



Predicting the necessity of oxygen therapy in the early stage of COVID-19 using machine learning

Sara Saadatmand¹ · Khodakaram Salimifard¹ · Reza Mohammadi² · Maryam Marzban³ · Ahmad Naghibzadeh-Tahami⁴

Received: 29 July 2021 / Accepted: 1 February 2022 / Published online: 11 February 2022
© International Federation for Medical and Biological Engineering 2022, corrected publication 2022

Abstract

Medical oxygen is a critical element in the treatment process of COVID-19 patients which its shortage impacts the treatment process adversely. This study aims to apply machine learning (ML) to predict the requirement for oxygen-based treatment for hospitalized COVID-19 patients. In the first phase, demographic information, symptoms, and patient's background were extracted from the databases of two local hospitals in Iran, and preprocessing actions were applied. In the second step, the related features were selected. Lastly, five ML models including logistic regression (LR), random forest (RF), XGBoost, C5.0, and neural networks (NNs) were implemented and compared based on their accuracy and capability. Among the variables related to the patient's background, consuming opium due to the high rate of opium users in Iran was considered in the models. Of the 398 patients included in the study, 112 (28.14%) received oxygen-based treatment. Shortness of breath (71.42%), fever (62.5%), and cough (59.82%) had the highest frequency in patients with oxygen requirements. The most important variables for prediction were shortness of breath, cough, age, and fever. For opioid-addicted patients, in addition to the high mortality rate (23.07%), the rate of oxygen-based treatment was twice as high as non-addicted patients. XGBoost and LR obtained the highest area under the curve with values of 88.7% and 88.3%, respectively. For accuracy, LR and NNs achieved the best and same accuracy (86.42%). This approach provides a tool that accurately predicts the need for oxygen in the treatment process of COVID-19 patients and helps hospital resource management.

Keywords COVID-19 · Opioid addiction · Oxygen treatment · Prediction · Machine learning · XGBoost

✉ Khodakaram Salimifard
salimifard@pgu.ac.ir

Sara Saadatmand
sara.saadatmand@mehr.pgu.ac.ir

Reza Mohammadi
a.mohammadi@uva.nl

Maryam Marzban
marzbanh@gmail.com

Ahmad Naghibzadeh-Tahami
anaghibzadeh61@gmail.com

¹ Computational Intelligence and Intelligent Optimization Research Group, Persian Gulf University, 75169 Bushehr, Iran

² Department of Operation Management, Amsterdam Business School, University of Amsterdam, Amsterdam, Netherlands

³ Department of Public Health, School of Public Health, Bushehr University of Medical Science, Bushehr, Iran

⁴ Modeling in Health Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran

1 Introduction

Since the spread of COVID-19 in December 2019, the healthcare sector has played a key role in combating the disease. The World Health Organization (WHO) on March 11, 2020, declared the COVID-19 a pandemic [1]. Hospitals as one of the main players have allocated a large part of their resources to deal with this disease. But, the growing number of patients due to the various variants of disease caused the lack of hospital resources such as ICU beds, medicine, and oxygen. Oxygen is a critical element in the treatment process of COVID-19 patients which according to WHO about 15% of cases require medical oxygen [2]. Decreasing respiratory failures caused by COVID-19 depends on the availability of oxygen and ventilation [3]. India was one of the countries, which lack of medical oxygen influenced its hospital service for COVID-19 patients [4].

To avoid the lack of supplies in hospitals during this pandemic, it is necessary to have an accurate and in-time

prediction from required equipment like oxygen and ventilators. Artificial intelligence (AI) which has been used widely in medicine can detect and learn the non-linear relationship among variables and diagnose, treat, and predict the outcomes [5]. In healthcare, AI assists to realize the unknown patterns in data and make effective decisions accordingly [6]. This field of science has been applied to the COVID-19 pandemic for screening, analyzing, tracking patients, and making medical predictions [7]. Furthermore, machine learning (ML) as a subset of AI is used for computational epidemiology, early detection, diagnosis, and disease progression of COVID-19 disease as well as clinical management issues of this illness such as ICU admission, mechanical ventilation, multi-organ failure, and death [8, 9]. To make valid predictions in medicine, a supervised ML model requires a dataset containing a number of features and a relevant outcome [10]. For COVID-19, these features can be demographic information, symptoms, lab results, and the background of the patient.

From various variables related to COVID-19 patient's background, such as diabetes, cancer, smoking, and kidney and liver diseases [11], the effect of consuming opium on COVID-19 patients' needs more research [12]. Opium is one of the most common and popular drugs among Iranian people, which has been used for more than five centuries. This country has one of the highest rates of opium users in the world, which includes about 2.7% of its population [13]. In 2013 a study on a national scale was designed for evaluating the spread of substance abuse and opium addiction in Iran to zoning the country in low- to high-risk areas. The initial results of the study revealed that Kerman has the highest degree and is the province with the most opium users [14]. A survey on drug abuse in Kerman [15] indicated that the prevalence of substance abuse in the rural areas of Kerman was 22.5% and the rate of addiction was 6%.

As prior mentioned, since the spread of novel Coronavirus, issues and problems related to the capacity of the health system to service the increasing number of patients has grown [16]. One of the basic requirements in the time of pandemic is to accurately predict the required resources and likely outcomes. Among the numerous required resources in the treatment process of COVID-19 patients, medical oxygen is an essential one [17]. Some studies were conducted to apply ML prediction models to predict the requirement for medical oxygen. Lee et.al [18] developed a prediction model that specifies the COVID-19 patients with the risk of requiring medical oxygen. They used the information of 221 patients with C-reactive protein, hypertension, age, and neutrophil and lymphocyte count parameters. Their model achieved a high AUC. To predict the need for mechanical ventilation, [19] used a cohort of 1980 COVID-19 patients. Their data include demographics, patient's background, vital signs at the emergency room, and laboratory data.

Their results demonstrated that age and fever were associated with the risk of ventilator requirement. In another article [20], a machine learning approach was used to predict the mechanical ventilation for COVID-19 patients. The input data included 12 clinical features of 197 COVID-19 patients collected from US hospitals. Their model predicted the mechanical ventilation requirement by applying blood factors and other variables like blood pressure and heart rate.

In this research, a machine learning approach will be proposed to predict the requirement for oxygen-based treatment based on patient characteristics and clinical data. One of the main contributions of this research in compared to previous studies is to predict the outcome (oxygen requirement) in the initial time of patient admission at the hospital by only measuring the symptoms and patient's background without requiring lab results and further information. This can accelerate the process of resource planning, especially in the time of the peak of the disease and avoid shortage occurrences. The second novelty of this article, which is significant from a medical viewpoint, is assessing the impact of using opium on requiring the oxygen-based treatment and fatality rate of COVID-19 cases, using data collected from Kerman that have a high prevalence of opium users in Iran. The results can assist hospitals in forecasting the need for oxygen and managing this source effectively.

2 Data and methods

In this section, the characteristics of the applied dataset and preprocessing operations on raw data will be described. Figure 1 illustrates an overview of the taken steps to build the prediction model. In the first phase, the required data of hospitalized COVID-19 patients were collected from hospitals. Next, the raw data were cleaned and preprocessing operations were applied. In the third phase, the relevant features were selected. In the fourth step, prepared data were split into train and test sets. Then, the train set was used as input for the numbers of machine learning models to train. After model training, the test set was applied to predict the outcomes. The prediction models were compared based on their accuracy and capability.

2.1 Dataset population

Data for this study were collected from two local hospitals in Kerman province in the south of Iran. The data were acquired from the hospital database and written records of 398 hospitalized patients with positive COVID-19 tests (PCR) in a period of 6 months from February to July 2020. The admitted patients' information contained demographic data, patient's background, and symptoms of the disease. The average age of patients was 41.11 years old with the

Fig. 1 Steps of building machine learning approach for predicting oxygen-requirement treatment in hospitalized COVID-19 patients

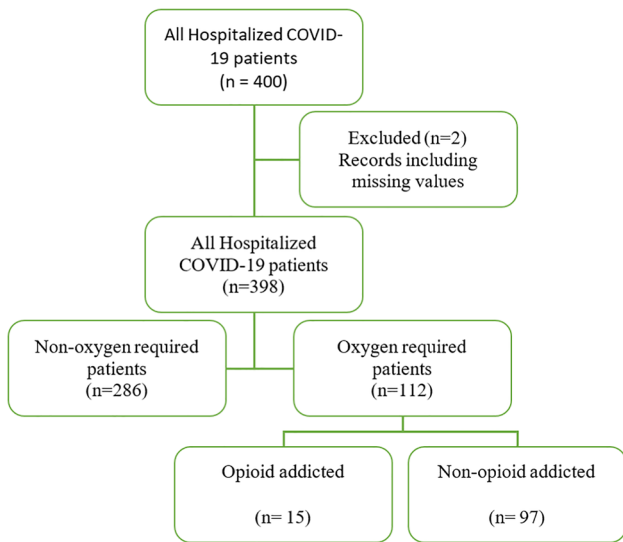
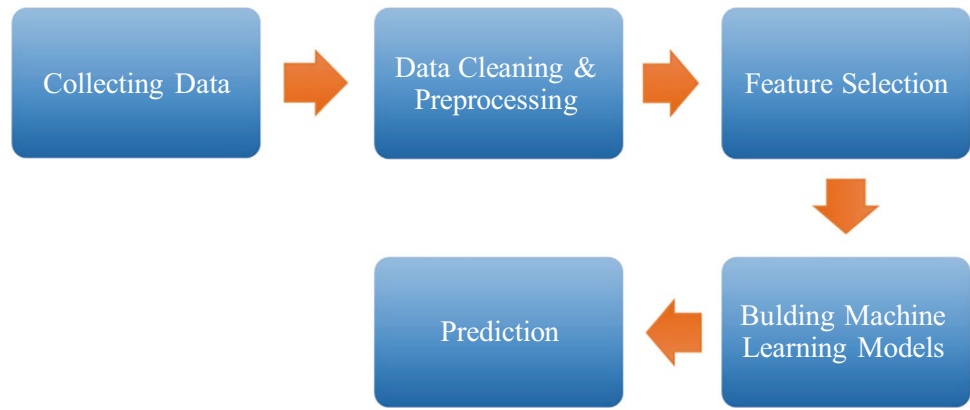


Fig. 2 Flow chart of selecting COVID-19 patients for predicting the necessity of oxygen-based treatment considering opioid and non-opioid addicted cases

median and mode of 39 and 33 years old. The frequency of hospitalized male cases in the dataset was more than female and comprises 54% of the total records. The number of discharged cases and deaths were 377(94.72%) and 21(5.27%), respectively.

The severity of disease, based on the patient’s condition and symptoms, was divided into three categories: mild, moderate, and severe, in which 6% of the patients experienced severe disease conditions, while 94% experienced mild to moderate severity. Patients with mild condition were received only medication, the moderate group received medication and mask oxygen, and the severe cases, besides medication, used ventilators. Figure 2 demonstrates the flowchart of selecting cases for this study. From a total of 400 cases, two records including missing values were excluded. Of the 398 hospitalized COVID-19 patients, 28.18% received oxygen-based treatment which

13.39% of them were opioid-addicted. Non-oxygen-required treatments were applied to 286 patients.

2.2 Data preparation and feature selection

The original dataset included 57 features. At the data cleaning phase, non-required data such as patient ID were removed; also the job variable due to variation and difficulty in job classification was omitted. After specifying the study’s objectives, a consultant with medical specialists was conducted to determine the most relevant characteristics and features. The final features included demographic characteristics (two variables), patient’s background (nine variables), disease symptoms (eight variables), and a target variable (type of treatment). Demographic information comprises gender and age. History of other diseases such as diabetes, blood pressure, and lung disease is in the patient’s background class, and the last group includes initial symptoms of COVID-19 like cough, fever, and shortness of breath.

The type of treatment was divided into two classes: oxygen-required treatment and non-oxygen treatment, which the first group included patients who used oxygen masks and ventilators besides medication, but the second group only received medication. Opium and its extracts and heroin were the four addiction-related variables that for each of them the start age of consumption, the amount of daily consumption, the number of daily usages, and type of use (orally taken or smoked) were collected. To add the opioid addiction variable to the database, these drugs, which are common among addicted people in Kerman, were combined as one binary variable and added to the dataset. Since the data were exclusively collected for scientific purposes, in order to be more accurate, the missing values in the electronic dataset were filled by available data in written records of patients, and only two incomplete medical records were omitted. Except for the age variable, the rest of the variables are binary and no outliers were observed.

2.3 Machine learning models

Five machine learning algorithms including logistic regression, neural networks, decision tree C5.0, random forest, and XGboost were applied to predict the requirement for oxygen-based treatment in COVID-19 patients.

Logistic regression (LR) is one of the qualified models for binary outcomes in such fields as medical science especially in exploring the relationship between risk factors and the incidence of disease [21, 22]. In this paper, a multivariable LR with 19 predictors used to predict the oxygen and non-oxygen treatment for hospitalized patients with positive COVID-19. To fit the LR to the dataset, the iteratively reweighted least squares method was applied [23].

In a neural network (NN) algorithm, which is based on the nervous systems, the neurons represent the nodes in the algorithm that learn from the input data to optimize its final output [24]. The NNs for the purpose of this study were applied using one hidden layer, an output layer, and one input layer including 19 factor variables. The entropy fitting method is used to fit the NNs to the dataset. The maximum number of iterations and the maximum number of weight were set to 100 and 1000, respectively.

C5.0 is the improved version of the C4.5 decision tree algorithm developed by Quinlan [25]. This algorithm is based on the ID3 algorithm which decreases the misclassification errors caused by noise in the training data set [26].

For this algorithm in our prediction model, the boosting iterations were set to ten and the trees decomposed into the rule-based model.

Random forest (RF) [27] is a machine learning method that is normally used for classification and regression. The capability of matching with a wide range of prediction problems and the simplicity of parameter tuning are the two main reasons for the popularity of the RF algorithm [28]. To set the parameters for the proposed RF prediction model, the number of trees and the minimum size of terminal nodes were set to 200 and 1, respectively. The number of variables randomly sampled as candidates at each branch was set to ten.

Implementing the Gradient Boosting concept, the XGBoost, provides a parallel tree boosting to solve a wide range of regression and classification problems fast and accurately. This algorithm applies a more regularized formalization to control over-fitting [29]. In this study, the maximum depth, number of rounds, and subsample ratio of columns for XGBoost were set to 1, 150, and 0.8, respectively. All parameters related to the five ML algorithms are displayed in Table 1.

3 Results

The 19 predictor features were categorized into three classes (as shown in Table 2). Patients in the age category of 19–60 years old consisted 74.37% of all cases and

Table 1 Parameters' values of five applied ML algorithms

Algorithm	Parameter	Value/setting
Logistic regression	Fitting method	Iteratively reweighted least squares
Neural networks	Hidden layer	1
	Input layer	1
	Output layer	1
	Fitting method	Entropy
	Maximum number of iteration	100
C5.0	Maximum number of weight	1000
	Boosting iterations	10
Random forest	Decomposition model	Rule-based
	Number of trees to grow	200
XGboost	Number of variables randomly sampled	10
	Minimum size of terminal nodes	1
	Max depth	1
	Number of rounds	150
	Minimum child weight	1
	Eta	0.3
	Subsample ratio of columns	0.8
	Subsample	0.5

Table 2 Statistics of oxygen and non-oxygen required patients based on model variables

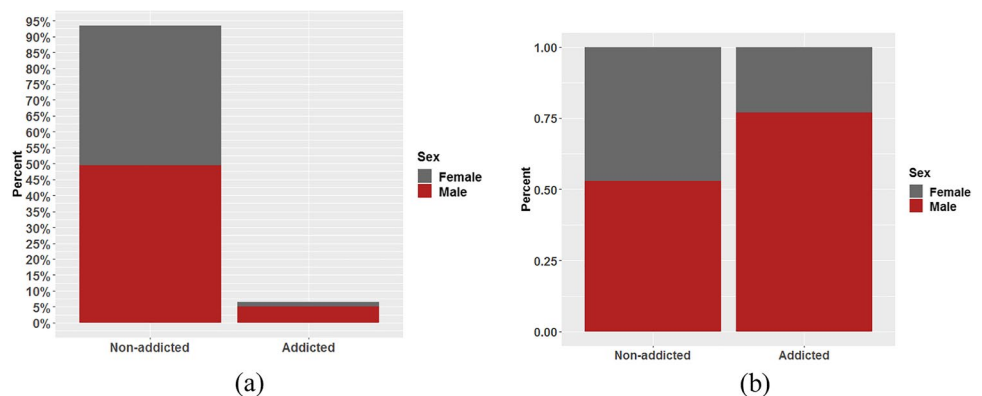
	Variable	All hospitalized patients (n = 398)	Non-oxygen required patients (n = 286)	Oxygen required patients (n = 112)
Demographics	Age < = 18	34 (8.54%)	31 (10.83%)	3 (2.67%)
	19–60	296 (74.37%)	227 (79.37%)	69 (61.60%)
	Age > = 60	68 (17.08%)	28 (9.79%)	40 (35.71%)
	Sex = male	217 (54.52%)	154 (53.84%)	63 (56.25%)
	Sex = female	181 (45.47%)	131 (45.80%)	50 (44.64%)
Patient background	Hypothyroidism or hyperthyroidism	28 (7.03%)	15 (5.24%)	13 (11.60%)
	Diabetes	53 (13.31%)	22 (7.69%)	31 (27.67%)
	Blood pressure	61 (15.32%)	25 (8.74%)	36 (32.14%)
	Obesity	15 (3.76%)	14 (4.89%)	1 (0.89%)
	Lung disease	49 (12.31%)	17 (5.94%)	32 (28.57%)
	Cardiovascular disease	37 (9.29%)	11 (3.84%)	26 (23.21%)
	Acute renal impairment	5 (1.25%)	2 (0.69%)	3 (2.67%)
	Cigarettes	18 (4.52%)	11 (3.84%)	7 (6.25%)
	Opioid addiction	26 (6.53%)	11 (3.84%)	15 (13.39%)
	Symptoms	Fever	197 (49.49%)	127 (44.40%)
Cough		145 (36.43%)	78 (27.27%)	67 (59.82%)
Sore throat		72 (18.09%)	66 (23.07%)	6 (5.35%)
Runny nose		35 (8.79%)	24 (8.39%)	11 (9.82%)
Chest pain		81 (20.35%)	47 (16.43%)	34 (30.35%)
Diarrhea		65 (16.33%)	53 (18.53%)	12 (10.71%)
Shortness of breath		140 (35.17%)	60 (20.97%)	80 (71.42%)
Nausea and vomiting		81 (20.35%)	75 (26.22%)	6 (5.35%)

61.60% of oxygen-required patients. Of the patients, 8.54% were under 18 years old in which 2.67% of them received oxygen-based treatment. Among the patient’s background features, blood pressure, lung disease, and diabetes with ratios of 32.14%, 28.57%, and 27.67%, respectively, had the most frequency in the oxygen-based treatment group. In the symptom category, shortness of breath with the rate of 71.42% had the highest frequency in patients with oxygen requirements. Fever (62.5%) and cough (59.82%) were in second and third places.

3.1 Opioid addiction

In this study, information about the use of opium and its subsequences for hospitalized patients with positive COVID-19 tests was collected. Because of the prevalence of opium consumption in this province, opium-addicted cases included 6.53% of total samples. As it is illustrated in Fig. 3, compared to females, the frequency of opioid addiction is higher in the male group. From the total number of hospitalized patients, 26 individuals were addicted which included 4

Fig. 3 (a) Percentage of opium-addicted and non-addicted hospitalized COVID-19 patients among females and males in two local hospitals in Iran. (b) Normalized percentage of opium-addicted and non-addicted hospitalized COVID-19 patients among female and male



females and 22 males. The average age of women drug users was 65.25 and for men was 60.5 years. The minimum age of initiation of consumption was 15 years, and the maximum was 79 years with a mean of 44.11 years.

A point to consider is the high rate of mortality among opioid-addicted cases. The fatality ratio among non-addicted patients is 4.03%, while for opioid-addicted cases is 23.07%. The survival and death of both groups (addicted and non-addicted) are illustrated in Fig. 4.

To determine the relationship between opioid addiction and oxygen-required treatment, the Chi-square test was conducted. The Chi-square test of independence applies to determine whether there is a relationship between two categorical variables. In this case, the null hypothesis assumes that there is not any relationship between addiction to opioids and the requirement of oxygen, and the alternative hypothesis assumes that there is an association between these two nominal variables. After computing the test statistics, it is found that $p < 0.001$. Hence, the null hypothesis was rejected considering the confidence interval of 95%. It can be concluded that the relationship between these two variables is statistically significant.

3.2 Prediction models

The basic dataset was randomly split into training and test set in a ratio of 80:20 with considering balanced data distribution. Ten-fold cross-validation method was applied to the training set to validate and evaluate the reliability of the developed models. ML algorithms including LR, NNs, C5.0, XGBoost, and RF were applied to build five different prediction models. To compare and evaluate the performance of models, accuracy, receiver operating characteristic (ROC) curves, Cohen's Kappa, balanced accuracy, confusion matrix, and the area under the curve (AUC) were calculated. ROC curve displays the performance of a classifier system when discrimination cut-off value changes over the range of the predictor variable. Higher points above the diagonal line refer to the better predictive value of the test [30].

The accuracy and kappa of applied ML models in the train set are displayed in Fig. 5. For accuracy metric, LR and XGBoost obtained the maximum (90.90%), and RF and NNs with the value of 90.62% were in second place. The mean of accuracy for all models except NNs, which is 78.28%, was just above 80%. The kappa measurement was also calculated for all models. Kappa is a measure of inter-rater agreement [31]. In machine learning, it measures the level of agreement between the true values and the predicted values. The LR algorithm with the kappa of 0.7924 achieved the highest value, and the XGBoost with 0.7785 was in second place. RF and NNs with values of 0.0350 and 0.1428 obtained the minimum kappa, respectively.

Figure 6 illustrates the ROC curves of five ML models in the test set. In comparing the performance of algorithms, XGBoost has the highest AUC followed by LR. NNs, C5.0, and RF was in third, fourth, and fifth positions, respectively. All five models demonstrate a desirable confidence interval result, ranging from 74.1 to 96.5%. The proposed approach was implemented in R software using libraries such as Caret [32], ggplot2 [33], and Liver [34]. The Caret environment consists of various machine learning models like NNs, RF, and LR. The overall runtime of the proposed model was 3.35 min, which makes it an appropriate tool for deciding on rush times. XGBoost had the longest and LR had the shortest runtime among all algorithms. In comparison to other schemes like making decisions based on historical data and previous experience, this approach assists decision-makers to decide more accurately in a shorter time.

Sensitivity and specificity are two statistical performance metrics in using classification models or a diagnostic test. Sensitivity is the ability of the model to predict the true positives, while specificity evaluates the prediction of the true negatives by model [35]. To calculate them, the models' confusion matrices (Fig. 7) were used. Table 3 demonstrates the performance of models, using accuracy, kappa, sensitivity, specificity, and balanced accuracy. LR and NNs

Fig. 4 (a) Fatality percentage for opium-addicted and non-addicted hospitalized COVID-19 patients in two local hospitals in Iran. (b) Normalized Fatality percentage for opium-addicted and non-addicted hospitalized COVID-19 patients

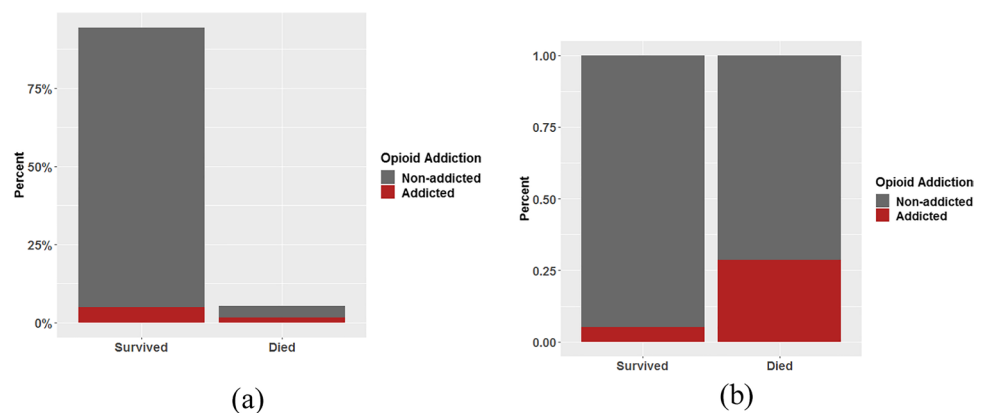


Fig. 5 Comparison of accuracy and kappa of five applied ML algorithms in train set using box plot

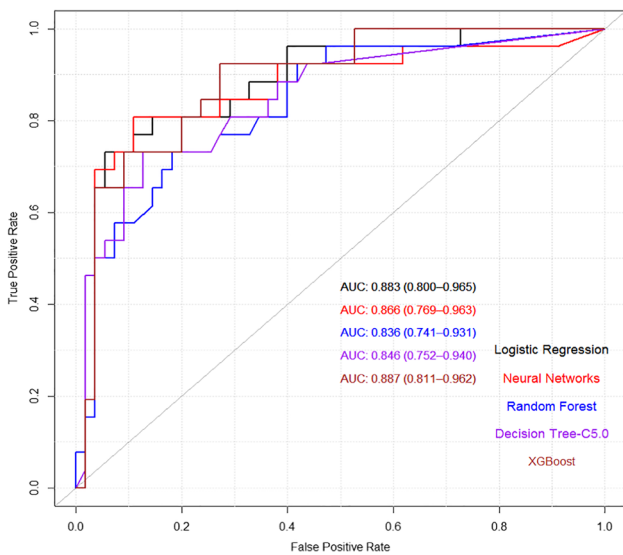
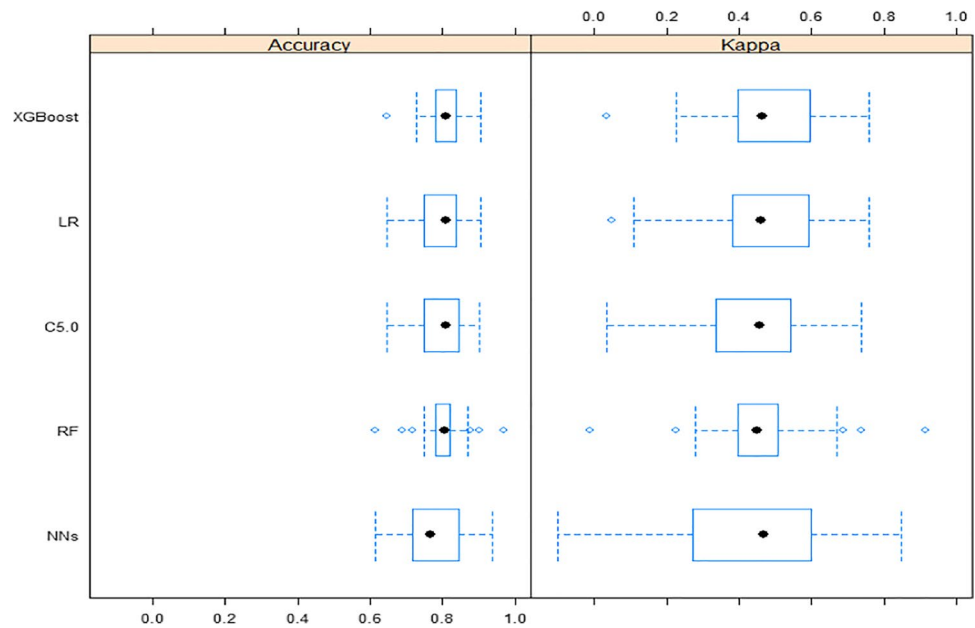


Fig. 6 Receiver operating characteristic (ROC) curves of five applied ML models to predict the oxygen-based treatment for COVID-19 patients. Area under the curve (AUC) and confidence interval are specified for each model

obtained the highest and the same performance in five metrics. Compared to C5.0, RF and XGBoost presented better performance.

In a classification model, each variable has a specific impact on making predictions. Variable importance is a technique that indicates the relative importance of each input variable in a model prediction. The more important a variable, the more a model depends on it to make an accurate prediction [36]. The variable importance can be used

to determine the most and least important variables to the model and improve the model’s performance by dropping ineffective features. In NNs, XGBoost, and RF, age had the most score (100%), while in LR and C5.0, shortness of breath and cough with 100% relative importance were the most effective variables in prediction. LR and C5.0 algorithms have found the variable importance for feature *age*, 59.97% and 91.48%, respectively. Shortness of breath and cough are among the five most important features in four of the classification models (RF, LR, C5.0, NNs). Four variables include age, shortness of breath, cough, and fever which are common in the top five features in NNs, LR, and RF. For the shortness of breath, the relative importance in LR, C5.0, NNs, RF, and XGBoost were 100%, 100%, 82.06%, 42.46%, and 31.88%, respectively. Using C5.0, LR, NNs, RF, and XGBoost, the relative importance of *cough* were 100%, 75.75%, 70.95%, 70.95%, and 14.65%, respectively. Opioid addiction variable with scores of 5.694% in NNs, 32.578% in LR, 1.903% in RF, 92.11% in C5.0, and 0.8475% in XGBoost showed different behaviors in each model.

4 Discussion

Oxygen therapy is one of the main treatment choices for COVID-19 patients which reduces the fatality rate among critical cases [37, 38]. In our proposed approach, four features including shortness of breath, cough, fever, and age were identified as the most important variables in predicting the requirement for oxygen-based treatment in the early stages of admission.

Fig. 7 Confusion matrices of five ML models (a-e). Each plot represents the true positive (sensitivity) and true negatives (specificity) for predicting oxygen-based treatment for hospitalized COVID-19 patients

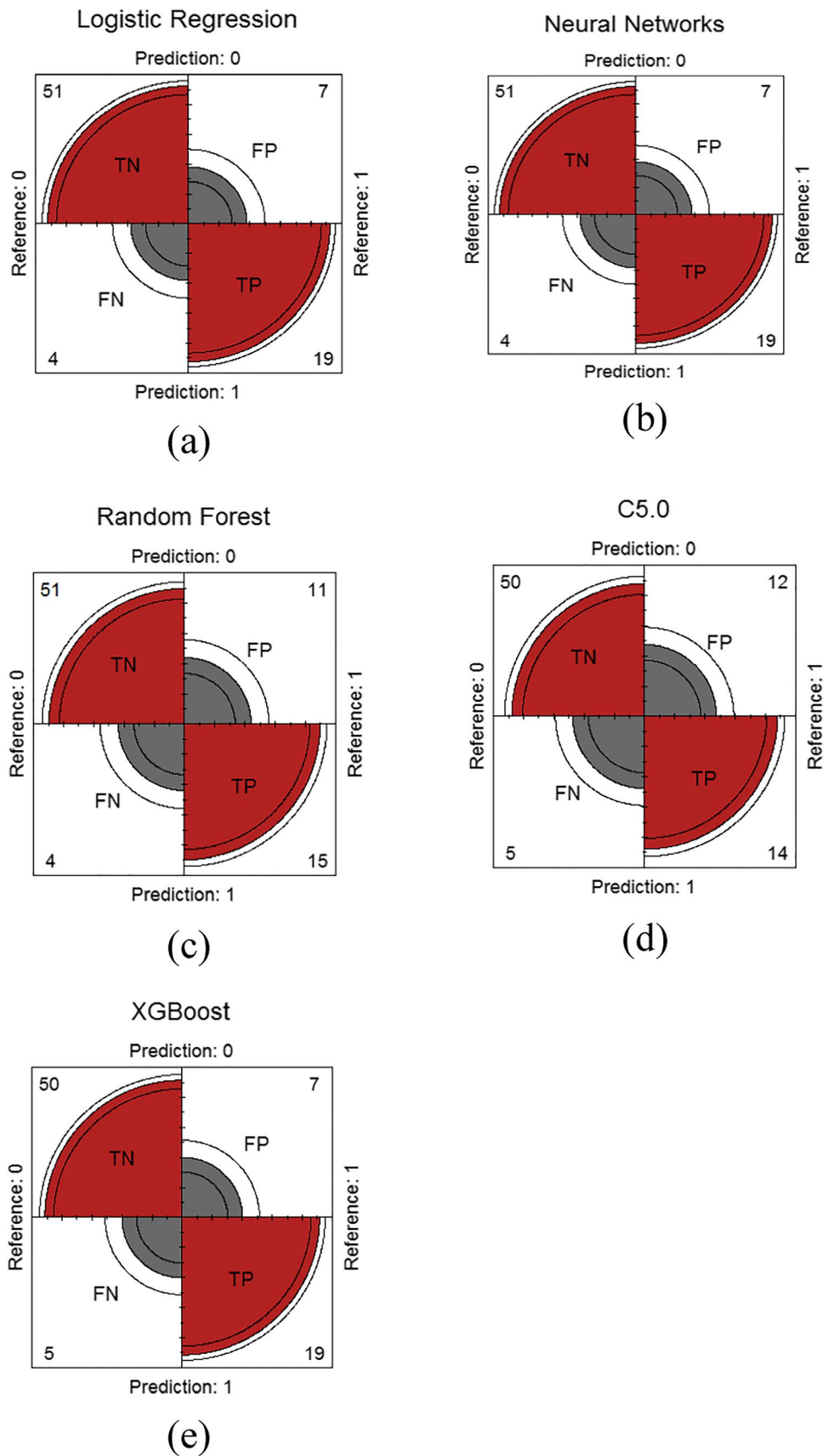


Table 3 Performance measures of the five ML models in test set

Model	Accuracy	Kappa	Sensitivity	Specificity	Balanced accuracy
Logistic regression	0.8642	0.6787	0.9273	0.7308	0.8290
Neural networks	0.8642	0.6787	0.9273	0.7308	0.8290
Random forest	0.8148	0.5427	0.9273	0.5769	0.7521
DecisionTree-C5.0	0.7901	0.4817	0.9091	0.5385	0.7238
XGBoost	0.8519	0.6531	0.9091	0.7308	0.8199

The association of shortness of breath and cough with receiving oxygen-based treatment has been addressed in various studies, which are the same as our results. Long et al. [38] analyzed the clinical information of 1362 COVID-19 patients of a local hospital in Wuhan. They found that most of the patients who experienced breathlessness, like shortness of breath, dyspnea, and chest tightness, received oxygen therapy. In another study in Ethiopia [39], the longer duration of supplemental oxygen requirement was associated with shortness of breath. They also found that compared to patients without this symptom on admission, the degree of ending oxygen therapy was 29.5% lower in cases with shortness of breath. Ni et.al [40] concluded that dyspnea is among the related factors to oxygen therapy for COVID-19 patients under 65 years and can increase their need for oxygen. They found that 59.5% of patients with dry cough received oxygen therapy.

Fever is one of the most common symptoms in patients with COVID-19 [41]. According to [42], 43.8% of COVID-19 patients on admission and 88.7% during the hospitalization experienced fever. In our ML models, fever was an important feature in the prediction of oxygen requirement that can be considered as an early sign on admission. [40] found the relationship between fever and oxygen therapy. Among COVID-19 patients, 70.9% of those with fever symptom received oxygen therapy.

It is shown that as the patient’s age increases, the severity of COVID-19 cases increases [43] and also the risk of in-hospital death [44]. In our algorithms, age was an effective factor in the prediction of oxygen-based treatment. Also, [39] recognized age as an important factor in the starting time of oxygen therapy, which is related to the longer duration of oxygen requirement among COVID-19 patients.

The effect of consuming opium on the requirement for oxygen-based treatment was analyzed in this study. The data were gathered from Kerman province in the south-east of Iran, which has a high number of opium consumers [45]. This research had the opportunity to evaluate the impact of opioid addiction in the prediction model. About 58% of patients addicted to opioid received oxygen-based treatment including ventilator and oxygen mask. This rate for non-addicted individuals was much less (26.34%). Additionally, compared to non-addicted patients, the fatality rate among opium-addicted cases was high (28.57%). It proves

the previous claim [46] that there is a higher death rate for COVID-19 opium users’ patients. It is, probably, due to the negative impact of opium on the immune system and respiratory cells.

There were several limitations to this study. First, the sample size of COVID-19 patients was small, especially for oxygen therapy and mechanical ventilation. Second, the data were collected from two hospitals in one province, which may influence the model reliability due to the variability in symptoms and other factors of disease between different populations. Not defining the exact time of needing oxygen-based requirement for each patient in the prediction model was the third limitation. It is due to the fact that data were limited. Fourth, available features were limited to the patient’s background and symptoms with no information related to lab results.

Future research can consider more features like lab results, vital signs of the patient, and CT images. In addition, a larger set of data needs to be used to build models that are more reliable. Apart from oxygen, the need for other COVID-19-related supplies such as medications and beds can be predicted. Further studies also may try to collect the data, based on a specific time interval to specify the demand time of supplies and equipment.

5 Conclusion

In this study, information of hospitalized COVID-19 patients from two local hospitals in Iran was applied to predict the requirements for oxygen-based treatment. First, relevant attributes were selected based on experts’ opinions, and then the model performed five ML classifications to predict oxygen requirement. The proposed approach found that the most important variables in predicting the need for oxygen therapy were age, shortness of breath, cough, and fever. One of the main objectives of this research was to predict the oxygen-based treatment in the early stages of patient admission, which, according to the results, the model indicated high accuracy and sensitivity in predicting the outcome. Among five ML algorithms, NNs and LR achieved high sensitivity (0.9273) and specificity (0.7308) that demonstrate their capability in predicting the need for oxygen-based treatment for COVID-19 patients. XGBoost showed

the highest AUC (0.887). Another aim was to analyze the effect of consuming opium on the requirement for oxygen in COVID-19 cases. The results revealed the high rate of the requirement to this medical resource and high fatality ratio in this group of patients compared to other cases.

In conclusion, the availability of medical resources especially in times of pandemic and the peak of the number of infected is an essential issue in managing hospital resources. Artificial intelligence tools like ML can help to accurately predict the need for medical supplies such as oxygen and avoid shortages.

Acknowledgements The authors would like to appreciate the anonymous reviewers' valuable and profound comments on an earlier version of the manuscript that have resulted in significant improvements to the article.

Data availability Due to the sensitive nature of data used in this study, the hospital authority was assured raw data should remain confidential and should not be shared.

Declarations

Ethics approval This study was approved by the Ethical committee of Sirjan University of Medical Science in Iran (IR.SIRUMS.REC.1399.008). The study was conducted in accordance with the ethical standards of the Helsinki declaration [47].

Conflict of interests The authors declare no competing interests.

References

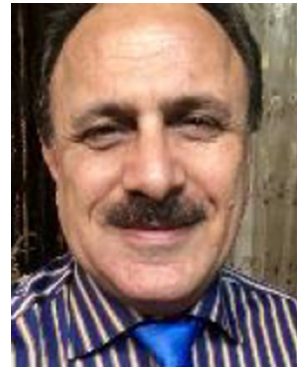
1. WHO, "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," WHO, 2020. <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>.
2. WHO, "Oxygen sources and distribution for COVID-19 treatment centres." [Online]. Available: <https://apps.who.int/iris/rest/bitstreams/1274720/retrieve>.
3. Stein F, Perry M, Banda G, Woolhouse M, Mutapi F (2020) Oxygen provision to fight COVID-19 in sub-Saharan Africa. *BMJ Glob Heal* 5(6):e002786. <https://doi.org/10.1136/bmjgh-2020-002786>
4. J. Wise, "Covid-19: Countries rally to support India through 'storm that has shaken the nation,'" *BMJ*, p. n1086, Apr. 2021 <https://doi.org/10.1136/bmj.n1086>.
5. Chan Y-K, Chen Y-F, Pham T, Chang W, Hsieh M-Y (2018) Artificial Intelligence in Medical Applications. *J Healthc Eng* 2018:1–2. <https://doi.org/10.1155/2018/4827875>
6. Islam MM et al (2021) Application of Artificial Intelligence in COVID-19 Pandemic: Bibliometric Analysis. *Healthcare* 9(4):441. <https://doi.org/10.3390/healthcare9040441>
7. Vaishya R, Javaid M, Khan IH, Haleem A (2020) Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes Metab Syndr Clin Res Rev* 14(4):337–339. <https://doi.org/10.1016/j.dsx.2020.04.012>
8. Syeda HB et al (2021) Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review. *JMIR Med Informatics* 9(1):e23811. <https://doi.org/10.2196/23811>
9. Chee ML et al (2021) Artificial Intelligence Applications for COVID-19 in Intensive Care and Emergency Settings: A Systematic Review. *Int J Environ Res Public Health* 18(9):4749. <https://doi.org/10.3390/ijerph18094749>
10. Sidey-Gibbons JAM, Sidey-Gibbons CJ (2019) Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 19(1):64. <https://doi.org/10.1186/s12874-019-0681-4>
11. CDC (2021) People with Certain Medical Conditions, Centers for Diseases Control and Prevention. <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html>. Accessed 1 Nov 2021
12. M. Saedi et al., "Opium Addiction and COVID-19: Truth or False Beliefs," *Iran. J. Psychiatry Behav. Sci.*, vol. 14, no. 2, Apr. 2020, <https://doi.org/10.5812/ijpbs.103509>.
13. Moradinazar M et al (2020) Prevalence of drug use, alcohol consumption, cigarette smoking and measure of socioeconomic-related inequalities of drug use among Iranian people: findings from a national survey. *Subst Abuse Treat Prev Policy* 15(1):39. <https://doi.org/10.1186/s13011-020-00279-1>
14. Mohebbi E et al (2018) Awareness and Attitude Towards Opioid and Stimulant Use and Lifetime Prevalence of the Drugs: A Study in 5 Large Cities of Iran. *Int J Heal Policy Manag* 8(4):222–232. <https://doi.org/10.15171/ijhpm.2018.128>
15. H. Ziaaddini and . M. R. Z., "The Household Survey of Drug Abuse in Kerman, Iran," *J. Appl. Sci.*, vol. 5, no. 2, pp. 380–382 2005, <https://doi.org/10.3923/jas.2005.380.382>.
16. Piccialli F, di Cola VS, Giampaolo F, Cuomo S (2021) The Role of Artificial Intelligence in Fighting the COVID-19 Pandemic. *Inf Syst Front*. <https://doi.org/10.1007/s10796-021-10131-x>
17. WHO (2021) WHO's Science in 5 on COVID-19: Episode #13 Medical oxygen, World Health Organization. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/media-resources/science-in-5/episode-33---medical-oxygen>. Accessed 9 Apr 2021
18. Lee EE et al (2021) Predication of oxygen requirement in COVID-19 patients using dynamic change of inflammatory markers: CRP, hypertension, age, neutrophil and lymphocyte (CHANeL). *Sci Rep* 11(1):13026. <https://doi.org/10.1038/s41598-021-92418-2>
19. Yu L et al (2021) Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19. *PLoS ONE* 16(4):e0249285. <https://doi.org/10.1371/journal.pone.0249285>
20. Burdick H et al (2020) Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. *Comput Biol Med* 124:103949. <https://doi.org/10.1016/j.compbiomed.2020.103949>
21. Stoltzfus JC (2011) Logistic Regression: A Brief Primer. *Acad Emerg Med* 18(10):1099–1104. <https://doi.org/10.1111/j.1553-2712.2011.01185.x>
22. Nick TG, Campbell KM (2007) Logistic regression. In: Ambrosius WT (eds) *Topics in Biostatistics. Methods in Molecular Biology*, vol 404. Humana Press. https://doi.org/10.1007/978-1-59745-530-5_14
23. Wolke R, Schwetlick H (1988) Iteratively Reweighted Least Squares: Algorithms, Convergence Analysis, and Numerical Comparisons. *SIAM J Sci Stat Comput* 9(5):907–921. <https://doi.org/10.1137/0909062>
24. K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Nov. 2015, [Online]. Available: <http://arxiv.org/abs/1511.08458>.
25. Salzberg SL (1994) *C4.5: Programs for Machine Learning* by J. Ross Quinlan. Morgan Kaufmann Publishers Inc, 1993. *Mach Learn* 16(3):235–240. <https://doi.org/10.1007/BF00993309>
26. Batra M, Agrawal R (2018) Comparative analysis of decision tree algorithms. In: Panigrahi B, Hoda M, Sharma V, Goel S (eds) *Nature Inspired Computing. Advances in Intelligent Systems and*

- Computing, vol 652. Springer, Singapore. https://doi.org/10.1007/978-981-10-6747-1_4
27. Tin Kam Ho, “Random decision forests,” in Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282, <https://doi.org/10.1109/ICDAR.1995.598994>.
 28. Biau G, Scornet E (2016) A random forest guided tour. *TEST* 25(2):197–227. <https://doi.org/10.1007/s11749-016-0481-7>
 29. T. Chen and C. Guestrin, “XGBoost,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, pp. 785–794 <https://doi.org/10.1145/2939672.2939785>.
 30. Yang S, Berdine G (2017) The receiver operating characteristic (ROC) curve. *Southwest Respir Crit Care Chronicles* 5(19):34. <https://doi.org/10.12746/swrccc.v5i19.391>
 31. Cohen J (1960) A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* 20(1):37–46. <https://doi.org/10.1177/001316446002000104>
 32. M. Kuhn, “Building Predictive Models in R Using the caret Package,” *J. Stat. Softw.*, vol. 28, no. 5, 2008 <https://doi.org/10.18637/jss.v028.i05>.
 33. H. Wickham, *ggplot2*. New York, NY: Springer New York, 2009.
 34. Mohammadi R, Burke K (2021) Liver package: eating the liver of data science (version 1.10), CRAN. <https://cran.rproject.org/web/packages/liver/>
 35. E. Martin et al., “Sensitivity and Specificity,” in *Encyclopedia of Machine Learning*, Boston, MA: Springer US, 2011, pp. 901–902.
 36. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. New York, NY: Springer New York, 2013.
 37. Jin Y-H et al (2020) A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version). *Mil Med Res* 7(1):4. <https://doi.org/10.1186/s40779-020-0233-6>
 38. Long L et al (2021) Effect of early oxygen therapy and antiviral treatment on disease progression in patients with COVID-19: A retrospective study of medical charts in China. *PLoS Negl Trop Dis* 15(1):e0009051. <https://doi.org/10.1371/journal.pntd.0009051>
 39. T. B. J. Tigist W. Leulseged, Ishmael S. Hassen, Mesay G. Edo, Daniel S. Abebe, Endalkachew H. Maru, Wuletaw C. Zewde, Nigat W. Chamiso, “Duration of Supplemental Oxygen Requirement and Predictors in Severe COVID-19 Patients in Ethiopia: A Survival Analysis,” 2021. <https://doi.org/10.1101/2020.10.08.20209122>.
 40. Ni Y-N, Wang T, Liang B, Liang Z-A (2021) The independent factors associated with oxygen therapy in COVID-19 patients under 65 years old. *PLoS ONE* 16(1):e0245690. <https://doi.org/10.1371/journal.pone.0245690>
 41. WHO (2021) Coronavirus disease (COVID-19), World Health Organization. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19>. Accessed 11 Nov 2021
 42. Guan W et al (2020) Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* 382(18):1708–1720. <https://doi.org/10.1056/NEJMoa2002032>
 43. Davies NG et al (2020) Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat Med* 26(8):1205–1211. <https://doi.org/10.1038/s41591-020-0962-9>
 44. Bahl A et al (2020) Early predictors of in-hospital mortality in patients with COVID-19 in a large American cohort. *Intern Emerg Med* 15(8):1485–1499. <https://doi.org/10.1007/s11739-020-02509-7>
 45. Naghibzadeh-Tahami A et al (2020) Is opium use associated with an increased risk of lung cancer? A case-control study. *BMC Cancer* 20(1):807. <https://doi.org/10.1186/s12885-020-07296-0>
 46. Dolati-Somarin A, Abd-Nikfarjam B (2021) The Reasons for Higher Mortality Rate in Opium Addicted Patients with COVID-19: A Narrative Review. *Iran J Public Health*. <https://doi.org/10.18502/ijph.v50i3.5587>
 47. “World Medical Association Declaration of Helsinki,” *JAMA*, vol. 310, no. 20, p. 2191, Nov. 2013, <https://doi.org/10.1001/jama.2013.281053>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Sara Saadatmand is a researcher in operations research and member of CHIORG at Persian Gulf University. Her research interests include machine learning and intelligent optimization.



Khodakaram Salimifard is Associate Professor of Operations Research and leader of CHIORG at Persian Gulf University. His research interests include computational intelligence & intelligent algorithms.



Reza Mohammadi is an Assistant Professor of Statistic at the Business School, University of Amsterdam. His research interests include business analytics and machine learning.



Maryam Marzban is an Assistant Professor of Epidemiology at Bushehr University of Medical Science. Her research interests include cancer and HIV/AIDS.



Ahmad Naghibzadeh-Tahami is PhD candidate in Epidemiology at School of Public Health of Kerman University of Medical Sciences. His research interests include cancer and high-risk behaviors.