International Journal of Emergency Medicine

RESEARCH

Open Access

Development of decision tree classification algorithms in predicting mortality of COVID-19 patients

Zahra Mohammadi-Pirouz[1], Karimollah Hajian-Tilaki[2,3*], Mahmoud Sadeghi Haddat-Zavareh[4], Abazar Amoozadeh[3] and Shabnam Bahrami[1]

Abstract

Introduction The accurate prediction of COVID-19 mortality risk, considering influencing factors, is crucial in guiding effective public policies to alleviate the strain on the healthcare system. As such, this study aimed to assess the efficacy of decision tree algorithms (CART, C5.0, and CHAID) in predicting COVID-19 mortality risk and compare their performance with that of the logistic model.

Methods This retrospective cohort study examined 5080 cases of COVID-19 in Babol, a city in northern Iran, who tested positive for the virus via PCR from March 2020 to March 2022. In order to check the validity of the findings, the data was randomly divided into an 80% training set and a 20% testing set. The prediction models, such as Logistic regression models and decision tree algorithms, were trained on the 80% training data and tested on the 20% testing data. The accuracy of these methods for the test samples was assessed using measures like ROC curve, sensitivity, specificity, and AUC.

Results The findings revealed that the mortality rate for COVID-19 patients who were admitted to hospitals was 7.7%. Through cross validation, it was determined that the CHAID algorithm outperformed other decision tree and logistic regression algorithms in specificity, and precision but not sensitivity in predicting the risk of COVID-19 mortality. The CHAID algorithm demonstrated a specificity, precision, accuracy, and F-score of 0.98, 0.70, 0.95, and 0.52 respectively. All models indicated that factors such as ICU hospitalization, intubation, age, kidney disease, BUN, CRP, WBC, NLR, O2 sat, and hemoglobin were among the factors that influenced the mortality rate of COVID-19 patients.

Conclusions The CART and C5.0 models had outperformed in sensitivity but CHAID demonstrates a better performance compared to other decision tree algorithms in specificity, precision, accuracy and shows a slight improvement over the logistic regression method in predicting the risk of COVID-19 mortality in the population under study.

Keywords Decision tree, CART, C5.0, CHAID, Logistic regression, COVID-19 mortality, Predictive factors

*Correspondence:
Karimollah Hajian-Tilaki
drhajian@yahoo.com
Full list of author information is available at the end of the article

Introduction

At the end of December 2019, emerging coronavirus respiratory disease (COVID-19) spread in Wuhan, China. This disease was named as COVID-19 by The World Health Organization (WHO) on February 11, 2020 [1]. According to clinical symptoms, COVID-19 is divided into four types: mild, moderate, severe and critical [2]. There is substantial evidence indicating that numerous individuals with COVID-19 show no symptoms, yet are capable of spreading the virus to other people [3]. The global impact of this epidemic has resulted in a concerning number of deaths. There are still many aspects of this disease's nature and risk factors that are not yet understood [4]. A study conducted in the Caucasus region found that the mortality rate for patients over the age of 80 was 18.8%, while the overall mortality rate was estimated to be 5% [5, 6]. A Chinese study that observed a group of patients found that the presence of other illnesses, advanced age, and being male were linked to a higher likelihood of experiencing severe disease and death [7]. In another study conducted in China, it was found that hospital death was associated with older age and a lower number of lymphocytes [8]. Furthermore, individuals over the age of 70 experienced a shorter time period between the onset of symptoms and death compared to younger individuals [9]. Additionally, the mortality rate for COVID-19 in hospitals was 28% in Spain, 29.7% in Northern Italy, and 32% in the Caucasus region [10–12]. These findings suggest that patients over the age of 65 have a higher prevalence of underlying health conditions, more severe symptoms, abnormal laboratory results, and are at a greater risk of multiple organ failure and mortality [13].

Many studies have been conducted to predict COVID-19 mortality and assess its related risk factors. These studies have utilized traditional models like logistic regression and Cox regression models [10, 11, 14–19], employing a limited number of predictor variables for causal analysis and variable selection[20]. Logistic regression has considerable limitations for analyzing structured questionnaire data with multiple exposures and missing values [20, 21]. Correlation between predictor variables (multiple collinearity) and a large number of predictors can be considered as some common challenges in traditional models [22, 23]. On the other hand, machine learning (ML) methods can use a larger number of predictors, requiring fewer assumptions, combining "multidimensional correlations" and creating a more flexible relationship between predictor variables and response variables [20]. In addition, ML models can create models for diagnosing and predicting the desired outcome [24], disease modeling [25] and predicting disease and mortality [26]. Despite, the presence of several algorithms of ML, decision tree is a popular method for classification and regression and its analysis is more efficient for visualization of data in clinical decision making and it is interest of clinicians for purpose of classification. Decision tree algorithms are known as one of the most appropriate ML models for effective and reliable decision-making with high accuracy in the classification [27]. In the decision tree, both discrete and continuous variables can serve as the target independent variable. Additionally, this algorithm is non-parametric and does not make any assumptions about the normality of the data [28]. It is used to select variables, evaluate the relative importance of variables, manage missing values, predict, manipulate data, and classify [29].

Four important criteria of sensitivity, specificity, accuracy and precision are used to compare the results of statistical models. In certain studies, both decision trees and logistic regression models demonstrated equal levels of sensitivity. However, when it comes to accuracy, specificity, and precision, the decision tree outperformed the logistic regression model [30]. Among the advantages of decision tree, their simplicity and self-explanation are mentionable, in other words, if they have a reasonable number of leaves, non-professional users and clinicians can understand them, and they can be converted into a set of rules. They can also handle both nominal and numeric input attributes. Decision trees have the capability to handle data sets that contain missing values [31]. Decision tree and neural network models are appropriate alternatives for stepwise regression models in understanding patterns and forecasting. By developing data mining approach for modeling, different types of models can be used to implement different modeling techniques, evaluate the performance of different models and choose the most suitable model for prediction [32]. There are different algorithms for tree classification while the most significant of them are C4.5, ID3, CART, CHAID and SPRINT. C4.5 is the best algorithm for small data sets because it provides better accuracy and efficiency than other algorithms [33].

According to some studies, the decision tree model proved higher diagnostic accuracy in comparison to the logistic regression model [34]. Although decision tree algorithms and logistic regression models may yield different results due to variations in the data. In this regard, the investigation of COVID-19 mortality through this algorithm has not been compared to the logistic regression model. On the other hand, the mortality predictors of COVID-19 are not well known clearly. Furthermore, classical models are often used for this purpose. Therefore, the purpose of the present study is to predict the mortality of patients suffering from COVID-19 and investigate the related factors using decision tree

algorithms and compare them to the logistic regression model.

Methods

Study design and population

This study is a historical cohort. The studied population were COVID-19 PCR-positive cases, who were admitted with clinical and paraclinical diagnosis by an infectious disease specialist in Rouhani Hospital in Babol, north of Iran during the years 2020–2021.

Sample and inclusion/exclusion criteria

The studied sample included 5080 COVID-19 PCR-positive cases. All demographic, clinical, paraclinical information and their discharge status were collected in databases. Men and women over 18 years were eligible for the study. People who were admitted to the emergency room less than 24 h and were discharged, as well as the failure to match the national patient code in the database link, were the conditions for exclusion from the study. Figure 1 shows the flowchart of the selection of patients to enter the study and the statistical analysis of the data.

Data collection

The data was gathered from two registered databases for hospitalized patients with COVID-19. These databases include the Hospital Information System (HIS) of Rouhani Hospital, as well as the database of the Medical Care Monitoring Center (MCMC). These data have been linked using R 4.2.1 software and integration of the national code of patients in these two databases. These databases contain 5845 records of information of COVID-19 PCR-positive patients hospitalized in 2020–2022, in which biomarkers such as Erythrocyte Sedimentation Rate (ESR), C-reactive protein (CRP), Blood Urea Nitrogen (BUN), Alkaline Phosphatase (ALP), Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), White Blood Cell count (WBC), Neutrophil-to-Lymphocyte Ratio (NLR), O2 saturation, Red Blood Cells (RBC), and hemoglobin on the first day of hospitalization, comorbidities such as type 2 diabetes, asthma, heart disease, kidney disease, liver disease, HIV, nervous disorders, immunodeficiency, HTN, hematologic diseases and history of cancer, clinical symptoms such as seizures, diarrhea, dizziness, fever, cough, muscular pain, respiratory distress, olfactory, loss of consciousness, loss of taste, abdominal pain, nausea, vomiting, anorexia, headache, chest pain, hemiparesis, hemiplegia, dermatitis and body temperature, demographic variables such as age, gender, pregnancy, ICU hospitalization, cigarettes, drug use, intubation, length of hospitalization, and discharge status (alive/dead) of all individuals have extracted. Of the 5845 PCR-positive records, 5080 records were eligible for our study and statistical analysis was performed on them. It should be noted that the biomarkers mentioned in this study were collected from all patients on the first day of admission.

Ethical considerations

The information of all patients was collected through registered files of database. This study was approved by the ethics committee of Babol University of Medical Sciences, Babol, Iran with the ethics ID IR.MUBABOL.REC.1401.071. For this study, the informed consent has been obtained by all hospitalized patients to include the data of their hospital charts to the data-based of electronic file for this research.

Imputation of missing values

If the missing values are random, multiple imputation can be done in different methods. Fully conditional specification (FCS) and joint modeling (JM) methods are the most used. In the JM method, the missing values of all variables are performed simultaneously using a statistical model of joint probability functions. The FCS method differs from JM as it does not rely on the joint distribution of variables, but rather on a collection of individual conditional models. In contrast, the JM method utilizes only one multivariate model, making it more straightforward to employ. In contrast, the FCS method, because we consider a separate conditional model for each variable, is more flexible while it has a large number of variables and is more suitable [35, 36]. In the present study, imputation was performed by the FCS method using Mice package. The method creates multiple imputations with replacement values for multivariate missing data. The missing data of each variable is imputed by a separate conditional model. This method can handle in imputing continuous, binary, categorical and order categorical data.

Statistical analysis

R 4.2.1 and SPSS 26 software were used for statistical analysis. In the first step, bivariate analysis, descriptive statistical indices and frequency distribution were performed on all data. In the analysis, the data were classified into two groups: death and discharge, then the chi-square test was used to determine the relationship between qualitative variables and the t-test of two independent samples was used to determine the relationship between quantitative variables related to mortality of patients. In the second step, we have randomly divided the data into two categories: training and testing. In this research, 80% of the data were randomly assigned to the training group and 20% of the data were assigned to the testing group. We have fitted the models on the training

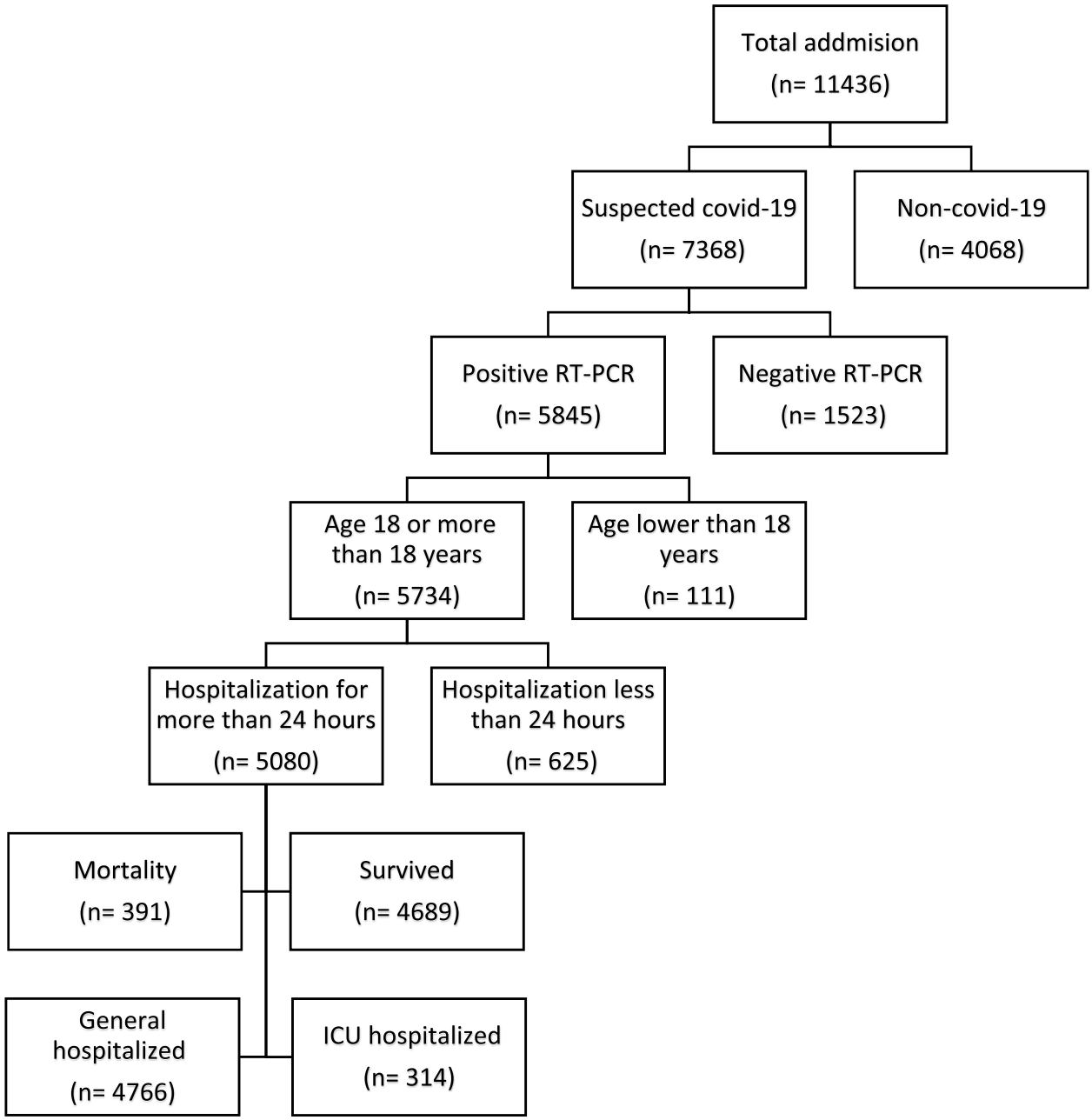

Fig. 1 Flowchart describing patient selection

data and then in the third step on the testing data, models were evaluated and cross-validated based on accuracy, sensitivity, specificity, precision as well as ROC curve.

In real conditions, the data of classification model often encounters with an imbalanced dataset problem, when the number of majority class is much higher than the minority class. This may lead the model unable to learn enough from minority class. In order to overcome this problem, we used SMOT-Tomek technique to balance data. The method that combines oversampling by duplicating some randomness from minority class to balance the data is much popular. Because of an imbalanced data of mortality o COVID-19 (7.7% hospital mortality versus 92.3% survival), we used the SMOTE-Tomek algorithm to balance two strata in the training dataset, but not for testing dataset. Ultimately, the DT

models were developed with a balanced dataset using different DT algorithms and their predictive performance was evaluated in imbalanced testing dataset.

Logistic regression model for predicting COVID-19 mortality

For a binary event, such as mortality, logistic regression is the usual classical method of choice. Similar to linear regression, logistic regression may include only one or more independent variables, and multiple logistic model coefficients reveal the unique contribution of each variable after adjusting for other variables. The probability of occurrence of the outcome with the inclusion of independent variables in the logistic regression was shown by the following equation [37]:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i}} \quad (1)$$

If p is the probability of the outcome, i.e., being in the class of a binary response, in this model, it is assumed that logit(p) has a linear relationship with the variables predicting the outcome.

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \quad (2)$$

The reason for this logit scale transformation lies in the basic parameters of the logistic regression model. The framework of this equation includes independent variables (X) and beta coefficients (β) in linear regression. Indeed, a major advantage of logistic regression is that it retains many of the features of linear regression in its analysis for binary outcomes. Logistic regression after iteration identifies the strongest linear combination of independent variables that increase the probability of detecting the observed outcome, a process known as maximum likelihood estimation, and the β_i coefficients in the model indicate the log OR. In other words, the odds ratio (OR) is equal to e^{β_i} [38, 39].

Decision tree

In a decision tree, each internal node, divides the sample into two or more node according to a specific discrete function of input attribute values. As a result, the algorithm searches for the best attribute to split on. There are various univariate measures [29]. Some of them are based on impurity, while others are normalizers of these criteria. Purity is measured using entropy.

Entropy

One choice to measure the degree of purity is the entropy of information. Entropy is a theoretical measure of the uncertainty in the training data that expresses how random an event is. Entropy is calculated from the following formula [40]:

$$Entropy = -\sum_{i=1}^{c} P_i log_2 P_i \quad (3)$$

where P_i is the probability of a data sample belonging to the i-th class, and c is the number of classes in the target attribute. The higher the entropy, the higher the probability that a sample of data belongs to a class by chance, and that attribute does not express much information about the target attribute.

Information gain

This measure uses the entropy as a criterion of impurity. The variable with the most information gain is selected for the root node, and the variable with less entropy has more information gain [27].

$$InformationGain(A) = Entropy(D) - Entropy_A(D) \quad (4)$$

that,

$$Entropy(D) = -\sum_{i=1}^{c} P_i log_2 P_i \quad (5)$$

$$Entropy_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Entropy(D_j) \quad (6)$$

where D_j is the number of samples at level j of attribute A, D is the number of training data, c is the number of available classes,P_i is the probability that a data sample belongs to the i-th class, and v is the number of domain members of attribute A.

Gini Index

The Gini index is a measure based on impurity. Binary classification is done for each variable and the variable with the lowest Gini is selected for the root node [40].

$$Gini(A) = Gini(D) - Gini_A(D) \quad (7)$$

$$Gini(D) = 1 - \sum_{i=1}^{c} P_i^2 \quad (8)$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (9)$$

where D is the number of training data, c is the number of available classes, P_i is the probability that a sample of the data belongs to the i-th class, and $D_{1,2}$ is the data set D for attribute A, which is divided into 2 parts.

Gain Ratio

This criterion normalizes the information gain as follows [27]:

$$Gainratio_A(D) = \frac{InformationGain(A)}{Entropy_A(D)} \quad (10)$$

This ratio is not defined when the denominator is zero. Also, this ratio may be in favor of adjectives whose denominator is very small. It has been shown that gain ratio performs better in comparison to information gain, both in terms of accuracy and complexity [41].

Decision tree algorithms

According to recent years, several algorithms have been developed for diagnostic classifications with decision trees, among which the most important ones include the following:

CART

Classification and regression tree (CART) make binary tree, that is, each internal node has exactly two branches. Partitions are selected using the Gini criterion. One of the important features of CART is its ability to generate regression trees. Regression trees are the trees whose leaves predict a real number instead of a class. CART looks for partitions that minimize the prediction error. The prediction in each leaf is based on the weighted average of the nodes [41].

C5.0

This algorithm uses gain ratio as a splitting criterion. When the number of samples to be split is less than a certain threshold, the split stops. C5.0 can generate missing values from a training set using the modified gain ratio criteria presented above. C5.0 algorithm can understand discrete or continuous values [41].

CHAID

CHAID is designed for nominal attributes. For each input attribute A, CHAID finds values that have the least significant difference from the target attribute. The significant difference is measured by the *p*-value obtained from the statistical test. The statistical test depends on the type of target attribute. If the target attribute is continuous, the F test, if it is nominal, the chi-square test, and if it is ordinal, the likelihood ratio test is used. Then the best input attribute is selected to be used to split the current node. This method also stops when one of the following conditions is met:1) The maximum depth of the tree has been reached. 2) The minimum number of cases in a node to be a parent has been reached, so it can no longer be split. This algorithm handles missing values by treating them all as a single category [41].

Evaluation of the performance of fitted models in testing data

Comparison of models using ROC curve

In this study, the diagnostic accuracy of decision tree algorithms was compared with the traditional logistic regression model on the logit scale using the ROC curve, and the area under the curve (AUC), sensitivity (Recall), specificity, accuracy, precision and F-score were used. In dichotomous (positive/negative) diagnostic tests, the conventional approach to test evaluation uses sensitivity and specificity compared to the gold standard status. In situations where the test results are reported in an ordinal or continuous scale, the sensitivity and specificity scale can be calculated in all possible threshold values. Hence, the sensitivity and specificity vary at different thresholds. A plot of sensitivity versus 1 minus specificity is called the receiver operating characteristic (ROC), and the area under the curve (AUC) is considered an effective measure of accuracy with meaningful interpretations.

This curve plays the main role in evaluating the diagnostic ability of tests to detect the true condition of people, finding the optimal cutoff and comparing two diagnostic methods. This predictive model is commonly used to estimate the risk of any adverse outcome based on the patient's risk profile in medical research. There are various methods to determine the optimal cutoff, including the method that maximizes the sum of sensitivity and specificity (or equally minimizes the sum of false positive and false negative errors). This criterion can be used to consider a cutoff as optimal. In this context, the Youden index is an index that maximizes the vertical distance between the ROC curve and the diagonal line (representing the chance level). It is defined as T*P*-FP and can be calculated as follows [42]:

$$Youden'sindex = sensitivity + specificity - 1 \quad (11)$$

While the other two indices—positive predictive value (PPV) and negative predictive value (NPV) may have interesting interpretations from a clinical standpoint, they are influenced by the disease's prevalence and are less accurate as diagnostic tests. The area under the curve (AUC) summarizes the entire area of the ROC curve instead of relying on a specific operating point. AUC is an effective and comprehensive measure of both sensitivity and specificity, providing valuable information about the intrinsic validity of diagnostic tests. AUC ranges between 0 and 1, with 1 indicating perfect discrimination between sick and healthy individuals. When the maximum AUC

is equal to 1, it means that the test has successfully differentiated between sick and non-sick individuals, as the distributions of test results for these two groups are completely distinct from each other [43].

Performance indicators of diagnostic accuracy

In this study, besides AUC, other indices were also utilized. These include diagnostic accuracy, sensitivity, specificity, and F-score, all of which were determined using the following formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \quad (14)$$

$$Specificity = \frac{TN}{TN + FP} \quad (15)$$

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (16)$$

that TP expresses true positive, TN true negative, FP false positive and FN false negative values. The higher the accuracy, sensitivity, and specificity of the model, the better the model will be [44]. In our study, ROC curve using AUC, sensitivity, specificity, accuracy, precision and F-score was used to evaluate the models. Figure 2 displays the flowchart depicting the process of statistical analysis for the training and testing datasets.

Results

Missing data imputation

In this study, imputation of missing data was performed using R 4.1.2 software from the Mice package through multiple imputation. To confirm that the missing values were randomly distributed and not biased, a comparison was conducted between the mortality ratio in the available data of the variables and the ratio in the missing data. No significant difference was found. Out of the 50 variables included in the study, 14 variables had missing records. Twelve variables had missing values that were less than 10%, while two variables, namely body temperature and NLR, had missing values greater than 10% (18% and 33% respectively).

Descriptive demographic and clinical findings and univariate tests

In this study, among the 5734 PCR positive patients of COVID-19, 5080 individuals were eligible to participate in the study. Out of the total 5080 patients, 4689 (92.3%) were discharged and 391 (7.7%) patients died. Among these patients, 2314 (45.6%) were men with an average age of 57.79 ± 16.83 years, while 2766 (54.4%) were women with an average age of 55.22 ± 16.06 years. In the male group, 199 individuals (8.6%) died, while in the female group 192 individuals (6.9%) died ($p = 0.027$). The average age in the group of deceased patients was 68.56 ± 14.36 years, whereas in the discharged group it was 55.38 ± 16.22 years ($p = 0.001$). Among all the patients

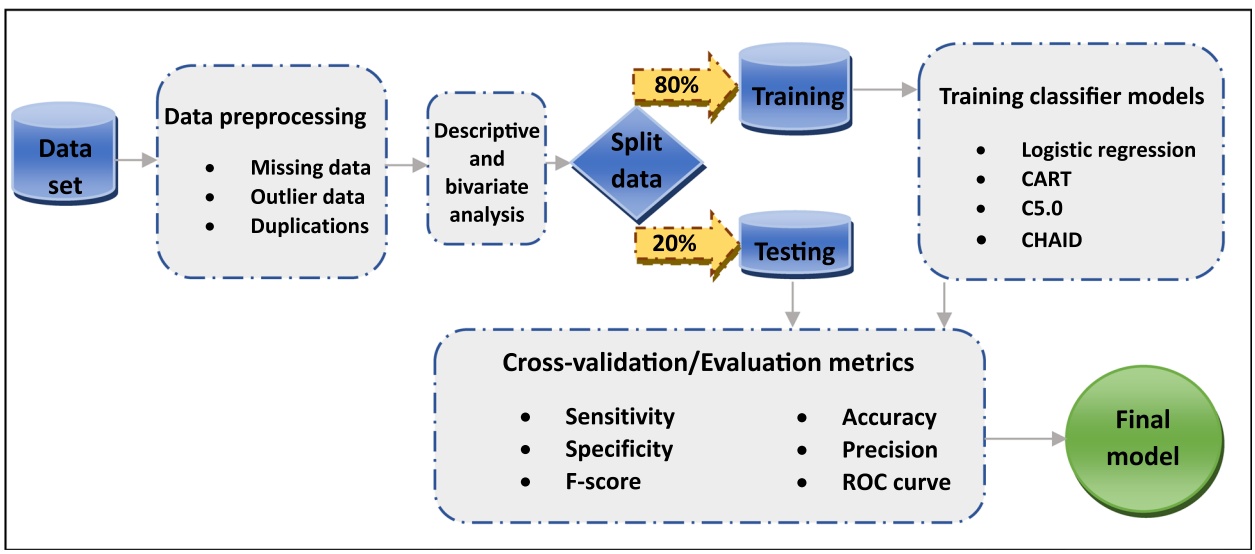

Fig. 2 Flowchart of describing the steps of analysis progress

studied, there were 100 cases (2%) of cancer, 20 cases (4%) of liver diseases, 1159 cases (22.8%) of diabetes, 23 cases (0.5%) of hematologic diseases, 15 cases (0.3%) of HIV, 21 cases (0.4%) of immunodeficiency, 670 cases (13.2%) of heart diseases, 58 cases (1.1%) of kidney diseases, 95 cases (1.9%) of asthma, 48 cases (0.9%) of nervous disorders, and 1145 cases (22.5%) of HTN. The Chi-square test was employed to examine the correlation between comorbidities and the mortality rate of COVID-19 patients. The findings revealed a significant association between cancer ($p=0.001$), diabetes ($p=0.001$), hematologic disease ($p=0.011$), heart disease ($p=0.001$), kidney disease ($p=0.001$), neurological disorders ($p=0.001$), and HTN ($p=0.001$) with an elevated mortality rate attributed to COVID-19 (Table 1).

Table 2 displays the median and interquartile range (IQR) of two patient groups, those who deceased and those who were discharged. The mean of all biological markers, excluding ALT, in the deceased group significantly differs when compared to the discharged group. To confirm this significance, we employed the U-Man Whitney test.

Findings of the logistic model

To fit the models, the data was first randomly divided into training and testing data in the ratio of 80% to 20%, respectively. Then the stepwise logistic regression model was applied. The coefficients of the model and their odds ratios (95% confidence intervals) are displayed in Table 3. Out of 50 variables, 14 variables entered the final model. According to these results, the variables of age, ICU hospitalization, fever, loss of consciousness, intubation, diabetes, O2 saturation, kidney disease, ESR, BUN, CRP, NLR, and AST were found to be statistically significant. These results indicate that patients in the age group above 65 years have a 5.7 times higher chance of death compared to the age group of 18–44 years. Similarly, patients in the age group of 64–45 years have a 93% greater chance of death than the age group of 18–44 years. Additionally, the chance of death in hospitalized patients in the ICU is 10.36-fold higher than in patients admitted to the general ward. Furthermore, patients who underwent intubation had a 25.1-fold higher chance of death. In patients with a fever, the risk of death was 41% lower compared to patients without a fever. Additionally, patients with a decreased level of consciousness had a 2.41-fold higher risk of death. Individuals with diabetes and kidney diseases had a 59% and 3.95-fold higher risk of death, respectively.

Findings of decision tree models

The results of identifying the risk factors that are effective in predicting the mortality of COVID-19 patients, using three methods to rank the relevant attributes—Information gain, Gain ratio, and Gini index—are shown in Table 4. These results indicate that adjectives with higher ranks have a greater impact on predicting mortality caused by COVID-19. The variables that were of higher importance in relation to the mortality of COVID-19 patients are loss of consciousness, BUN, ALP, CRP, WBC, NLR, O2 sat, age, ICU hospitalization, and intubation—these are the ten variables.

CART algorithm findings

According to Fig. 3, the results of this algorithm indicate that among all the variables in this model, ICU hospitalization, BUN, intubation, age, WBC, hemoglobin, and CRP are included. The algorithm was able to classify 16% of individuals in the death group and 84% of individuals in the discharge group. This algorithm revealed that 6% of the patients who died were those admitted to the ICU. Additionally, 4% of the patients who were not admitted to the ICU had BUN levels above 27, were not intubated, had WBC count below 11,000, but their hemoglobin level was less than 11. Furthermore, 3% of the patients who were not admitted to the ICU had BUN levels above 27, were not intubated, and had a WBC count exceeding 11,000. Moreover, 1% of the patients who were not hospitalized in the ICU had BUN levels above 27 and were intubated, and 1% of those same patients had BUN levels below 27 and were intubated. Lastly, 1% of the patients who were not hospitalized in the ICU had BUN levels below 27, were not intubated, were over 64 years old, and had a CRP reading over 201.

C5.0 algorithm findings

Figure 4 displays the results obtained from the C5.0 algorithm. In this particular model, the variables considered included intubation, ICU hospitalization, BUN levels, kidney disease, WBC, fever, length of hospitalization, and CRP. Within the model, it was determined that 6.1% of individuals in the group died, while 93.9% were classified as being part of the discharge group. Moreover, it was further observed that 3.8% of the patients who passed away had been intubated, whereas 1.4% of those who did not require intubation had been admitted to the ICU and had a BUN level greater than 27.

CHAID algorithm findings

In this model, intubation, ICU hospitalization, BUN, age, and kidney diseases were the most important variables included. Figure 5 shows that 5.7% of individuals were classified in the death group, while 94.3% were classified in the discharge group. Out of the deceased patients, 80 were those who were intubated and hospitalized in the ICU. Additionally, 61 individuals were intubated but not

Table 1 Demographic and comorbidity of study participants according to COVID-19 mortality

Characteristic	Mortality *n* (%)	Survived *n* (%)	Total *n* (%)	*p*-value (χ^2)
Age (year)				0.001
18- 44	24 (1.8)	1301 (98.2)	1325 (26.1)	
45- 64	111 (5.3)	1968 (94.7)	2079 (40.9)	
65 ≤	256 (15.3)	1420 (84.7)	1676 (33.0)	
Gender				0.027
Male	199 (8.6)	2115 (91.4)	2314 (45.6)	
Female	192 (6.9)	2574 (93.1)	2766 (54.4)	
Cigarette user				0.163
Yes	(11.5) 11	(88.5) 85	96 (1.9)	
No	380 (7.6)	4604 (92.4)	4984 (98.1)	
Drug user				0.094
Yes	13 (11.9)	96 (88.1)	109 (2.1)	
No	378 (7.6.)	4593 (92.4)	4971 (97.9)	
Cancer				0.001
Yes	17 (17.0)	83 (83.0)	100 (2.0)	
No	374 (7.5)	4606 (92.5)	4980 (98.0)	
Liver disease				0.699
Yes	2 (10.0)	18 (90.0)	20 (0.4)	
No	389 (7.7)	4671 (92.3)	5060 (99.6)	
Type 2 diabetes				0.001
Yes	148 (12.8)	1011 (87.2)	1159 (22.8)	
No	243 (6.2)	3678 (93.8)	3921 (77.2)	
Hematologic disease				0.011
Yes	5 (21.7)	18 (78.3)	23 (0.5)	
No	386 (7.6)	4671 (92.4)	5057 (99.5)	
HIV				0.632
Yes	0 (0.0)	15 (100.0)	15 (0.3)	
No	391 (7.7)	4674 (92.3)	5065 (99.7)	
Immunodeficiency				0.216
Yes	3 (14.3)	18 (85.7)	21 (0.4)	
No	388 (92.3)	4671 (92.3)	5059 (99.6)	
Heart disease				0.001
Yes	87 (13.0)	583 (87.0)	670 (13.2)	
No	304 (6.9)	4106 (4106)	4410 (86.8)	
Kidney disease				0.001
Yes	19 (32.8)	39 (67.2)	58 (1.1)	
No	372 (7.4)	4650 (92.6)	5022 (98.9)	
Asthma				0.904
Yes	7 (7.4)	88 (92.6)	95 (1.9)	
No	384 (7.7)	4601 (92.3)	4985 (98.1)	
Chronic Nervous Disorders				0.001
Yes	11 (22.9)	33 (77.1)	48 (0.9)	
No	380 (7.6)	4652 (92.4)	5032 (99.1)	
HTN				0.001
Yes	135 (11.8)	1010 (88.2)	1145 (22.5)	
No	256 (6.5)	3679 (93.5)	3935 (77.5)	

Table 2 The median (IQR) of biomarkers according to COVID-19 mortality

Biomarkers	Mortality Median (IQR)	Survived Median (IQR)	Total	*p*-value (χ^2)
NLR	(5.10) 5.66	(3.17) 3.78	Median (IQR)	0.001
Blood urea nitrogen	(22) 28	(9) 17	(3.10) 3.9	0.001
(mg/dl)	(25) 29	(25) 29	(10) 17	0.99
Alanine transferase	(39) 52	(26) 39	(25) 29	0.001
(U/L)	(122) 195	(81.5) 165	(26) 40	0.001
Aspartate aminotransferase	(45) 40	(35) 31	(84.75) 167	0.001
(U/L)	(62) 78	(56) 49	(35) 32	0.001

Table 3 The regression coefficients and odds ratio (OR) of the stepwise logistic regression in COVID-19 mortality

Independent variable	B	SE (B)	OR (95% CI)	*p*-value
Age (year)				
18- 44	-		-	-
45- 64	0.73	0.32	(1.07–3.67) 1.93	0.022
65 ≤	1.82	0.31	(3.25–10.56) 5.7	0.001
ICU hospitalization (yes vs no)	2.38	0.20	(6.96–15.45) 10.36	0.001
Fever (yes vs no)	0.52-	0.16	(0.43–0.81) 0.59	0.001
Loss of consciousness (yes vs no)	0.88	0.37	(1.15–4.83) 2.41	0.016
Intubation (yes vs no)	3.18	0.24	(15.11–39.59) 24.2	0.001
Type 2 diabetes (yes vs no)	0.46	0.17	(1.14–2.21) 1.59	0.004
O2 sat(mg) (> 93% vs < 93%)	0.84-	0.16	(0.32–0.58) 0.43	0.001
Kidney disease (yes vs no)	1.37	0.43	(1.65–9.01) 3.95	0.001
ESR (mm/h)	0.007-	0.003	(0.986–0.998) 0.992	0.013
BUN (mg/dl)	0.027	0.004	(1.02–1.036) 1.028	0.001
CRP (mg/L)	0.007	0.001	(1.0049–1.0095) 1.007	0.001
NLR	0.033	0.016	(1.001–1.067) 1.034	0.05
Hemoglobin (g/dL)	0.072-	0.038	(0.86–1.002) 0.93	0.05
AST (U/L)	0.006	0.002	(1.003–1.009) 1.006	0.001

hospitalized in the ICU. Furthermore, 70 of the deceased patients were not intubated, but they were hospitalized in the ICU and had a BUN greater than 24. Lastly, 14 individuals were not intubated, not hospitalized in the ICU, had a BUN greater than 24, and had kidney disease.

Comparison of predictive performance of fitted models in testing data

In this section, we cross-validate the performance of the fitted model predictors using 20% of the testing data, and the results are presented in Table 5. To assess the predictive power of the algorithms, we consider sensitivity as a measure. The CART and C5.0 algorithms performed the best, achieving sensitivities of 0.77 and 0.75, respectively. On the other hand, the logistic model and CHAID performed better in terms of specificity, obtaining a specificity of 0.98. When it comes to precision, the CHAID model performed the best, with a precision of 0.70. In terms of diagnostic accuracy, both the CHAID and C5.0 models achieved a score of 0.95. The F-score indicated that the C5.0 algorithm and CHAID had similar performance, outperforming other models. As for the area under the ROC curve, both the logistic and CART models displayed similar performance, surpassing other models.

In this study, the logistic regression model, the C5.0, and CART algorithms have higher specificity but CART has a better performance in sensitivity. However, C5.0 has much ability in predicting outcome (precision) compared to CART and CHAID algorithms. The accuracy of all models was >=0.90. So, the question now is which model performs better. Since the desired outcome in this study is mortality, it is important to accurately diagnose the death rate, which is better in the CART model compared to other models. However, other indicators in these models should also be considered. Therefore, although

Table 4 Ranking from low to high importance of each attribute in predicting COVID-19 mortality using decision tree indices

Attributes	Information gain	Gini index	Gain ratio	Average ranks
Chest pain	4	2	2	2.67
Olfactory	3	1	4	2.67
Abdominal pain	5.5	3.5	5.5	4.83
Asthma	5.5	3.5	5.5	4.83
Nausea	9	7	3	6.33
Anorexia	10	8	1	6.33
Seizure	7	5	11	7.67
Liver disease	8	6	13	9
Vomiting	11	9	10	10
Muscular pain	12	12	7	10.33
Headache	13	11	9	11
Cough	15	14	8	12.33
Dizziness	17	16	15	16
Body temperature	1.5	34	16	17.17
Immunodeficiency	14	15	24	17.67
Drug	19	17	18	18
Cigarettes	20	19	19	19.33
Diarrhea	22	20	17	19.67
HIV	18	10	35	21
Gender	29	23	12	21.33
Hemiparesis	16	18	30	21.33
ALT	1.5	40	23	21.5
Fever	30	27	14	23.67
Pregnancy	21	13	37	23.67
Cancer	27	25	28	26.67
HTN	31	29	21	27
Respiratory distress	33	30	20	27.67
Loss of taste	26	21	39	28.67
Hematologic disease	23	22	42	29
Heart disease	32	31	26	29.67
Hemiplegia	24	24	44	30.67
ESR	34	38	22	31.33
Type 2 diabetes	36	32	27	31.67
Nervous disorder	28	26	41	31.67
Dermatitis	25	28	48	33.67
Hemoglobin	37	41	25	34.33
AST	41	42	29	37.33
RBC	39	43	31	37.67
Kidney disease	35	33	47	38.33
Length of hospitalization	42	36	40	39.33
Loss of consciousness	38	35	46	39.67
BUN	48	35	38	40.33
ALP	40	48	34	40.67
CRP	46	44	33	41
WBC	45	46	32	41
NLR	43	47	36	42
O2 sat	44	37	45	42
Age	47	39	43	43
ICU hospitalization	49	49	49	49
Intubation	50	50	50	50

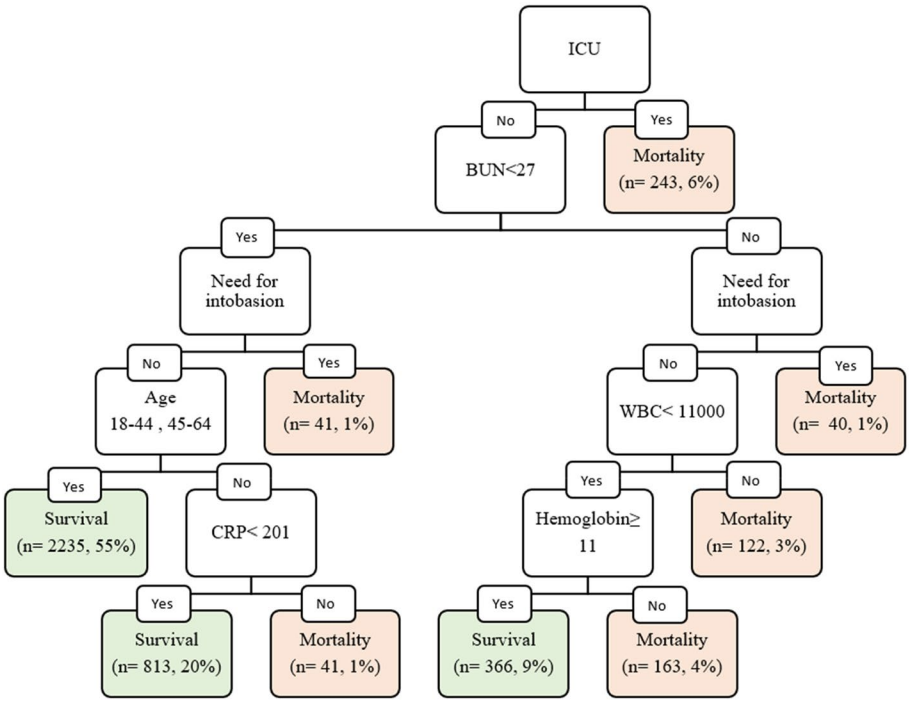

Fig. 3 CART algorithm for predictors of mortality

these models have higher sensitivity, they should not be ignored because the sensitivity and specificity cannot both be high at the same time. Additionally, based on the AUC, the CART model is slightly different from the logistic model. The ROC curve was used to compare the prediction performance of the models, where the higher the levels under the curve, the higher the AUC and the better the model performs. The ROC curve for all three decision tree models and the logistic regression model is shown in Fig. 6. In terms of AUC, the logistic models and CART models performed better and had almost similar performance.

Comparison of performance of DT predictive model using balanced dataset

We applied the SMOTE-Tomek algorithm to develop the DT models with a balanced training dataset. Then, the performance of predictive models was evaluated in both training and testing data sets. The results showed that the CART model has sensitivity, specificity, and accuracy of 0.99, 0.46, and 0.74 in training balanced dataset, while these indexes were 0.98, 0.51, and 0.55 respectively in the testing datasets. The fitted C5.0 model had the sensitivity, specificity, and accuracy of 0.93, 0.93, and 0.93 in a balanced training data set, but these indexes were 0.70, 0.85, and 0.84 respectively in testing datasets. Finally, the fitted CAHID model showed sensitivity, specificity, accuracy of 0.49, 0.98, and 0.94 respectively in a balanced training dataset. However, these measures were 0.41, 0.98, and 0.94 in the testing dataset. Thus, the sensitivity of DT models was decreased in the testing dataset.

Discussion

In this study, we identified the factors that affect COVID-19 mortality using a logistic regression model and decision tree algorithms. Understanding the factors that influence mortality is essential for clinicians and health policymakers when monitoring hospitalized COVID-19 patients. According to the results of this study, among these factors, we can mention ICU hospitalization, intubation, age, kidney diseases, hemoglobin level, and biological markers such as NLR, WBC, O2 sat, CRP, and BUN. These factors were found to have a significant relationship with the mortality rate. The predictors of mortality caused by COVID-19 have been widely reported in traditional classical models in different regions. These models include the findings of biological and radiological markers, co-morbidities, and demographic variables. Numerous studies that predicted the effective factors in COVID-19 mortality mainly used classical statistical methods, which is somewhat consistent with the results of our study.

In the present study, ICU hospitalization is identified as a key factor influencing the mortality rate among patients with COVID-19. This variable has consistently been included in all four proposed models of this

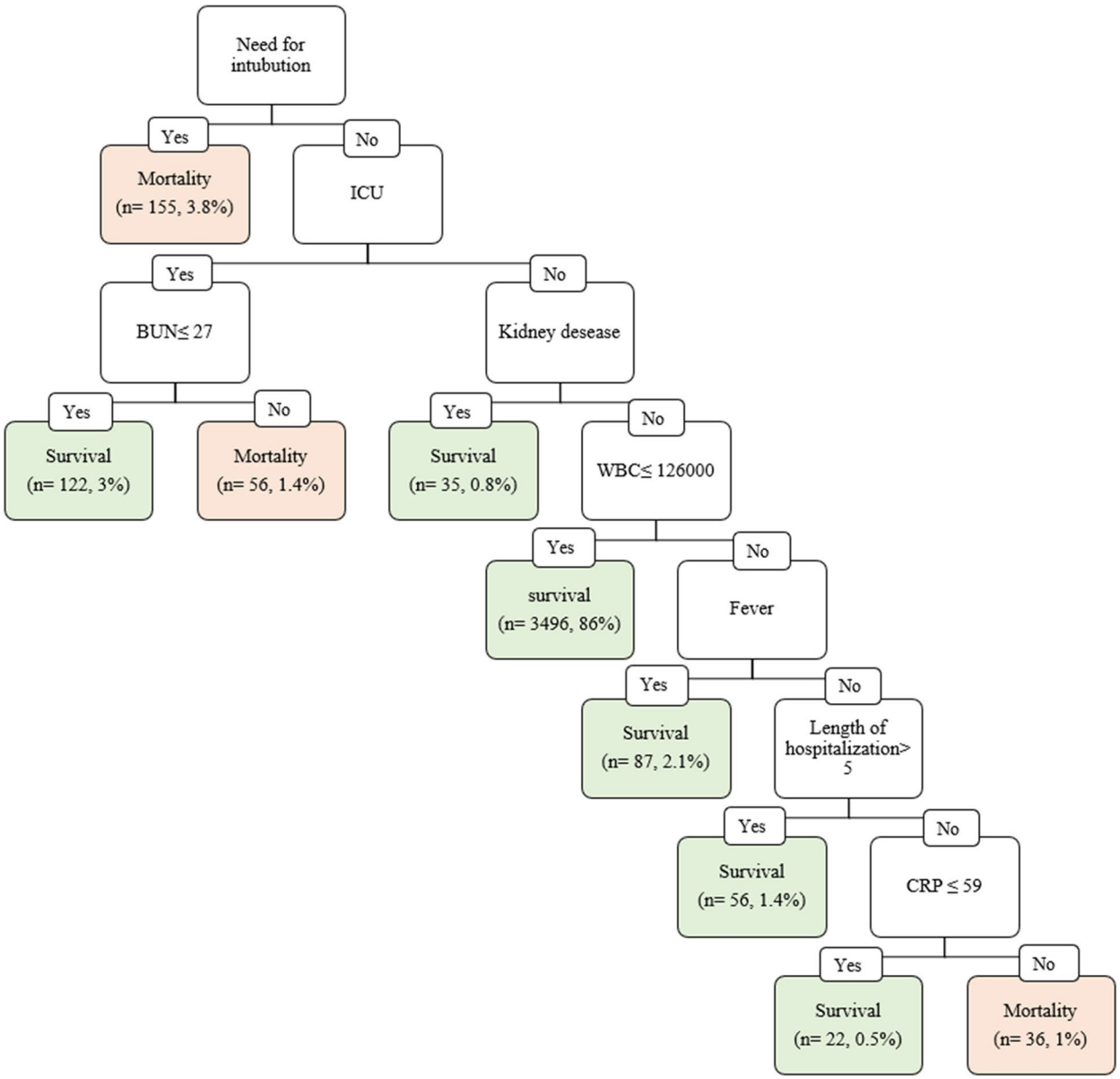

Fig. 4 C5.0 algorithm for predictors of mortality

study, highlighting its significance in increasing the risk of mortality. This escalation in risk could potentially be attributed to the severity of patients' conditions within the ICU. This variable had the strongest impact on mortality in numerous studies. For instance, in a study conducted by Dawood Adham in Ardabil, Iran [37], it was found to have a significant effect. Another study by Karaca-Mandic in the United States demonstrated that a 1% increase in ICU bed utilization is linked to a 2.84-fold increase in COVID-19 mortality [45]. Among other significant variables in our three models, we can mention old age and CRP. In a study by Nasser Malekpour in Tehran, Iran, which examined 396 surviving patients and 63 deceased patients, it was shown that the likelihood of death in the hospital is influenced by age and CRP levels upon admission [38]. A prospective study was also conducted in Iran by Ruhollah Alizadeh. Three hundred and nineteen patients with COVID-19 were followed up after two months to assess their health status. Fever, CRP, and age were identified as the most significant symptoms of COVID-19 infection [39]. These findings align with our study. Another study conducted in Birjand, Iran, by Qodsieh Azarkar revealed significant differences in clinical parameters

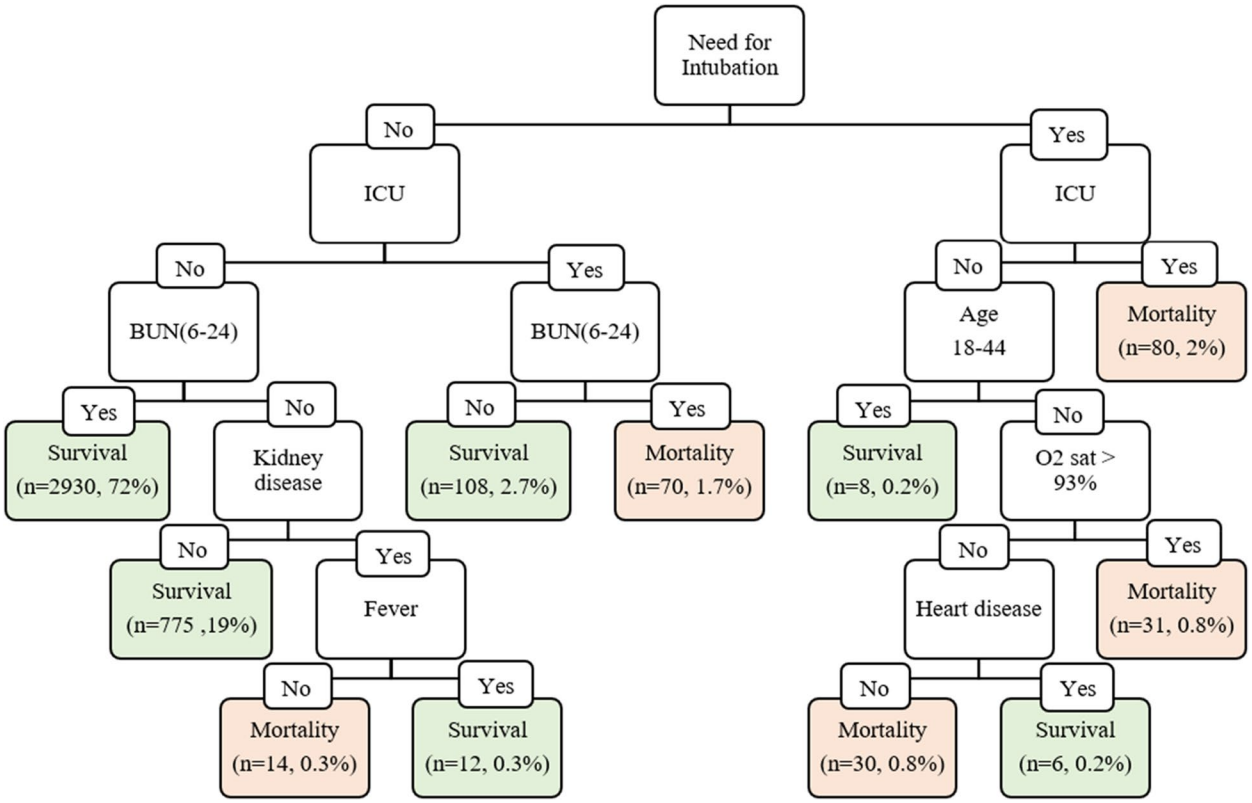

Fig. 5 CHAID algorithm for predictors of mortality

Table 5 Comparing the performance of logistic regression and decision tree algorithms to predict the COVID-19 mortality in testing dataset with model fitted to imbalanced data

Models	Confusion matrix				Sensitivity (Recall)	Specificity	Precision	Accuracy	F-score	AUC
	Predicted	Actual	Mortality	Survived						
	Mortality		TP	FP						
	Survived		FN	TN						
Logistic			26 42	13 935	0.38	0.98	0.66	0.94	0.48	0.93
CART			51 17	89 859	0.75	0.91	0.36	0.90	0.49	0.92
C5.0			28 8	40 940	0.41	0.99	0.78	0.95	0.41	0.78
CHAID			28 40	12 936	0.41	0.98	0.70	0.95	0.52	0.87

and comorbidities between the death and discharge groups. Parameters such as O2 saturation, lymphocyte and platelet count, hemoglobin level, CRP, and liver and kidney function displayed statistical differences. These differences hold meaningful importance. The results indicate that comorbidities, the number of lymphocytes and CRP, may increase the risk of death in hospitalized patients with COVID-19. Patients with lower lymphocyte counts in their hemogram and high levels of CRP, as well as those with comorbidities, are more likely to face a higher risk of death [40]. In another study carried out by Javanian in Babol, Iran, it was found that older age, length of hospital stays, ICU stay, kidney failure, and lymphocyte count were associated with mortality [41]. This aligns with our findings. In a study conducted by Fabiana Tezza in Italy,

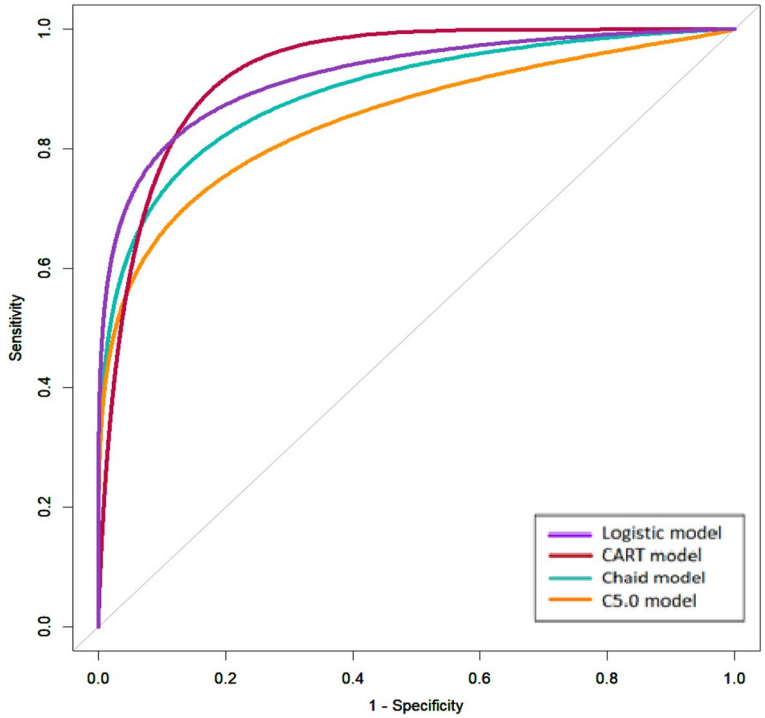

Fig. 6 Comparison of logistic model and decision tree algorithms performance in terms of ROC curve

the identification of predictors for COVID-19 mortality revealed that age and hemoglobin were among the most significant predictors of in-hospital mortality [46]. Because the mechanism of kidney dysfunction caused by COVID-19 is still unknown, it has been shown that SARS-CoV-2 plays a pathogenic role in COVID-19 patients by binding to the angiotensin-converting enzyme (ACE) 2 receptor [47]. A study conducted by Bertsimas in America identified increasing age, O2 saturation, increased CRP, and BUN as the most important predictors of COVID-19 mortality [4]. Similarly, a study conducted by Maryam Kabotari in Iran found that age and O2 saturation were significantly related to mortality caused by COVID-19 in hospital settings [48]. In a systematic review and meta-analysis conducted by Zhao Zheng et al. in China, which included 3027 patients with COVID-19, it was found that age over 65 years and smoking were identified as risk factors for disease progression [3]. Furthermore, a study carried out by Fabiana Tezza in Italy examined 341 patients with an average age of 74 years, and determined that age, along with vital signs and laboratory parameters such as lymphocyte count and hemoglobin, were the most significant predictors of in-hospital mortality. These findings align with our own study.

According to our recent study, we discovered that the mortality rate among COVID-19 patients is 7.7%, which is very similar to the findings reported by Dawood Adham in his study that indicated a mortality rate of 8.5% [37]. However, another systematic review and meta-analysis conducted by John J Y Zhang in China reported a lower mortality rate of 4.3% [44]. This variance could possibly be attributed to variations in the level of specialized care, treatment protocol, and the criteria for including patients in each respective study. Additionally, in a study conducted in Birjand, Iran by Qudsieh Azarkar, the mortality rate was approximately 17.4% [40], potentially attributed to the limited sample size of 360 participants and specific inclusion criteria. Furthermore, this study also indicated a hospitalization rate of 6.2% in intensive care units (ICUs). In contrast, a systematic review and meta-analysis study conducted by John JV Zhang found a higher ICU admission rate of 10.9% [44]. The difference in rates could be explained by variations in the severity of the disease or the adequacy of ICU beds in the healthcare system of China.

Our study has revealed that decision tree algorithms can effectively serve as an alternative in developing mortality prediction models for COVID-19 patients. Similar to Mostafa Shanbezadeh's study in Iran, the use of the Gini index facilitated an investigation into the criteria for diagnosing COVID-19. The findings demonstrated that the J-48 algorithm displayed the highest performance, with an accuracy of 0.85, in detecting COVID-19 [49].

In our study, we found that both the C5.0 and J-48 algorithms performed exceptionally well, displaying a high accuracy rate of 0.95. These results indicate that, despite having fewer variables and assumptions compared to the logistic regression model, the decision tree model shows a remarkable predictive accuracy. By utilizing a reduced number of variables, the decision tree algorithm can achieve comparable levels of accuracy and sensitivity as logistic regression, while also demonstrating a higher level of specificity. In the current study, the input variables in the models derived from three decision tree algorithms are highly similar. The variables used in the CART and CHAID models are identical to those in the logistic model. Likewise, the variables employed in the C5.0 model are all present in the logistic model, except for the length of hospitalization. To compare the models' performance, the ROC curve was utilized, specifically examining the AUC. A larger AUC, indicating a broader area under the curve, is indicative of superior performance.

Machine learning (ML) models are being increasingly employed in the arena of diagnosing and predicting health-related outcomes. Among these models, DT, Random Forests, and neural networks have a significant legacy. DT has the advantage of automatically identifying predictor variables, making it easier for clinicians to interpret results and identify non-linear relationships. This stands in contrast to the multiple logistic regression model, which depends on assumptions of linearity on the scale. The logistic regression model and the collinearity of variables are potential challenges to the accuracy and validity of the results. The results of this study demonstrate that even though there are defaults in the logit model, the DT method has nearly identical diagnostic accuracy with the added benefits of easy interpretation of results and sequential analysis of variables. In fact, certain parameters may even outperform the logit model. Our findings highlight the efficiency of DT analysis, a method that relies on straightforward algorithmic rules, for predicting mortality outcomes, as opposed to the logit regression model, which focuses on establishing the relationship between independent variables and outcomes.

The factors affecting the death of COVID-19 patients were often discussed in classical models and only a few studies focused on predicting mortality. This current study has several advantages. Firstly, it utilized a large dataset with over 5 thousand records of hospitalized COVID-19 patients from the hospital and health database in the northern region of Iran. Secondly, the study analyzed high-dimensional data, including demographic, clinical, and paraclinical variables. Thirdly, the statistical analysis involved multiple algorithms and a simultaneous logistic model in educational data. Fourthly, the models created were tested and cross-validated. The study aimed to predict the death of COVID-19 patients and the findings were derived from decision tree modeling, identifying the death rate and influential factors. Based on the study's findings, it is expected that by controlling these predictors of mortality, the costs associated with this widespread disease on families and the healthcare system could be minimized.

The results indicate that the fitted DT models, have relatively good performance in diagnostic accuracies both in training and testing imbalanced data sets for predicting COVID-19 mortality. This high performance may be explained by a big dataset of over 5000 records in our study. Despite the presence of imbalanced data with respect to mortality versus survival. Our results show no evidence of overfitting in unbalanced training data sets, because of the presence of rather closed performance of diagnostic accuracies in training and testing datasets. In our findings with imbalanced training data of big data sets and suitable pruning, the fitted model of C5.0 and CART algorithms had outperformed in sensitivity while CHAID had better performance in specificity and precision than other algorithms. However, when the DT models were developed on balanced training datasets, the performance of sensitivity of all algorithms decreased surprisingly in testing datasets but not in training datasets. This may imply the synthetic and oversampling used to deal with imbalanced data of minority class have create the possibility of overfitting and generation of synthetic case that might not be accurate representative the minority class. Moreover, oversampling in SMOTE-Tomek method may introduce sampling errors which can lead to bias and also increase the risk of overfitting, where the model learns the noise in datasets [50].

In this study, we unfortunately faced limitations in terms of time and costs, which prevented us from gathering robust data from multiple centers for model training. As a result, data was collected solely from one hospital, and since it was retrospective, there were several variables with missing data. To overcome this, advanced statistical methods were employed, and future longitudinal studies will aim to minimize the occurrence of missing data. However, the clinical significance of lactate dehydrogenase variation in COVID-19 patients is noteworthy. Unfortunately, it has been excluded from the study due to the fact that it was only measured in 3% of the patients. The study was conducted between March 2020 and March 2022. However, the COVID-19 vaccination rollout began in Iran in January 2021. Therefore, it is plausible that some of the participants in our survey might have been vaccinated, but we did not have access to their vaccination information. Moreover,

although the model built on tree classification in our analysis has been validated independently by testing datasets of study regions, testing the model with external datasets would strengthen the generalizability of results. However, we did not access external datasets of other countries or other regions of Iran. This may rather limit the generalizability of predictive models of tree-based classification.

Conclusion

The findings from this study reveal that factors including ICU hospitalization, intubation, age, kidney diseases, O2 sat, WBC, BUN, CRP, NLR, and hemoglobin play a significant role in determining the mortality rate of COVID-19 patients. To establish the reliability of these results and assess the role of Decision Tree (DT) analysis in the diagnostic process, it is essential to conduct further longitudinal studies involving multiple hospital centers. Specialists in statistics who focus on prediction and classification should carefully consider the potential of decision tree models, which can be equally or even more effective than traditional regression methods in identifying predictive patterns without making as many assumptions. Furthermore, the authors recommend the future research directions, such as exploring ensemble methods or deep learning model for predicting COVID-19 mortality.

Acknowledgements
The authors acknowledge the deputy of Research and Technology of Babol University of Medical Sciences for their support.

Authors' contributions
Z.M. and K.H. conceptualized and designed the study.M.S.H., S.B. and A.A. provided data file from data -based. Z.M. and K.H. analyzed the data and wrote the first draft of manuscript. All authors have read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The data that support the findings of this study are available from the authors but restrictions apply to the availability of these data, which were used under license from Babol University of Medical Sciences, Babol, Iran, for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission from Babol University of Medical Sciences.

Declarations

Ethics approval and consent to participate
All methods were carried out in accordance with relevant guidelines and regulations. This study was approved by the Institutional Review Board of Ethics Committee of the Babol University of Medical Sciences, Babol, Iran. All patients had given a written consent at hospitalization to include the data of their hospital charts to the data-based of electronic file for this research.

Consent for publication
Not applicable.

Competing interests
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details
[1]Student Research Center, Research Institute, Babol University of Medical Sciences, Babol, Iran. [2]Department of Biostatistics and Epidemiology, School of Public Health, Babol University of Medical Sciences, Babol, Iran. [3]Social Determinants of Health Research Center, Research Institute, Babol University of Medical Sciences, Babol, Iran. [4]Department of Infectious Diseases, Ayatollah Rohani Hospital, Babol University of Medical Sciences, Babol, Iran.

Received: 19 March 2024 Accepted: 18 August 2024
Published online: 27 September 2024

References

1. World Health Organization. Clinical management of severe acute respiratory infection (SARI) when COVID-19 disease is suspected: interim guidance, 13 March 2020. World Health Organization; 2020.
2. General Office of National Health Commission of People's Republic of China OoNAoTCM. Diagnosis and Treatment of Corona Virus Disease-19 (7th Trial Edition). 2020(6):801–5.
3. Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J, et al. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. J Infect. 2020;81(2):e16–25.
4. Bertsimas D, Lukin G, Mingardi L, Nohadani O, Orfanoudaki A, Stellato B, et al. COVID-19 mortality risk assessment: An international multi-center study. PLoS ONE. 2020;15(12): e0243262.
5. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507–13.
6. Mahase E. Coronavirus covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. BMJ. 2020;368: m641.
7. Chen T, Wu D, Chen H, Yan W, Yang D, Chen G, et al. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. BMJ. 2020;368: m1091.
8. Sun H, Ning R, Tao Y, Yu C, Deng X, Zhao C, et al. Risk Factors for Mortality in 244 Older Adults With COVID-19 in Wuhan, China: A Retrospective Study. J Am Geriatr Soc. 2020;68(6):E19-e23.
9. Wang W, Tang J, Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan. China J Med Virol. 2020;92(4):441–7.
10. Bellan M, Patti G, Hayden E, Azzolina D, Pirisi M, Acquaviva A, et al. Fatality rate and predictors of mortality in an Italian cohort of hospitalized COVID-19 patients. Sci Rep. 2020;10(1):20731.
11. 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.
12. Mendes A, Serratrice C, Herrmann FR, Genton L, Périvier S, Scheffler M, et al. Predictors of In-Hospital Mortality in Older Patients With COVID-19: The COVIDAge Study. J Am Med Dir Assoc. 2020;21(11):1546–54.e3.
13. Chen T, Dai Z, Mo P, Li X, Ma Z, Song S, et al. Clinical Characteristics and Outcomes of Older Patients with Coronavirus Disease 2019 (COVID-19) in Wuhan, China: A Single-Centered, Retrospective Study. J Gerontol A Biol Sci Med Sci. 2020;75(9):1788–95.
14. Hippisley-Cox J, Coupland CA, Mehta N, Keogh RH, Diaz-Ordaz K, Khunti K, et al. Risk prediction of covid-19 related death and hospital admission in adults after covid-19 vaccination: national prospective cohort study. BMJ. 2021;374: n2244.
15. Atkins JL, Masoli JAH, Delgado J, Pilling LC, Kuo CL, Kuchel GA, et al. Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort. J Gerontol A Biol Sci Med Sci. 2020;75(11):2224–30.
16. Josephus BO, Nawir AH, Wijaya E, Moniaga JV, Ohyver M. Predict Mortality in Patients Infected with COVID-19 Virus Based on Observed Characteristics of the Patient using Logistic Regression. Procedia Comput Sci. 2021;179:871–7.

17. Du RH, Liang LR, Yang CQ, Wang W, Cao TZ, Li M, et al. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. Eur Respir J. 2020;55(5).
18. Mahendra M, Nuchin A, Kumar R, Shreedhar S, Mahesh PA. Predictors of mortality in patients with severe COVID-19 pneumonia - a retrospective study. Adv Respir Med. 2021;89(2):135–44.
19. Trecarichi EM, Mazzitelli M, Serapide F, Pelle MC, Tassone B, Arrighi E, et al. Clinical characteristics and predictors of mortality associated with COVID-19 in elderly patients from a long-term care facility. Sci Rep. 2020;10(1):20834.
20. Knol MJ, Vandenbroucke JP, Scott P, Egger M. What Do Case-Control Studies Estimate? Survey of Methods and Assumptions in Published Case-Control Research. Am J Epidemiol. 2008;168(9):1073–81.
21. Gu W, Vieira AR, Hoekstra RM, Griffin PM, Cole D. Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. Epidemiol Infect. 2015;143(13):2786–94.
22. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017;38(23):1805–14.
23. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res. 2017;121(9):1092–101.
24. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007;160(1):3–24.
25. Ayer T, Chhatwal J, Alagoz O, Kahn CE Jr, Woods RW, Burnside ES. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics. 2010;30(1):13–22.
26. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D. Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. Ai Magazine. 2012;33(4):33-.
27. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. J Med Syst. 2002;26(5):445–63.
28. Zhao Y, Zhang Y. Comparison of decision tree methods for finding active objects. Adv Space Res. 2008;41(12):1955–9.
29. Song Y-Y, Ying L. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130.
30. Chern CC, Chen YJ, Hsiao B. Decision tree-based classifier in providing telehealth service. BMC Med Inform Decis Mak. 2019;19(1):104.
31. Rokach L, Maimon O. Decision trees. Data mining and knowledge discovery handbook: Springer; 2005. p. 165–92.
32. Tso GK, Yau KK. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy. 2007;32(9):1761–8.
33. Batra M, Agrawal R. Comparative analysis of decision tree algorithms. Nature inspired computing: Springer; 2018. p. 31–6.
34. Alkhadar H, Macluskey M, White S, Ellis I, Gardner A. Comparison of machine learning algorithms for the prediction of five-year survival in oral squamous cell carcinoma. J Oral Pathol Med. 2021;50(4):378–84.
35. Liu Y, De A. Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. International journal of statistics in medical research. 2015;4(3):287.
36. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16(3):219–42.
37. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med. 2011;18(10):1099–104.
38. Tabachnick BG, Fidell LS, Ullman JB. Using multivariate statistics: pearson Boston, MA; 2007.
39. Hosmer DW Jr. Lemeshow S, Sturdivant RX. Applied logistic regression: John Wiley & Sons; 2013.
40. Njoku OC. Decision trees and their application for classification and regression problems. 2019.
41. Rokach L, Maimon O. Decision Trees. 62005. p. 165–92.
42. Hajian-Tilaki K. The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation. Stat Methods Med Res. 2018;27(8):2374–83.
43. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med. 2013;4(2):627.
44. Baratloo A, Hosseini M, Negida A, El Ashal G. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. 2015.
45. Karaca-Mandic P, Sen S, Georgiou A, Zhu Y, Basu A. Association of COVID-19-related hospital use and overall COVID-19 mortality in the USA. Journal of general internal medicine. 2020:1–3.
46. Tezza F, Lorenzoni G, Azzolina D, Barbar S, Leone LAC, Gregori D. Predicting in-Hospital Mortality of Patients with COVID-19 Using Machine Learning Techniques. J Pers Med. 2021;11(5).
47. Xiang S, Li L, Wang L, Liu J, Tan Y, Hu J. A decision tree model of cerebral palsy based on risk factors. J Matern Fetal Neonatal Med. 2021;34(23):3922–7.
48. Kabootari M, Habibi Tirtashi R, Hasheminia M, Bozorgmanesh M, Khalili D, Akbari H, et al. Clinical features, risk factors and a prediction model for in-hospital mortality among diabetic patients infected with COVID-19: data from a referral centre in Iran. Public Health. 2022;202:84–92.
49. Shanbehzadeh M, Kazemi-Arpanahi H, Nopour R. Performance evaluation of selected decision tree algorithms for COVID-19 diagnosis using routine clinical data. Med J Islam Repub Iran. 2021;35:29.
50. Alkhawaldeh IM, Albalkhi I, Nawwham AJ. Challenges and limitations of synthetic minority oversampling techniques in machin learning. World J Methodol. 2023;13(5):375–8.

Publisher's Note