19 ICU admission. The results of our study demonstrated that three variables of old age (ORs = 3.565) 95% ORs CI = [2.227, 5.708], activated partial thromboplastin time (ORs = 3.004) 95% ORs CI = [1.977, 5.031] and history of diabetes

```
J48 pruned tree
------------------

Activated partial thromboplastin time <= 31: 1 (64.0/10.0)
Activated partial thromboplastin time > 31
|   Consolidation = No: 1 (46.0/8.0)
|   Consolidation = Yes
|   |   Activated partial thromboplastin time <= 41
|   |   |   Diabetes = No
|   |   |   |   Loss of smell = No
|   |   |   |   |   Pleural fluid = No
|   |   |   |   |   |   Activated partial thromboplastin time <= 34: 0 (2.0)
|   |   |   |   |   |   Activated partial thromboplastin time > 34: 1 (3.0)
|   |   |   |   |   Pleural fluid = Yes: 0 (189.0/25.0)
|   |   |   |   Loss of smell = Yes
|   |   |   |   |   Length of hospitalization <= 2: 1 (5.0)
|   |   |   |   |   Length of hospitalization > 2: 0 (57.0/17.0)
|   |   |   Diabetes = Yes
|   |   |   |   Loss of taste = No: 0 (66.0/20.0)
|   |   |   |   Loss of taste = Yes
|   |   |   |   |   Cardiac Disease = No
|   |   |   |   |   |   Activated partial thromboplastin time <= 33: 0 (3.0)
|   |   |   |   |   |   Activated partial thromboplastin time > 33: 1 (8.0/1.0)
|   |   |   |   |   Cardiac Disease = Yes
|   |   |   |   |   |   Activated partial thromboplastin time <= 33: 1 (3.0)
|   |   |   |   |   |   Activated partial thromboplastin time > 33
|   |   |   |   |   |   |   Length of hospitalization <= 6: 1 (3.0/1.0)
|   |   |   |   |   |   |   Length of hospitalization > 6: 0 (3.0)
|   |   Activated partial thromboplastin time > 41
|   |   |   Runny nose = No
|   |   |   |   Length of hospitalization <= 3: 1 (2.0)
|   |   |   |   Length of hospitalization > 3: 0 (14.0/3.0)
|   |   |   Runny nose = Yes: 1 (14.0/2.0)

Number of Leaves  :     16

Size of the tree :      31
```

Fig. 4. Pruned J-48 decision tree algorithm.

Table 4
Confusion matrix for external dataset.

	Predicted ICU admitted	Predicted non-ICU admitted	Total
Real ICU admitted	53	8	61
Real non-ICU admitted	17	30	47
Total	70	38	108

Based on Table 4, we obtained the predictive model performance criteria as PPV = 75.7%, NPV = 32%, sensitivity = 86.9%, specificity = 63.8%, accuracy = 76.8% and F-score = 80.9%. The ROC of the J-48 for the external dataset is depicted in Fig. 5.

(ORs = 2.776) 95% ORs CI = [1.437, 3.285] had the top variables according to odds ratio.

Many studies have been focused on determining the key risk factors for ICU admission. COVID-19 patients with the underlining diseases such as hypertension [27], diabetes [28], cancer [29] and lung diseases [30] were considered to be susceptible to having poor prognosis. They had higher risk of admission to an ICU, invasive ventilation or death. Results of prior studies have also shown that older age [31], decreased oxygen saturation [32], high sequential organ failure assessment score [33], higher D-dimer [34], leukocytosis [35] and high fever [36] are regarded as the most effective factors for predicting COVID-19 ICU risk. In general, high compliance is observed from classifying and prioritizing variables in the reviewed studies with the most common variables in our study.

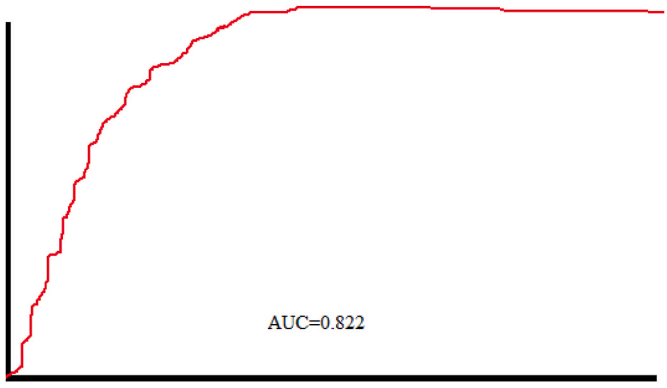


Fig. 5. The ROC of J-48 for the external dataset.

4.2. Developed predictive models

In our study, the DT algorithms were trained using the selected top variables as input data. The results of comparing the different selected DT algorithms demonstrated using the J-48 generally, with F-score = 0.816 and AUC = 0.845 had the best performance in classifying the ICU and non-ICU COVID-19 hospitalized patients.

In some of the related studies, the functionality of these algorithms in COVID-19 prediction has been investigated. Goncalves et al. (2020) retrospectively studied 827621 confirmed COVID-19 patients' data from Centers for Disease Control and Prevention (CDC) of COVID-19 case surveillance database. They tested 10 DT–based ensemble ML methods

on the selected dataset for predicting COVID-19 deterioration in ICU hospitalized patients. Finally, the best significant results were observed from the AdaBoost model (AUC of 91.91%) [22]. Castiglioni et al. (2021) also conducted a retrospective analysis on data of 270 COVID-19 and non-COVID-19 cases. They then developed an intelligent model based on DT algorithms to predict the need for hospitalization of COVID-19 patients. Their results showed that the model developed using J-48 with 0.81 of AUC gained the best performance [37]. Besides, Famiglini (2021) compared three DT classifiers' performance based on 4995 CBC tests to predict ICU admission in COVID-19 patients. The experimental results showed that the ensemble decision tree (EDT) was introduced as the most suitable algorithm (AUC of 88%) [38]. Ahmad et al. (2021) retrospectively assessed the 600 laboratory findings of confirmed and negative COVID-19 patients using 18 variables. Ten DT algorithms were tested and the XGBoost DT algorithm gained the best predictive performance with the AUC of 0.873 [39]. Another work by Vetrugno et al. (2020) analyzed the data of 198 COVID-19 hospitalized patients and showed that the DT achieved the highest accuracy to predict the need for hospitalization or home monitoring of confirmed or suspected cases with the ROC of 0.75 [40]. Finally, Talebi et al. (2020) designed a DT-based model for predicting the COVID-19 patient status using chest x-ray data of 1078 COVID-19 confirmed patients. The result showed classification and regression tree (CART) gained optimum predictive performance with accuracy, sensitivity and specificity of 93.3%, 72.8% and 97.1%, respectively [41]. The result of comparing the DT algorithms demonstrated that J-48 with the F-score of 0.816 and AUC of 0.845 had the best performance.

4.3. Strength and limitations

The developed models in our study had several opportunities for clinical use as a screening tool for potential infectious disease outbreaks such as the current COVID-19 crisis. These models reduced the current uncertainty and ambiguity in the COVID-19 clinical practice by providing measurable, non-subjective and evidence-based approaches [42,43]. Accurate prediction of patient admission to the ICU could support the optimal allocation of limited hospital resources, improve the quality of care and reduce patients mortality [43]. Early identification of at-risk patients may potentially reduce the need for imminent ICU beds and invasive mechanical ventilators. In addition, the use of these predictive models can increase the rate of timely transfer to the ICU, lead to a reduction in mortality and result in shorter stay in the ICU. This could reduce ambiguity by providing quantitative, objective and evidence-based models for risk classification, forecasting and ultimately care planning [44,45].

This study had some limitations that need to be addressed. First, because of analyzing a single-center and retrospective database, we were not able to include even more patients in the analysis. However, the used dataset was collected at Ayatollah Talleghani Hospital that delivered only special care to COVID-19 patients. Even so, the data of another COVID-19 hospital center was used to perform external validation of the proposed models for increasing the accuracy prediction. The small sample size could be acceptable criticism, but the dataset analyzed in our study were manually gathered and adjusted. The data were not exported electronically from the database, in which missing data is common, and the validity of the information was not verified. Second, this study only included 12 clinical variables available at the initial time of admission. It does not mean these should be the only criteria for predicting ICU admission. However, according to the aim of the present study, it is sufficient to consider only the routine clinical features of patients at the beginning of hospitalization. Although the limitation of using data at the point of admission encourages adopting the models in patients' triage, events that occur during patients' hospitalization period may change their clinical course, which is not understood by the available admission data. Third, the dynamic variations of some significant variables must be followed up to recognize patients at higher risks of poor outcomes in a better and timely manner. Finally, the selected dataset lacked important clinical variables such as radiological and imaging indicators. In future, the performance accuracy of our model and its generalizability will be enhanced if we test more ML techniques in a larger, multicenter and prospective dataset, which is equipped with more qualitative and validated data.

5. Conclusions

This study identified the highly ranked clinical predictors that can predict the likelihood of ICU admission more precisely. Based on these findings, we developed and compared some DT-driven prediction models. In particular, it was observed that the J-48 model performed best on classification accuracy among other DT algorithms. This method had the potential to provide frontline clinicians with an objective instrument to manage COVID-19 patients more efficiently in such time-sensitive, resource-demanding, and potentially resource-constrained situations. Finally, the comparison results of prediction models' performance in this study were satisfactory to some extent and we believe further investigations are needed to validate our model in the larger, multi-central and more qualitative dataset.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank the Abadan University of Medical Sciences research deputy for financially supporting this project. (IR.ABADANUMS.REC.1400.110).

References

[1] Gheysarzadeh A, Sadeghifard N, Safari M, Balavandi F, Falahi S, Kenarkoohi A, et al. Report of five nurses infected with severe acute respiratory syndrome coronavirus 2 during patient care: case series. Elsevier; 2020. p. 100694.
[2] Wang Z, Deng H, Ou C, Liang J, Wang Y, Jiang M, et al. Clinical symptoms, comorbidities and complications features in severe and non-severe patients with COVID-19: a systematic review and meta-analysis without cases duplication. 2020.
[3] Falahi S, Abdoli A, Kenarkoohi A. Claims and reasons about mild COVID-19 in children. New Microbes New Infect 2021;41:100864.
[4] Lechien JR, Chiesa-Estomba CM, Place S, Van Laethem Y, Cabaraux P, Mat Q, et al. Clinical and epidemiological characteristics of 1420 European patients with mild-to-moderate coronavirus disease 2019. J Intern Med 2020;288(3):335–44.
[5] Smith EM, Lee ACW, Smith JM, Thiele A, Zeleznik H, Ohtake PJ. COVID-19 and post-intensive care syndrome: community-based care for ICU survivors. Home Health Care Manag Pract 2021;33(2):117–24.
[6] Abate SM, Ali SA, Mantfardo B, Basu B. Rate of intensive care unit admission and outcomes among patients with coronavirus: a systematic review and Meta-analysis. PLoS One 2020;15(7 July).
[7] Sadeghi A, Eslami P, Moghadam AD, Pirsalehi A, Shojaee S, Vahidi M, et al. COVID-19 and ICU admission associated predictive factors in Iranian patients. Casp J Intern Med 2020;11:S512–9.
[8] Supady A, Curtis JR, Abrams D, Lorusso R, Bein T, Boldt J, et al. Allocating scarce intensive care resources during the COVID-19 pandemic: practical challenges to theoretical frameworks. Lancet Respir Med 2021;9(4):430–4.
[9] Lichtner G, Balzer F, Haufe S, Giesa N, Schiefenhövel F, Schmieding M, et al. Predicting lethal courses in critically ill COVID-19 patients using a machine learning model trained on patients with non-COVID-19 viral pneumonia. Sci Rep 2021;11(1):1–10.
[10] Yazdani A, Sharifian R, Ravangard R, Zahmatkeshan M. COVID-19 and information communication technology: a conceptual model. J Adv Pharm Educ Res 2021;11(S1). Jan-Mar.
[11] Agieb R. Machine learning models for the prediction the necessity of resorting to icu of covid-19 patients. Int J Adv Trends Comput Sci Eng 2020:6980–4.
[12] Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. PLoS One 2020;15(7):e0236618.
[13] Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 2020;584(7821):430–6.

[14] 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):1054–62.
[15] Berenguer J, Ryan P, Rodríguez-Baño J, Jarrín I, Carratalà J, Pachón J, et al. Characteristics and predictors of death among 4035 consecutively hospitalized patients with COVID-19 in Spain. Clin Microbiol Infect 2020;26(11):1525–36.
[16] Grasselli G, Greco M, Zanella A, Albano G, Antonelli M, Bellani G, et al. Risk factors associated with mortality among patients with COVID-19 in intensive care units in Lombardy, Italy. JAMA Intern Med 2020;180(10):1345–55.
[17] Zhou Y, He Y, Yang H, Yu H, Wang T, Chen Z, et al. Exploiting an early warning Nomogram for predicting the risk of ICU admission in patients with COVID-19: a multi-center study in China. Scand J Trauma Resuscitation Emerg Med 2020;28(1):1–13.
[18] Mehrdad R, Farzaneh S, Shahab F, Azra K, Samad ABO, Seyed Younes H, et al. Prediction of hepatitis B virus lamivudine resistance based on YMDD sequence data using an artificial neural network model. 2011.
[19] Moulaei K, Shanbehzadeh M, Mohammadi-Taghiabad Z, Kazemi-Arpanahi H. Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med Inf Decis Making 2022;22(1):1–12.
[20] Shanbehzadeh M, Orooji A, Kazemi-Arpanahi H. Comparing of data mining techniques for predicting in-hospital mortality among patients with COVID-19. J Biostat Epidemiol 2021;7(2):154–69.
[21] Gao Y, Cai GY, Fang W, Li HY, Wang SY, Chen L, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nat Commun 2020;11(1).
[22] Goncalves CP, Rouco J. Comparing decision tree-based ensemble machine learning models for COVID-19 death probability profiling. medRxiv; 2020.
[23] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: a data perspective. ACM Comput Surv 2017;50(6):1–45.
[24] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507–17.
[25] Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. NPJ Digit Med 2021;4(1):1–7.
[26] Cheng F-Y, 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.
[27] Roncon L, Zuin M, Zuliani G, Rigatelli G. Patients with arterial hypertension and COVID-19 are at higher risk of ICU admission. Br J Anaesth 2020;125(2):e254–5.
[28] Fernández García L, Puentes Gutiérrez AB, García Bascones M. Relationship between obesity, diabetes and ICU admission in COVID-19 patients. Med Clínica 2020;155(7):314–5.
[29] Boilève A, Stoclin A, Barlesi F, Varin F, Suria S, Rieutord A, et al. COVID-19 management in a cancer center: the ICU storm. Support Care Cancer 2020;28(10):5037–44.
[30] Nagra D, Russell M, Yates M, Galloway J, Barker R, Desai SR, et al. COVID-19: opacification score is higher in the right lung and right lung involvement is a better predictor of ICU admission. Eur Respir J 2020;56(6).
[31] Flaatten H, Beil M, Guidet B. Prognostication in older ICU patients: mission impossible? Br J Anaesth 2020;125(5):655–7.
[32] Kjellberg A, Douglas J, Pawlik MT, Kraus M, Oscarsson N, Zheng X, et al. Randomised, controlled, open label, multicentre clinical trial to explore safety and efficacy of hyperbaric oxygen for preventing ICU admission, morbidity and mortality in adult patients with COVID-19. BMJ Open 2021;11(7):e046738.
[33] Bels JLM, van Kuijk SMJ, Ghossein-Doha C, Tijssen FH, van Gassel RJJ, Tas J, et al. Decreased serial scores of severe organ failure assessments are associated with survival in mechanically ventilated patients; the prospective Maastricht Intensive Care COVID cohort. J Crit Care 2021;62:38–45.
[34] Hachim MY, Hachim IY, Naeem KB, Hannawi H, Salmi IA, Hannawi S. D-Dimer, troponin, and urea level at presentation with COVID-19 can predict ICU admission: a single centered study. Front Med 2020;7.
[35] Yamada T, Wakabayashi M, Yamaji T, Chopra N, Mikami T, Miyashita H, et al. Value of leukocytosis and elevated C-reactive protein in predicting severe coronavirus 2019 (COVID-19): a systematic review and meta-analysis. Clin Chim Acta 2020;509:235–43.
[36] Choron RL, Butts CA, Bargoud C, Krumrei NJ, Teichman AL, Schroeder ME, et al. Fever in the ICU: a predictor of mortality in mechanically ventilated COVID-19 patients. J Intensive Care Med 2021;36(4):484–93.
[37] Castiglioni I, Ippolito D, Interlenghi M, Monti CB, Salvatore C, Schiaffino S, et al. Machine learning applied on chest x-ray can aid in the diagnosis of COVID-19: a first experience from Lombardy, Italy. Eur Radiol Exper 2021;5(1):1–10.
[38] Famiglini L, Bini G, Carobene A, Campagner A, Cabitza F, editors. Prediction of ICU admission for COVID-19 patients: a machine learning approach based on complete blood count data. IEEE 34th International Symposium on computer-based medical systems (CBMS); 2021. IEEE; 2021.
[39] Ahmad A, Safi O, Malebary S, Alesawi S, Alkayal E. Decision tree ensembles to predict coronavirus disease 2019 infection: a comparative study. Complexity 2021:2021.
[40] Vetrugno G, Laurenti P, Franceschi F, Foti F, D'Ambrosio F, Cicconi M, et al. Gemelli decision tree Algorithm to Predict the need for home monitoring or hospitalization of confirmed and unconfirmed COVID-19 patients (GAP-Covid19): preliminary results from a retrospective cohort study. Eur Rev Med Pharmacol Sci 2021;25(6):2785–94.
[41] Talebi A, Borumandnia N, Jafari R, Pourhoseingholi MA, Jafari NJ, Ashtari S, et al. Predicting the COVID-19 patients' status using chest CT scan findings: a risk assessment model based on Decision tree. 2021.
[42] Yadaw AS, Li Y-c, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health 2020;2(10):e516–25.
[43] 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.
[44] Agieb RS. Machine learning models for the prediction the necessity of resorting to icu of covid-19 patients. Int J Adv Trends Comput Sci Eng 2020;9(5):6980–4.
[45] Assaf D, Gutman Ya, Neuman Y, Segal G, Amit S, Gefen-Halevi S, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med 2020;15(8):1435–43.