Patient Safety & Quality Improvement Journal

http://psj.mums.ac.ir

PSQI

# Potential Pathological, Clinical, and Symptomatic Findings of COVID-19 to Predict Mortality in Positive PCR Individuals Using Data Mining

Nasrin Talkhi[1,2], Nooshin Akbari sharak[1], Zahra Pasdar[3], Maryam Salari[4], Seyed Masoud Sadati[5], *Mohammad Taghi Shakeri[6]

*1. Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.*
*2. International UNESCO Center for Health-Related Basic Sciences and Human Nutrition, Mashhad University of Medical Sciences, Mashhad, Iran.*
*3. Institute of Applied Health Sciences, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen.*
*4. Assistant Professor in Biostatistics, Expert Management and Information Technology, Mashhad University of Medical Sciences, Mashhad, Iran.*
*5. Center of Statistics and Information Technology Management, Imam Reza Hospital, Mashhad University of Medical Sciences, Mashhad, Iran.*
*6. Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.*

**A R T I C L E I N F O**

***Article type:***
Original Article

***Article History*:**
**Received:** 15 Feb 2023
**Accepted:** 25 Mar 2023

***Key words:***
Conditional Inference Tree, COVID-19, Data Mining, Decision Trees, Machine Learning.

**A B S T R A C T**

***Introduction:***
COVID-19 has placed immense burdens on healthcare systems and medical staff. To avoid spread, the statistician's role and the use of appropriate predictive models -prediction of survivors versus non-survivors- is highly relevant. This study aimed to apply a model which avoids overfitting and selection bias towards selecting predictors to predict COVID-19 mortality.

***Materials and Methods:***
The Conditional Inference Tree (CIT) model was used. Data from 59,564 hospitalized individuals with positive polymerase chain reaction (PCR) test results were collected from February 20, 2020, to September 12, 2021, in the Khorasan Razavi province, Iran.

***Results:***
The sensitivity and specificity of the model were 88.7% and 88.1%, respectively, the accuracy was 88.2%, and the area under the curve (AUC) was 73.0% on test data. Therefore, the model had considerable accuracy in prediction. The potential predictors involved in predicting survivors versus non-survivors were intubation, age, PO2 level, decreased consciousness level, presence of distress, anorexia, drug use, and kidney diseases.

***Conclusion:***
According to the findings, the CIT model showed high accuracy by avoiding overfitting and selection bias toward selecting predictors. Thus, the results of this study and the efforts of healthcare systems to stop the spread of this pandemic prove helpful.

► ***Please cite this paper as:***
*Talkhi N, Akbari sharak N, Pasdar Z, Salari M, Sadati SM, *Shakeri MT. Potential Pathological, Clinical, and Symptomatic Findings of COVID-19 to Predict Mortality in Positive PCR Individuals Using Data Mining. Journal of Patient Safety and Quality Improvement. 2022; 11(1):13-21. Doi: 10.22038/PSJ.2023.70741.1390*

**** Corresponding Author:***
*Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.*
*E-mail: ShakeriMT@mums.ac.ir*

## Introduction

The unprecedented outbreak of the novel coronavirus, SARS-CoV-2, which epi-centered in the Hubei Province of the People's Republic of China, has ever since spread to numerous countries worldwide, making it a pandemic, as declared by the World Health Organization (WHO) (1,2). The WHO has announced that the disease will persist long (3).

The first defined cases of coronavirus disease 2019 (COVID-19) in Iran were announced on the 19th of February, 2021. After a short period, COVID-19 spread widely to all other provinces in Iran (4). Most infected individuals present with mild symptoms, whereas others may require a ventilator; or die quickly. Many infections of COVID-19 are asymptomatic, but the infected individual can still transmit the virus to others. It makes it difficult to accurately diagnose the disease based on clinical symptoms alone (5). COVID-19 causes respiratory diseases with many symptoms, including but not limited to cold symptoms, upper or lower respiratory tract symptoms, cough, fever, shortness of breath, muscular pain, and even death (6). The mortality rate in the elderly and high-risk groups, including individuals with cardiovascular disease, diabetes, chronic respiratory disease, and hypertension (HTN), is significantly higher than in healthy individuals (7).

Despite various interventions to manage the pandemic, the disease continues to spread and cause death in many people (8). Early diagnosis is critical. Recent studies suggest that delays in the diagnosis of COVID-19 may delay treatment and even lead to death (6).

It is high infectivity places immense burdens on healthcare systems and medical staff. To avoid the spread, the statistician's role and the use of appropriate predictive models -prediction of survivors versus non-survivors- can be very useful.

Nowadays, machine learning techniques have been applied to disease diagnosis and have yielded successful results (9). Machine learning is a subarea of artificial intelligence in computer science, and the main goal is to perform pattern detection in databases (6,10-12). On the other hand, Tree-based models in machine learning have been used to diagnose problems and for prediction (3).

A Decision Tree (DT) uses IF-THEN rules to interpret results. For this reason, it is one of the most popular classification methods. Using the hierarchical structure in the DT, interpreting the concepts is understandable and easy for humans. Many studies have used the DT to analyze COVID-19 data. Two studies showed that the DT and Naive Bayes methods performed better than other methods in determining COVID-19 in the patients based on their symptoms (13). Another study worked on the prediction of diagnosis and prognosis of COVID-19 using DT (14). Van Pelt et al. considered five strategies and analyzed data with COVID-19; one was classifying students with symptoms as having COVID-19 (15). To classify the symptoms, Rochmawati et al. used J48 and Hoeffding Tree (16). One study proposed clinical indicators for efficient screening and testing for COVID-19 infection by Classification and regression trees (CART) (17). In our research process, we found studies by Mesenburg et al. (18) and Venturini et al. (12). Mesenburg et al. just used the CIT model based on a limited number of variables, including sex, age, skin color/ethnicity, and being Indigenous on 18,000 instances. Venturini et al. compared a traditional decision trees model and the CIT model, and their results showed that the CIT model seems robust enough to derive a predictive model on COVID-19 data.

To the best of our knowledge, most conducted research and studies did not focus on Conditional Inference Trees (CIT). The CART model is susceptible to overfitting and selection bias toward selecting inputs or predictors. The CIT model has been proposed to solve these problems (19-22). Revealing potential predictors of a diagnosis of death is essential. Therefore, in the current study, we were interested in identifying the potential pathological, clinical, and symptomatic characteristics which can predict death among COVID-19 patients with positive PCR results and studying the importance of identified potential characteristics using the CIT model on a large population.

## *Materials and Methods*

Data were collected from questionnaires between February 20, 2020, and September 12, 2021. Necessary data and information of subjects were extracted from questionnaires filled out by the nurses and registered in Medical Care Monitoring Center (MCMC) database. In this study, all subjects referred to the hospital with symptoms of COVID-19 and, according to the doctor's diagnosis and the patient's condition, needed hospitalization were included. Due to the presence of the holy shrine of Imam Reza, there are many pilgrims in Khorasan Razavi from almost all Iranian cities, as well as some regional Muslim countries. The data cleaning process was done before performing analyses. All individuals were tested by PCR method. A comprehensive review of the data was done. After cleaning the inaccurate, irrelevant, and incomplete data, a dataset with 33 variables and 59,564 instances that had positive PCR test results remained. The registered variables include age, sex, presence of fever, cough, muscular pain, respiratory distress, decreased consciousness, decreased sense of smell, decreased sense of taste, convulsions, headache, confusion, chest pain, skin inflammation, stomachache, nausea, vomit, diarrhea, anorexia, smoking status, drug use, intubation, cancer, liver diseases, diabetes, blood diseases, Immunodeficiency, heart disease, kidney diseases, asthma, neurological disorders diseases, hypertension, partial pressure of oxygen (PO2). The dependent variable is a dichotomous variable with Survivor and non-survivor levels. Fever was defined as a temperature > $37^{\circ}C$. The location of the study area is depicted in Figure 1.

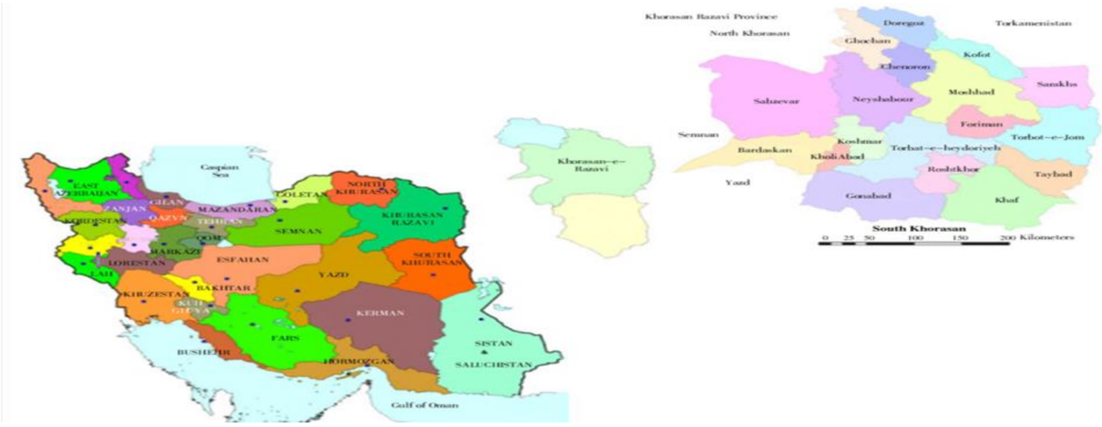


**Figure 1:** Location of the Khorasan-Razavi province and its cities in Iran.

Data processing, calculations, and analysis were performed using R statistical software version 4.1.1. The DT model is built on training data in the group with positive PCR test results (PCR+ group). The association between outcomes and predictors was assessed in the first step. The Kolmogorov–Smirnov test was used t check the normality assumption of age. Chi-square tests were used to study the association between two qualitative variables. A $P$-value of $< 0.05$ indicated a statistically significant difference.

Tree-based models are used in machine learning for classification and regression problems, and they use decision trees that are a framework for creating trees with a series of if-then rules (23). These rules are applied to represent the roles of the different input variables in predicting the target variable. The predictions are usually generated from one or more recursive decision trees (23).

In machine learning, tree-based models are very popular models. They have some advantages, such as (I) their interpretation and understanding are simple, (II) they handle both categorical and continuous variables as a target, and (III) they have good performance for large data (19).

CART models are a popular type of recursive partitioning method. Despite the widespread use and advantages of the tree-based model, they have some drawbacks. These include: (I) overfitting is common in decision tree models, including CART models, and this leads to poor predictions, and (II) with changes in the data, the results will not be stable (19,23). Thus, the CART model is susceptible to overfitting and selection bias toward selecting inputs or predictors. As a solution to these problems, the CIT model has been used (19-22). Hothorn et al. (2006) have previously described the CIT model in detail in reference (24).

CITs are a useful tool for understanding and predicting. For this reason, CITs are more appropriate for diagnostic purposes than specific recursive splitting procedures implemented in CART models. To communicate to practitioners, the resulting tree models are easier. Generally, due to the previously mentioned useful characteristics, CITs were chosen as the modelling tool for analyzing health conditions (20). After modeling, the confusion matrix and the relevant extracted indices assessed the model's accuracy, such as Sensitivity, Specificity, the Area Under the Curve (AUC), and the Receiver Operating Characteristics (ROC) curve.

## Results

During the pre-processing phase, cleaning, recoding, and selecting inputs with valid values were performed. The data comprised 29,283 (49.2%) females and 30,281 (50.8%) males. The mean ± SD for age was 55.84±19.00. The initial associations with the outcomes are reported in Table 1. Most of the predictors showed a significant difference.

**Table 1.** Descriptive statistics for characteristics and symptoms of the study population

| Variables [Reference level] | Positive PCR test (n=59,564) | | | | p-value |
|---|---|---|---|---|---|
| | Survivor | | Non-survivors | | |
| Age | 52.93±18.62 | | 67.15±16.01 | | <0.001 |
| **Levels of variables** | **Yes** | **No** | **Yes** | **No** | |
| Sex [Female] | 23366(49.4) | 23974(50.6) | 6915(56.6) | 5309(43.4) | <0.001 |
| Fever [No] | 16947(35.8) | 30393(64.2) | 3875(31.7) | 8349(68.3) | <0.001 |
| Cough [No] | 21718(45.9) | 25622(54.1) | 4249(34.8) | 7975(65.2) | <0.001 |
| Muscular pain [No] | 12081(25.5) | 35259(74.5) | 2134(17.5) | 10090(82.5) | <0.001 |
| Distress [No] | 28000(59.1) | 19340(40.9) | 9601(78.5) | 2623(21.5) | <0.001 |
| decreased consciousness [No] | 1250(2.6) | 46090(97.4) | 1960(16.0) | 1026484.0() | <0.001 |
| Decreased sense of smell [No] | 687(1.5) | 46493(98.5) | 63(0.5) | 12143(99.5) | <0.001 |
| Decreased sense of taste [No] | 414(0.9) | 46696(99.1) | 54(0.4) | 12132(99.6) | 0.006 |
| Convulsions [No] | 139(0.3) | 46971(99.7) | 29(0.2) | 12157(99.8) | 0.291 |
| Headache [No] | 2333(10.7) | 19511(89.3) | 130(2.6) | 4893(97.4) | <0.001 |
| Confusion [No] | 1271(2.8) | 44887(97.2) | 234(2.0) | 11613(98.0) | <0.001 |
| Chest pain [No] | 1427(3.1) | 44731(96.9) | 302(2.5) | 11545(97.5) | 0.002 |
| Skin inflammation [No] | 48(0.1) | 46110(99.9) | 11(0.1) | 11836(99.9) | 0.734 |
| Stomachache [No] | 1093(2.4) | 45199(97.6) | 146(1.2) | 11767(98.8) | <0.001 |
| Nausea [No] | 3425(7.4) | 42867(92.6) | 598(5.0) | 11315(95.0) | <0.001 |
| Vomit [No] | 1741(3.8) | 44551(96.2) | 335(2.8) | 11578(97.2) | <0.001 |
| Diarrhea [No] | 1265(2.7) | 45027(97.3) | 221(1.9) | 11692(98.1) | 0.008 |
| Anorexia [No] | 5250(11.3) | 41042(88.7) | 1267(10.6) | 10646(89.4) | 0.029 |
| Smoking status [No] | 767(1.6) | 46413(98.4) | 239(2.0) | 11967(98.0) | 0.011 |
| Drug use [No] | 979(2.1) | 46201(97.9) | 431(3.5) | 11775(96.5) | <0.001 |
| Intubation [No] | 995(2.1) | 46345(97.9) | 6176(50.5) | 6048(49.5) | <0.001 |
| PO2 [More than 93 %] | 25940(54.8) | 21400(45.2) | 10609(86.8) | 1615(13.2) | <0.001 |
| Cancer [No] | 367(0.8) | 46973(99.2) | 314(2.6) | 11910(97.4) | <0.001 |
| Liver disease [No] | 191(0.4) | 47149(99.6) | 86(0.7) | 12138(99.3) | <0.001 |
| Diabetes [No] | 5662(12.0) | 41678(88.0) | 2440(20.0) | 9784(80.0) | <0.001 |
| Blood diseases [No] | 133(0.3) | 47207(99.7) | 83(0.7) | 12141(99.3) | <0.001 |
| Immunodeficiency [No] | 45(0.1) | 47295(99.9) | 23(0.2) | 12201(99.8) | 0.007 |
| Heart disease [No] | 3876(8.2) | 43464(91.8) | 1904(15.6) | 10320(84.4) | <0.001 |
| Kidney disease [No] | 488(1.0) | 46852(99.0) | 357(2.9) | 11867(97.1) | <0.001 |
| Asthma [No] | 762(1.6) | 46578(98.4) | 261(2.1) | 11963(97.9) | <0.001 |
| Neurological disease [No] | 490(1.0) | 46850(99.0) | 251(2.1) | 11973(97.9) | <0.001 |
| HTN [No] | 7905(16.8) | 39275(83.2) | 3164(25.9) | 9042(74.1) | <0.001 |
| Abbreviation: HTN, hypertension; Distress, respiratory distress. Data were reported as mean ± standard deviation and n (%). The Mann-Whitney U and Chi-square tests were used. | | | | | |

The CIT model was constructed using a 5-Fold cross-validation with 90% training and 10% testing dataset. The final created tree was pruned and visualized in Figure 2. Next, models were evaluated by the extracted indices from the confusion matrix. The decision tree rules follow a general structure, including IF-THEN statements, semantically resembling how we think. This general structure expresses that IF some conditions are met, THEN a certain prediction occurs. Predictions provide from intelligible features. It should be noted that the ">0" and "<0" symbols in the extracted rules indicate the presence and absence of the desired symptom, respectively.

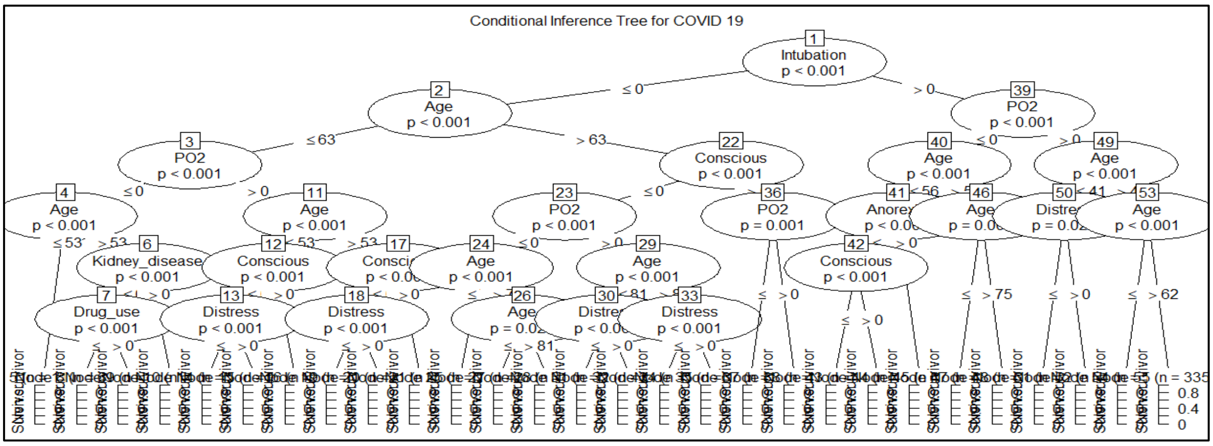


**Figure 2:** Visualization of trained DT model

Analyses were performed in the PCR+ group, and the final model classified subjects on the test data with sensitivity=88.7%, specificity=88.1%, AUC=73.0%, and accuracy=88.2%. These indices on the training dataset are indicated in Table 2. These indices demonstrate that the model performed very well in the prediction task. The extracted rules (28 rules) of the trained tree have been presented in Table 3.

**Table 2:** The model evaluation indices

| Phase | Sensitivity | Specificity | AUC | Accuracy |
|---|---|---|---|---|
| Train | 88.9% | 88.3% | 73.9% | 88.4% |
| Test | 88.7% | 88.1% | 73.0% | 88.2% |

**Table 3:** The 28 rules extracted through DT in the PCR+ group

| | | | | |
|---|---|---|---|---|
| R1 | Intubation <= 0; age <= 63; PO2 <= 0; age <= 53 | Then class is: | Survivor | 11756(98.8) |
| R2 | Intubation <= 0; age <= 63; PO2 <= 0; age > 53; kidney disease <= 0; drug use <= 0 | Then class is: | Survivor | 3274(96.2) |
| R3 | Intubation <= 0; age <= 63; PO2 <= 0; age > 53; kidney disease <= 0; drug use > 0 | Then class is: | Survivor | 53(86.6) |
| R4 | Intubation <= 0; age <= 63; PO2 <= 0; age > 53; kidney disease > 0 | Then class is: | Survivor | 49(79.6) |
| R5 | Intubation <= 0; age <= 63; PO2 > 0; age <= 53; decreased consciousness <= 0; distress <= 0 | Then class is: | Survivor | 2356(96.1) |
| R6 | Intubation <= 0; age <= 63; PO2 > 0; age <= 53; decreased consciousness <= 0; distress > 0 | Then class is: | Survivor | 7603(92.6) |
| R7 | Intubation <= 0; age <= 63; PO2 > 0; age <= 53; decreased consciousness > 0 | Then class is: | Survivor | 195(69.7) |
| R8 | Intubation <= 0; age <= 63; PO2 > 0; age > 53; decreased consciousness <= 0; distress <= 0 | Then class is: | Survivor | 1227(92.1) |
| R9 | Intubation <= 0; age <= 63; PO2 > 0; age > 53; decreased consciousness <= 0; distress > 0 | Then class is: | Survivor | 4342(84.3) |
| R10 | Intubation <= 0; age <= 63; PO2 > 0; age > 53; decreased consciousness > 0 | Then class is: | Survivor | 175(61.7) |
| R11 | Intubation <= 0; age > 63; decreased consciousness <= 0; PO2 <= 0; age <= 72 | Then class is: | Survivor | 2207(92.8) |
| R12 | Intubation <= 0; age > 63; decreased consciousness <= 0; PO2 <= 0; age > 72; age <= 81 | Then class is: | Survivor | 1337(88.0) |
| R13 | Intubation <= 0; age > 63; decreased consciousness <= 0; PO2 <= 0; age > 72; Age > 81 | Then class is: | Survivor | 833(82.1) |
| R14 | Intubation <= 0; age > 63; decreased consciousness <= 0; PO2 > 0; age <= 81; distress <= 0 | Then class is: | Survivor | 1826(85.3) |
| R15 | Intubation <= 0; age > 63; decreased consciousness <= 0; PO2 > 0; age <= 81; distress > 0 | Then class is: | Survivor | 6315(76.2) |
| R16 | Intubation <= 0; age > 63; decreased consciousness <= 0; PO2 > 0; age > 81; distress <= 0 | Then class is: | Survivor | 558(75.1) |
| R17 | Intubation <= 0; age > 63; decreased consciousness <= 0; PO2 > 0; age > 81; distress > 0 | Then class is: | Survivor | 1969(64.7) |
| R18 | Intubation <= 0; age > 63; decreased consciousness > 0; PO2 <= 0 | Then class is: | Survivor | 232(65.9) |
| R19 | Intubation <= 0; age > 63; decreased consciousness > 0; PO2 > 0 | Then class is: | Survivor | 890(50.8) |
| R20 | Intubation > 0; PO2 <= 0; age <= 56; anorexia <= 0; decreased consciousness <= 0 | Then class is: | Survivor | 326(64.1) |
| R21 | Intubation > 0; PO2 <= 0; age <= 56; anorexia <= 0; decreased consciousness > 0 | Then class is: | Death | 37(73.0) |
| R22 | Intubation > 0; PO2 <= 0; age <= 56; anorexia > 0 | Then class is: | Death | 34(79.4) |
| R23 | Intubation > 0; PO2 <= 0; age > 56; age <= 75 | Then class is: | Death | 350(72.0) |
| R24 | Intubation > 0; PO2 <= 0; age > 56; age > 75 | Then class is: | Death | 215(89.3) |
| R25 | Intubation > 0; PO2 > 0; age <= 41; distress <= 0 | Then class is: | Death | 112(64.3) |
| R26 | Intubation > 0; PO2 > 0; age <= 41; distress > 0 | Then class is: | Death | 418(79.4) |
| R27 | Intubation > 0; PO2 > 0; age > 41; age <= 62 | Then class is: | Death | 1569(87.8) |
| R28 | Intubation > 0; PO2 > 0; age > 41; age > 62 | Then class is: | Death | 3353(93.5) |
| Abbreviations: R is short of rule; The values are reported as N (Prob), that N is individual's number in the desired class and its percent. | | | | |

The strongest rules (with the highest probability) to predict each of the two states of non-survival (chance percent=98.8%) or survival (chance percent =93.5%) were rules number 1 and 28, respectively. In other words, from rule 1, it might be observed that patients who do not intubate, are younger than 63, do not get enough oxygen (less than 93%), and are especially younger than 53 are more likely to survive (percent of prob=98.8%). Also, from rule 28, it might be observed that patients who intubate, get enough oxygen (more than 93%), are younger than 41, and especially those younger than 62 are more likely to die (percent of prob=93.5%). DT's rule number 22 informs that if patients intubate, do not get enough oxygen (less than 93%), are under 56 years, and also have anorexia, then the patients' probability of death is 79.4% in this subgroup.

In a subgroup of patients with no incubation and age older than 63 and no decrease in the level of consciousness and not getting enough oxygen (less than 93%) and also, especially younger than 72, the chance of survival was 92.8% (rule 11). In addition, rules 2 and 5 might identify survivors' class with a probability of 96.2% and 96.1%, respectively, based on the ranges for patients' age, intubation status, PO2 level, having or not having kidney disease, drug use, decrease or no decrease in the level of consciousness, and involvement with respiratory distress (See the details of the rules in Table 3).

The other rules with different probabilities for predicting patients' survival and non-survival can be seen in Table 3. The feature importance also tells us how much a feature helped to improve the purity of all nodes' averages. Therefore, the feature importance analysis was done and visualized with a word cloud plot in Figure 3. In this plot, the more important variables have presented larger.

## Discussion

According to the high fatality rate of COVID-19 in Iran (25-27), early detection of the disease and the adoption of effective treatment is crucial to saving more lives, and it has been proven that the use of Machine learning algorithms improves the detection of high-risk groups, prediction, and

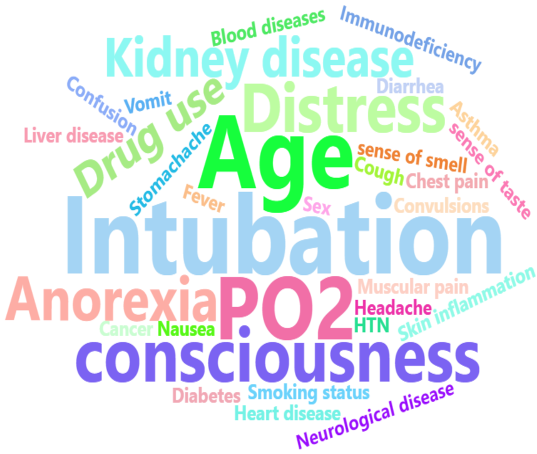


**Figure 3:** Feature Importance for the PCR+

diagnosis of patients with infectious diseases (28). In machine learning, tree-based models are popular as they are easy to understand and interpret and perform well in large datasets. However, some drawbacks lead to poor predictions and no stability (19, 23, 29). To address these problems, CIT models are recommended, and they have been used for analyzing data related to health conditions (19-22).

The most common important characteristics in the PCR+ group were intubation, age, PO2 level, decreased consciousness level, presence of distress, anorexia, drug use, and kidney diseases.

In the study of Mesenburg et al., the change in smell and taste, fever, and cough were identified as important variables by the CIT model (18). The study of Venturini et al. showed that the conditional inference tree model seems robust enough to derive a predictive model(12). The target variable had three levels: death, discharge, or transfer to ICU. Using the CIT model, the plasma sodium level, fever, hospitalization in the ward dedicated to COVID-19 patients only, the arterial pressure of oxygen, and the inspiratory fraction of oxygen ratio (PaO2/FiO2) were identified as important variables. The results of our study about the importance of PO2 align with the findings of Mesenburg et al. and Venturini et al. It is considered that the used predictor variables used in each study were different, and some were in common with us.

However, more studies have implemented DT algorithms in COVID-19 datasets for early diagnosis. Vinod et al. used a DT classifier to identify COVID-19-infected persons. Their results showed that this algorithm could diagnose infected cases with an accuracy of 93% and a precision of 88% (30). Elshazli et al., in a study on 6320 patients, employed the DT algorithm to identify important features related to COVID-19 infections. Their results showed high performance of the model with a sensitivity of 100% and specificity of 81% in the diagnosis of severe patients (31). Shanbehzadeh et al. compared seven DT algorithms (stump, Hoeffding tree, J-48, the logistic model tree, random tree, random forest, and REP-tree algorithms) using performance criteria to find out the best prognostic model. Based on these results, the J-48, with an accuracy of 0.85% and ROC of 0.926%, performed best in diagnosing COVID-19-infected cases (32).

Yeşilkanat et al. assessed the performance of the selected DT algorithms to estimate the near future COVID-19 cases. Their results indicated that the random forest algorithm performs well ($R2 = 0.959$ and RMSE = 259. 38) (33). Our findings indicated that CIT models performed very well in distinguishing between survival and non-survival patients, as shown by sensitivity results of 88.7% and accuracy of 88.2% in the PCR+ group on the test dataset.

## Conclusion

According to the experimental results, the intubation status was a key variable in predicting the patients' survival and non-survival who had positive PCR test results. Some characteristics such as intubation status, age ranges, PO2 level, kidney disease status, drug use status, decreased level of consciousness, involvement status with respiratory distress, and anorexia status contributed to predicting mortality of COVID-19. Being intubated, getting enough oxygen, and being over 62 years old, with a high probability, may lead to the death of patients. Not being intubated, not getting enough oxygen, and being younger than 63, with a high probability, may lead to the survival of patients. Intubation, and other factors, such as old age, anorexia, respiratory distress, and decreased level of consciousness, may lead to death. While there is a pattern in the data that shows if a patient is intubated and is young too and does not have other characteristics such as decreased oxygen level, decreased level of consciousness, and anorexia, it may lead to survival. We can confirm these results because the CIT model had high accuracy and avoided overfitting and selection bias toward predictor selection. The results of this study and the efforts of healthcare systems can be beneficial in mitigating the spread of this pandemic.

### Author statements

The study was approved by the Ethics Committee of the Mashhad University of Medical Sciences with the code of the ethics committee IR.MUMS.REC.1400.248 in 2021.

### References

1. Talkhi N, Akhavan Fatemi N, Ataei Z, Jabbari Nooghabi M. Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods. Biomedical Signal Processing and Control. 2021; 66:102494.
2. Clari M, Luciani M, Conti A, Sciannameo V, Berchialla P, Di Giulio P, et al. The Impact of the COVID-19 Pandemic on Nursing Care: A Cross-Sectional Survey-Based Study. 2021;11(10):945.
3. Kyriazos T, Galanakis M, Karakasidou E, Stalikas A. Early COVID-19 quarantine: A machine learning approach to model what differentiated the top 25% well-being scorers. Personality and Individual Differences. 2021; 181:110980.
4. Abdi M. Coronavirus disease 2019 (COVID-19) outbreak in Iran: Actions and problems. Infection Control & Hospital Epidemiology. 2020; 41(6):754-5.
5. Mohammadi F, Pourzamani H, Karimi H, Mohammadi M, Mohammadi M, Ardalan N, et al. Artificial neural network and logistic regression modelling to characterize COVID-19 infected patients in local areas of Iran. Biomedical journal. 2021.
6. Chen Y-J, Jian W-H, Liang Z-Y, Guan W-J, Liang W-H, Chen R-C, et al. Earlier diagnosis improves COVID-19 prognosis: a nationwide retrospective cohort analysis. Annals of Translational Medicine. 2021;9(11).
7. Nasr Esfahani BN, Ahadi AM, Shalibeik S. A Review of Novel Coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Iranian Journal of Medical Microbiology. 2020;14(2):154-61.

8. Ballı S. Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. Chaos, Solitons & Fractals. 2021;142:110512.
9. Rinivas MLY, Liao HYM. Deep dictionary learning for fine-grained image classification. International Conference on Image Processing 2017:17-20.
10. Sánchez-Montañés M, Rodríguez-Belenguer P, Serrano-López AJ, Soria-Olivas E, Alakhdar-Mohmara Y. Machine learning for mortality analysis in patients with COVID-19. International Journal of Environmental Research and Public Health. 2020;17(22):8386.
11. Alballa N, Al-Turaiki I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Informatics in Medicine Unlocked. 2021:100564.
12. Venturini S, Orso D, Cugini F, Crapis M, Fossati S, Callegari A, et al. Classification and analysis of outcome predictors in non-critically ill COVID-19 patients. Internal Medicine Journal. 2021; 51(4): 506-14.
13. Malik M, Iqbal MW, Shahzad SK, Mushtaq MT, Naqvi MR, Kamran M, et al. Determination of COVID-19 patients using machine learning algorithms. Intelligent Automation Soft Computing. 2022:207-22.
14. Huyut MT, Üstündağ H. Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study. Med Gas Res. 2022; 12(2): 60-6.
15. Van Pelt A, Glick HA, Yang W, Rubin D, Feldman M, Kimmel SE. Evaluation of COVID-19 Testing Strategies for Repopulating College and University Campuses: A Decision Tree Analysis. Journal of Adolescent Health. 2021;68(1):28-34.
16. Rochmawati N, Hidayati HB, Yamasari Y, Yustanti W, Rakhmawati L, Tjahyaningtijas HPA, et al., editors. Covid Symptom Severity Using Decision Tree. 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE); 2020 3-4 Oct. 2020.
17. Zimmerman RK, Nowalk MP, Bear T, Taber R, Clarke KS, Sax TM, et al. Proposed clinical indicators for efficient screening and testing for COVID-19 infection using Classification and Regression Trees (CART) analysis. Human Vaccines Immunotherapeutics. 2021; 17(4): 1109-12.
18. Mesenburg MA, Hallal PC, Menezes AMB, Barros AJ, Horta BL, Hartwig FP, et al. Prevalence of symptoms of COVID-19 in the state of Rio Grande do Sul: results of a population-based study with 18,000 participants. Revista de saude publica. 2021;55:82.
19. Guan H, Dong X, Yan G, Searls T, Bourque CP, Meng FR. Conditional inference trees in the assessment of tree mortality rates in the transitional mixed forests of Atlantic Canada. PloS one. 2021;16(6):e0250991.
20. Sardá-Espinosa A, Subbiah S, Bartz-Beielstein T. Conditional inference trees for knowledge extraction from motor health condition data. Engineering Applications of Artificial Intelligence. 2017;62:26-37.
21. Kaaber Rasmussen NE, Frank Hansen M, Stephensen P. Conditional inference trees in dynamic microsimulation-modelling transition probabilities in the SMILE model. Danish Rational Economic Agents Model, DREAM; 2013.
22. Ahrazem Dfuf I, Mira McWilliams J, M,, González Fernández M, C, . Multi-output conditional inference trees applied to the electricity market: Variable importance analysis. Energies. 2019;12(6):1097.
23. Gross K. Tree-Based Models: How They Work (In Plain English!) 2020 [Available from: https://blog.dataiku.com/tree-based-models-how-they-work-in-plain-english.
24. Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics. 2006;15(3):651-74.
25. Razeghi Nasrabad HB, Sasanipour M. Effect of COVID-19 Epidemic on Life Expectancy and Years of Life Lost in Iran: A Secondary Data Analysis. Iranian Journal of Medical Sciences. 2022; 47(3):210-8.
26. Rassouli M, Ashrafizadeh H, Shirinabadi Farahani A, Akbari ME. COVID-19 management in Iran as one of the most affected countries in the world: advantages and weaknesses. Frontiers in public health. 2020;8:510.
27. Ghafari M, Kadivar A, Katzourakis A. Excess deaths associated with the Iranian COVID-19 epidemic: a province-level analysis. medRxiv. 2020.
28. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals. 2020; 139: 110059.
29. Zhang Z. Decision tree modeling using R. Annals of translational medicine. 2016; 4(15): 275.
30. Vinod DN, Prabaharan S. Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19. Chaos, Solitons & Fractals. 2020; 140:110182.
31. Elshazli RM, Toraih EA, Elgaml A, El-Mowafy M, El-Mesery M, Amin MN, et al. Diagnostic and prognostic value of hematological and immunological markers in COVID-19 infection: A meta-analysis of 6320 patients. PloS one. 2020; 15(8): e0238160.

32. Shanbehzadeh M, Kazemi-Arpanahi H, Nopour R. Performance evaluation of selected decision tree algorithms for COVID-19 diagnosis using routine clinical data. Medical Journal of the Islamic Republic of Iran. 2021;35:29.

33. Yeşilkanat CM. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos, Solitons & Fractals. 2020;140:110210.