Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2021;7(2): 154-169

Original Article

Comparing of Data Mining Techniques for Predicting In-Hospital Mortality Among Patients with COVID-19

Mostafa Shanbehzadeh[1], Azam Orooji [2], Hadi Kazemi-Arpanahi [3,4]*

[1]Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran.
[2]Department of Advanced Technologies, School of Medicine, North Khorasan University of Medical Science, North Khorasan, Iran.
[3]Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran.
[4]Student Research Committee, Abadan University of Medical Sciences, Abadan, Iran.

ARTICLE INFO

Received 03.05.2021
Revised 23.05.2021
Accepted 01.06.2021
Published 19.06.2021

Key words:
COVID-19;
Coronavirus;
Artificial intelligence;
Machine learning;
Mortality

ABSTRACT

Introduction: The COVID-19 epidemic is currently fronting the worldwide health care systems with many qualms and unexpected challenges in medical decision-making and the effective sharing of medical resources. Machine Learning (ML)-based prediction models can be potentially advantageous to overcome these uncertainties.

Objective: This study aims to train several ML algorithms to predict the COVID-19 in-hospital mortality and compare their performance to choose the best performing algorithm. Finally, the contributing factors scored using some feature selection methods.

Material and Methods: Using a single-center registry, we studied the records of 1353 confirmed COVID-19 hospitalized patients from Ayatollah Taleghani hospital, Abadan city, Iran. We applied six feature scoring techniques and nine well-known ML algorithms. To evaluate the models' performances, the metrics derived from the confusion matrix calculated.

Results: The study participants were 1353 patients, the male sex found to be higher than the women (742 vs. 611), and the median age was 57.25 (interquartile 18-100). After feature scoring, out of 54 variables, absolute neutrophil/lymphocyte count and loss of taste and smell were found the top three predictors. On the other hand, platelet count, magnesium, and headache gained the lowest importance for predicting the COVID-19 mortality. Experimental results indicated that the Bayesian network algorithm with an accuracy of 89.31% and a sensitivity of 64.2 % has been more successful in predicting mortality.

Conclusion: ML provides a reasonable level of accuracy in predicting. So, using the ML-based prediction models facilitate more responsive health systems and would be beneficial for timely identification of vulnerable patients to inform appropriate judgment by the health care providers.

Abbreviation: Coronavirus Disease 2019 (COVID-19), World Health Organization (WHO), Machine Learning (ML), Artificial Intelligence (AI), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Locally Weighted Learning (LWL), Clinical Decision Support System (CDSS)

* Corresponding Author Email: H.kazemi@abadanums.ac.ir

Introduction

Since its occurrence in December 2019, the Coronavirus Disease 2019 (COVID-19) has rapidly become a worldwide public health and social security hazard. Soon afterward, the World Health Organization (WHO) officially confirmed this disease a pandemic with considerably extraordinary contagious power and death rates compared with its antecedents, including SARS and MERS (1, 2). This virus is a very contagious respiratory infection that transmits through multiple routes and till now remains to be scattered destructively across international borders(3). The clinical courses of COVID-19 are differing widely among different patients alternating from asymptomatic, or limited to a few simple flu-like symptoms, the disease may progress to severe respiratory illness in some patients or even leading to multi-organ failure and ultimately death (4-6).

In an epidemic, a country's health care system tolerates tremendous pressure due to the increase in the use of healthcare services and surge in hospitalizations. However, most COVID-19 patients usually are asymptomatic or have minor symptoms and can be advised to self-quarantine and get better under ambulatory or virtual care services. For severe or advanced ill patients who progressed to fast deterioration, instant hospitalization is of great significance to receive early interventions and supportive treatments that may increase the patient's survival chance (7-12). Furthermore, this pandemic has led to the shortage of hospital resources and the overtiredness of healthcare workers, which demands accurate forecast models to successfully triage hospitalized patients with poor prognoses and make the best use of restricted resources(13). Thus, using an Artificial Intelligence (AI)-based risk assessment tool is valuable to mitigate the burden of health systems from unnecessary hospital visits, charges, and mental and physical pressure of the health workers especially in countries with intensive medical resources shortages (5, 14, 15). Machine Learning (ML) is a sub-form of AI that provides new insight or knowledge via extract functional patterns and applicable models from large volumes of the raw dataset (16). ML is a valued solution that is ever more deployed in clinical researches to conduct deep analyses and make known new contributing factors of a specific target outcome (17, 18). ML algorithms consist of supervised and unsupervised methods, which we considered supervised methods. In the supervised approach, we use part of our data as training data set to develop our model, and then we test the model with a section of data that is new to the algorithm (19).

Accordingly, applying ML-based prediction models may aid decision-making by generating rapid and reliable predictions to determine the mortality risk of COVID-19 patients and effectively triage them. It can be beneficial to reduce the overwhelmed burden on healthcare systems by helping to predict the risk of deterioration and possible deaths (20, 21).

Therefore, this study aimed to train several ML algorithms using routine clinical variables extracted from a COVID-19 database registry. Then their performance was compared using confusion matrix evaluation criteria. Finally, the predictive factors ranked according to their importance for COVID-19 in-hospital mortality using six feature selection techniques.

Material and Methods

1.2 Data set description

This retrospective, single-center study was conducted between January 9, 2020, and January 20, 2021, in Ayatollah Taleqhani hospital, affiliated to Abadan University of Medical Sciences, which is the core center for COVID-19 diagnosis and treatment in South West of Khuzestan province, Iran. A total of

12885 supposed COVID-19 cases has been referred to this center during the study period. Of those, 3350 cases introduced as positive RT-PCR COVID-19 test. Finally, only hospitalized patients who were meeting our inclusion criteria were involved in this study (see Figure1).

2.2 Ethical consideration

This study was approved by the ethical committee board of the Abadan University of Medical Sciences (Ethics code: IR.ABADANUMS.REC.1400.222). To protect the privacy and confidentiality of patients, we concealed the unique identifying information of all patients in the process of data collection and presentation.

2.3 Study predictors

The initial feature set in predicting COVID-19 mortality was determined using an extensive literature review coupled with an expert consensus. Then a questionnaire was designed through initial features in five sections, including patient's demographic, comorbidities diseases, clinical presentation, laboratory tests, and treatment plans. The content validity of the questionnaire was assessed by an expert panel including two infectious diseases specialists and two virologists. In addition, a test-retest (at 10-day interval) was done to evaluate the reliability of the questionnaire. Finally, the proposed clinical features were validated using a two-round Delphi survey by a group of multidisciplinary expert team, including five infectious diseases specialists, three epidemiologists, and two virologists. The experts were asked to review the initial list of parameters to score each item according to their importance perceived by them based on a 5-point Likert scale, ranging from 1 to 5, where 1 indicated "not important" and 5 indicated "highly important". Only the variables with an average score of 3.75 (70%) or higher were allowed into the study.

Finally, a total of 54 variables, including sociodemographic characteristics (five variables), clinical presentation (14 variables), comorbidities diseases (seven variables), laboratory tests (26 variables), and treatment plans (two variables) were considered to extract potential inclusion in our models from patient's database registry. The outcome was represented by in-hospital mortality. A detailed list of variables can be found in Table 1.

Table1. Baseline predictor variables and outcomes measures

Data Classes	Risk factors
Demographic characteristics	Gender, age, weight, height, and blood type
Clinical presentation	Cough, contusion, nausea, vomit, headache, gastrointestinal symptoms, muscular pain, chill, fever, dyspnea, loss of taste, loss of smell, runny nose, and sore throat
Comorbidities diseases	Pneumonia, cardiac disease, hypertension, diabetes, smoking, alcohol addiction, and another underline disease
Laboratory tests	Creatinine, red-cell count, white cell count, hematocrit, hemoglobin, platelet count, absolute lymphocyte count, absolute neutrophil count, calcium, phosphorus, magnesium, sodium, potassium, blood urea nitrogen, total bilirubin, aspartate aminotransferase, alanine aminotransferase, albumin, glucose, lactate dehydrogenase, activated partial, thromboplastin time, prothrombin time, alkaline phosphatase, erythrocyte sedimentation rate, C-reactive protein, Hypersensitive troponin
Treatment	Oxygen therapy, ICU hospitalization
Outcome	Mortality (Death/Alive)

2.4 Data preparation

Inadequate case records which had a large proportion of omitted data (>70%) were discarded from the study. Also, the remaining missing values were imputed with the mean or mode of each variable. Noisy and abnormal values, errors, duplicates, and meaningless data were checked by two Health Information Management experts (M: SH and H: K-A) in collaboration with two experts of infectious diseases and epidemiology. For different interpretations about data preprocessing, we contacted the corresponding physicians.

2.5 Model development

In this work, Naive Bayes and Bayesian network classifiers are used, both of which are based on Bayes theory. This group of classifiers is built without the need for complex iterations and more parameters and is therefore suitable for applying to huge datasets(22, 23). From the group of functional classifiers, Multi-Layer Perceptron (MLP) and Support Vector Machines (SVM) are used. SVM is based on geometrical properties to find a hyperplane that best separates the features into different domains(24, 25), and MLP is a class of feedforward Artificial Neural Network (ANN) that converts or map input data into a set of outputs (input-process–output). This is an extended linear perceptron network that uses three or more layers of neurons with nonlinear stimulus functions and is, therefore, more powerful in categorizing nonlinear data (26, 27).

Extracting if-then rules from data is done by rule-based classifiers. If-Then rules are highly readable, so they are very useful in identifying relationships between variables. In this study, rule-based classifiers, including OneR, and PART have been used(28, 29). From the group of lazy learners, Kstar and Locally Weighted Learning algorithms were implemented. Lazy learners are a sub-type of incremental learning. These algorithms can model complex decision spaces that have hyper polygonal shapes and are not easily describable by others ML algorithms(3, 30).

J48 and Random Forest are widely used ML algorithms, which is a decision tree algorithm. Decision trees are both simple-to-make and easy-to-use. They provide a top-down approach to the decision process that has a tree structure, and each branch ultimately leads to a decision after applying a series of conditions to different variables. This makes these methods much more understandable and easier to figure out(29, 31, 32).

To construct the mortality prediction model, we applied 10 ML algorithms from five different categories including Bayes Net and Naive Bayes (from Bayes), MLP and SVM (from functions), Kstar and Locally Weighted Learning (LWL) (from Lazy learners), OneR and PART (from Rules), J48 and Random Forest (from Trees). The algorithms were implemented using WEKA application software.

2.6 Model Evaluation

To compare the performance of different ML algorithms in predicting disease mortality, the derived 10-fold crosses validation metrics along with some measures such as accuracy, sensitivity and specificity were calculated (Table 2).

Table2. The performance evaluation measures

Accuracy= (TP+TN)/(TP+TN+FP+FN)
Precision= TP/(TP+FP)
Sensitivity= TP/(TP+FN)
Specificity= TN/(TN+FP)

* True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)

2.7 Feature scoring

In this step, the selected important variables for the risk prediction model were scored from a total of 54 candidate clinical parameters through six feature scoring methods, including Correlation, Gain Ratio, Info gain, Symmetrical uncertainty, OneR, and Relief were simulated in the Weka application software environment.

(Supplementary Information) The rank of each variable is calculated from its average in the six methods according to the following equation:

Equitation 1: Averaged rank= (r1+r2+ …. + r6)/6

In equation1, ri represents the rank of each risk factor in the ith feature selection method.

Results

3.1 Patient selection criteria

The data on 2082 eligible patients were extracted from the Ayatollah Taleghani hospital registry database. Then, 228 incomplete case records which had a lot of missing data (more than 70%) were excluded from the analysis. Also, the missing values were imputed with the mean or mode of each variable. After applying exclusion criteria, ultimately the 1353 records were remained (Figure 1).

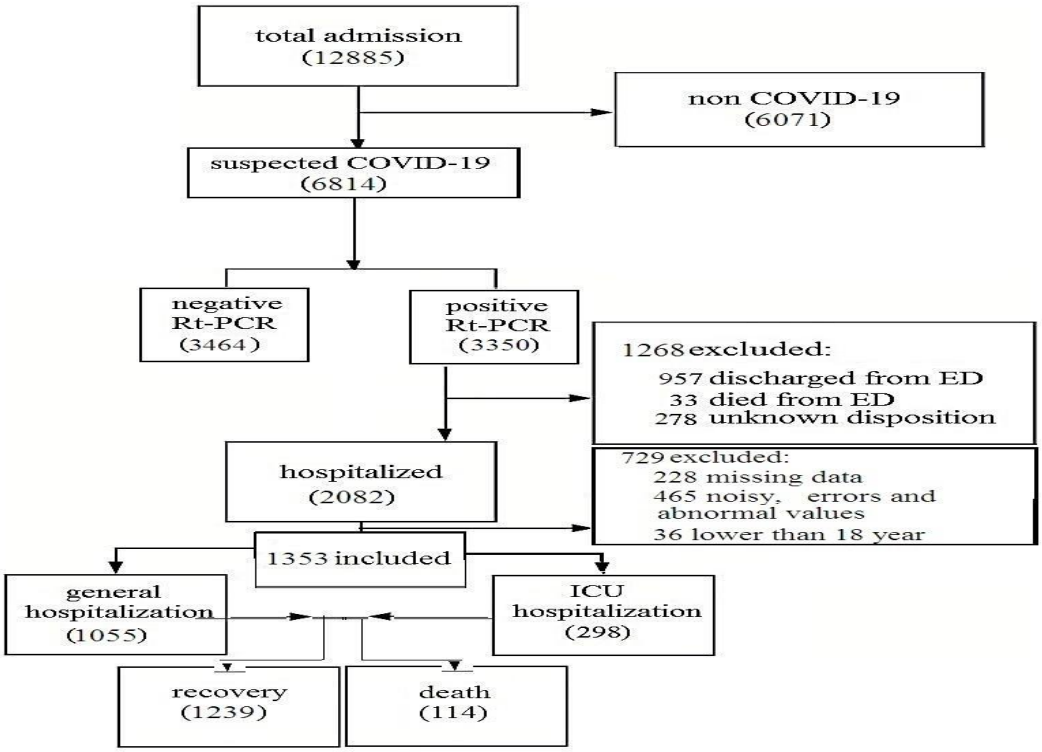

Figure1. Flowchart describing patient selection

3.2 The participant's characteristics

After applying the exclusion criteria and quantitative analysis of case records, finally, the number of 1353 patients met eligibilities. Of 1353 hospitalized COVID-19 patients in the retrospective study, 742 (54.85%) were male and 611 (45.15%) were women and the median age of participants was 57.25 (interquartile 18-100). 298 (22.02%) hospitalized in ICU and 1055 (77.98%) hospitalized in general wards. Of these, 1239 (91.57%) recovered, and 114 (8.43%) deceased. Descriptive statistics for the 1353 records in this dataset are shown in Tables 3 and 4.

Table3. The descriptive statistics of quantitative variables of the study after preprocessing

Variable name	Values	Frequencies
Blood Type	A-	37
	A+	592
	B-	33
	B+	156
	O-	39
	O+	421
	AB-	14
	AB+	61
Gender	Male	742
	Female	611
Cough	Yes	1036
	No	317
Contusion	Yes	487
	No	866
Nausea	Yes	479
	No	874
Vomit	Yes	399
	No	954
Headache	Yes	390
	No	963
Gastrointestinal symptoms	Yes	302
	No	1051
Muscular pain	Yes	701
	No	652
Chill	Yes	669
	No	684
Fever	Yes	706
	No	647
Pneumonia	Yes	1094
	No	259
Oxygen therapy	Yes	1103
	No	253
Dyspnea	Yes	1156
	No	197
Loss of taste	Yes	350
	No	1003
Loss of smell	Yes	355
	No	998
Runny Noise	Yes	487
	No	866
Sore throat	Yes	494
	No	859
Other underlining diseases	Yes	831
	No	522
Cardiac disease	Yes	346
	No	1007
Hypertension	Yes	445
	No	908
Diabetes	Yes	298
	No	1055
Smoking	Yes	61
	No	1292
alcohol addiction	Yes	21
	No	1332
C-reactive protein	Positive	1113
	Negative	240
Hypersensitive troponin	Positive	108
	Negative	1245
ICU hospitalization	Yes	298
	No	1055
Oxygen therapy	Yes	769
	No	584

Table4. The descriptive statistics of qualitative variables of the study after preprocessing

Variable name	**Range**	**Mean (SD)**
Age	18-100	57.25 (17.8)
Height	92-195	168.53 (8.5)
Weight	6.5-163	75.20 (13.0)
Creatinine	0.1-17.9	1.39 (1.4)
Red-cell count	1.38-13.1	4.56 (0.9)
White-cell count	1300-63000	8182.34 (4897.4)
Hematocrit	3.6-73.9	39.20 (6.7)
Hemoglobin	3.7-46	13.21 (2.4)
Platelet count	108000-691000	215493.66 (88380.1)

Absolute lymphocyte count	2-95	23.74 (11.8)
Absolute neutrophil count	8-98	74.52 (12.3)
Calcium	0.9-14.1	9.68 (0.8)
Phosphorus	2-12.4	3.50 (0.5)
Magnesium	1.14-19.1	2.16 (0.6)
Sodium	37-157	137.94 (5.3)
Potassium	2.5-14.2	3.98 (0.7)
Blood urea nitrogen	0.5-251	42.52 (31.7)
Total bilirubin	0.01-10	0.72 (0.7)
Aspartate aminotransferase	3.8-924	44.45 (53.5)
Alanine aminotransferase	2-672	38.29 (41.6)
Albumin	0.2-8.9	4.02 (0.5)
Glucose	18-994	136.09 (74.2)
Lactate dehydrogenase	4.6-6973	555.68 (339.0)
Activated partial thromboplastin time	1-120	28.56 (11.4)
Prothrombin time	0.9-46.8	12.82 (1.9)
Alkaline phosphatase	9.6-2846	213.12 (139.2)
Erythrocyte sedimentation rate	2-258	40.65 (28.8)

3.3 Feature selection, Model development, and evaluation

In addition, to construct a predictive model, this article attempts to identify important risk factors influencing COVID-19 mortality. After conducting an extensive literature review, 67 clinical features were identified as proposed predictors for determining the mortality risk of COVID-19 patients. A number of 14 clinical features were excluded after a two-round Delphi survey. Finally, following the reviewed studies and the expert's opinion, a total of 54 factors was remained to predict COVID-19 mortality. These clinical features, listed in Table3, were all clinically important and were divided into six categories, including demographics, risk factors, clinical manifestations, laboratory tests, and therapeutic plans.

After preparing the dataset, 10 ML algorithms were implemented, including LWL, Kstar, MLP, SVM, Naive Bayes, Bayesian network, OneR, PART, J48, and Random Forest. 10-fold cross-validation was utilized for evaluating ML methods in mortality prediction. Comparisons ML algorithms based on performance measures included accuracy, sensitivity, specificity was calculated which are shown in Table 5.

Table 5: Performance evaluation of selected ML algorithms for COVID-19 death prediction

ML method	**Accuracy**	**Sensitivity**	**Specificity**	**Confusion matrix** (Predicted / Actual: Alive, Death)	
				Alive: TN	Death: FP
				Death: FN	TP
Bayesian network	**89.31**	**64.2**	88.8	1003 95	129 126
Naive Bayes	83.51	38.6	91	1035 142	119 57
SVM	88.24	36.9	96.9	1066 141	53 85
MLP	85.88	42	93.2	1010 134	103 106
Kstar	82.78	17.6	93.7	1015 177	98 63
LWL	85.63	0	**100**	1109 236	0 0
OneR	84.65	13.6	96.6	1063 177	64 49
PART	84.08	42.6	91	1003 141	114 95
J48	83.84	33.5	92.3	1013 150	101 89
Random Forest	87.27	16.5	99.1	1075 170	39 69

The Bayesian network algorithm has higher accuracy and sensitivity and has been more successful in predicting mortality. On the other hand, based on the confusion matrix metrics, the LWL technique places all death class instances in the live class and is therefore not a suitable method for predicting (see Figure 2).

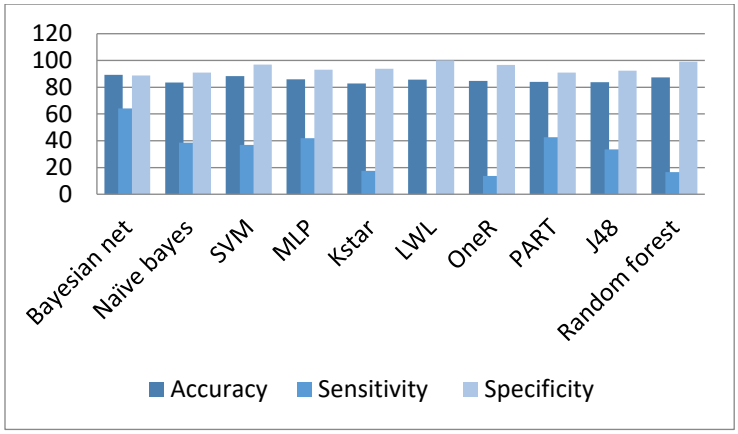

Figure 2. Performance visualization of selected ML algorithms

3.4 Feature Scoring

To identify the risk factors influencing the prediction of mortality of COVID-19 patients, six feature scoring methods were used, including Correlation, Gain Ratio, Info gain, Symmetrical uncertainty, OneR, and Relief (Table 6). In addition, the results of each feature scoring approach were portrayed in Appendix A.

Table6. Influential factors in predicting mortality in patients with COVID-19

Attributes	Correlation	Gain ratio	Info gain	OneR	Relief	Symmetrical Uncertainty	Averaged rank
Absolute neutrophil count	3	5	3	1	25	4	6.833333
Absolute lymphocyte count	5	7	4	3	27	5	8.5
Loss of taste	7	19	16	2	2	19	10.83333
Loss of smell	6	20	15	7	4	18	11.66667
Oxygen therapy	17	15	12	5	12	10	11.83333
White-cell count	2	1	2	52	23	1	13.5
Blood urea nitrogen	1	3	1	53	22	2	13.66667
Other underline disease	10	21	13	9	9	20	13.66667
Runny Noise	11	23	17	6	3	22	13.66667
Calcium	12	9	10	12	42	9	15.66667
Age	4	16	6	40	24	6	16
Creatinine	20	2	5	47	33	3	18.33333
Total bilirubin	22	12	14	21	37	11	19.5
Hypersensitive troponin	15	13	28	14	28	21	19.83333
Blood Type	21	28	8	38	1	24	20
Dyspnea	29	22	27	4	13	25	20
Lactate dehydrogenase	9	8	9	48	40	7	20.16667
Glucose	8	18	11	43	30	13	20.5
Albumin	24	14	7	33	41	8	21.16667
Hypertension	13	27	21	25	17	26	21.5
Cough	14	26	24	20	18	28	21.66667
Cardiac disease	19	29	29	13	21	29	23.33333
Contusion	23	31	30	18	8	31	23.5
Erythrocyte sedimentation rate	26	25	18	23	32	23	24.5
Activated partial thromboplastin time	25	10	22	46	36	14	25.5
Nausea	28	33	32	24	6	32	25.83333
Prothrombin time	18	17	19	39	49	16	26.33333
Phosphorus	30	6	25	51	34	17	27.16667
Sore throat	37	36	35	10	11	37	27.66667
Hemoglobin	16	24	26	36	39	27	28
Potassium	34	11	20	45	46	12	28
Chill	32	35	33	22	14	35	28.5
Sodium	51	4	23	32	48	15	28.83333
Diabetes	36	34	36	15	19	34	29
Muscular pain	35	37	34	26	7	36	29.16667
Gender	40	41	38	16	5	40	30
Vomit	41	40	37	27	10	39	32.33333
Alanine aminotransferase	27	30	31	50	31	30	33.16667
Aspartate aminotransferase	42	43	40	28	16	41	35
Fever	48	47	45	11	15	46	35.33333
Pneumonia	46	46	44	8	26	44	35.66667

Gastrointestinal symptoms	43	44	42	29	20	43	36.83333
C-reactive protein	49	39	39	35	29	38	38.16667
Alkaline phosphatase	33	32	41	49	52	33	40
Smoking	45	38	43	31	45	42	40.66667
Weight	52	48	49	19	35	49	42
Red-cell count	31	45	47	44	43	47	42.83333
Height	53	49	48	17	44	48	43.16667
alcohol addiction	50	42	46	30	53	45	44.33333
Hematocrit	47	51	52	41	38	51	46.66667
Magnesium	39	53	50	37	50	52	46.83333
Platelet count	44	52	51	34	51	50	47
Headache	38	50	53	42	47	53	47.16667

Based on Table 6, from a total of 54 predictors, absolute neutrophil (6.833333) and lymphocyte (8.5) count and loss of sense of smell (10.83333) and taste (11.66667) were determined as the top predictors of COVID-19 mortality. Besides platelet count, magnesium, and headache with an average rank of 46.66667, 46.83333, and 47.16667 respectively, were gained the lowest importance for predicting the COVID-19 mortality.

Discussion

Early death prediction in COVID-19 hospitalized patients can help in facilitating triage of critically ill patients and optimal planning of resource allocation to respond to the pandemic (21, 33). This study intent retrospectively building and evaluate ML models based on the most important variables in determining the risk of COVID-19 mortality. Therefore, for this aim, 10 most popular ML algorithms such as LWL, Kstar, MLP, SVM, Naive Bayes, Bayesian network, OneR, PART, J48, and Random forest were developed. For COVID-19 mortality risk prediction upholding correctness, warranting noise-free data, applying a balanced dataset, and decreasing analytic time are crucial topics (4). Since the dataset contains redundant or irrelevant features, feature selection as a preparatory stage to ML is greatly effective in eliminating insignificant and redundant data, diminishing data dimensionality and ambiguousness, lessen the analytical time, and improving model effectiveness (36, 37). In this work at first, after removing unrelated and redundant attributes through the expert consensus, 54 clinical factors that provide valuable prognostic information for death prediction were identified.

The current study in addition shows it is essential to select the most important features to maximize the capability of the model when compared to the use of whole attributes from the dataset (34). These noteworthy topics rest on the precise feature selection of COVID-19 due to the high volume of data in COVID-19 databases. It is acknowledged that the model's accuracy depends on the dataset, preprocessing, analytical tools, and techniques (38, 39). Hence by executing six feature scoring methods including, Correlation, Gain Ratio ،Info gain ، Symmetrical uncertainty ،OneR and Relief methods, an ideal feature list has been selected according to the average rank of each variable in all feature scoring techniques.

In bibliography a number of studies have been undertaken to identify important clinical risk factors affecting mortality prediction for COVID-19. The selected features are used as inputs for developing ML-based models for predicting severity, deterioration and mortality of COVID-19 patients. The ten top clinical variables predicting mortalities in reviewing studies, including age (high) (5,

36, 40-45), sense of taste (low / loss) (4, 5, 14, 41, 45-47), body temperature (high) (5, 14, 36, 40, 41, 44, 48), oxygen saturation (decreased) (20, 21, 41, 43, 48, 49), lymphocyte/neutrophil count (raised) (21, 41-43, 46, 49), C reactive protein (raised) (21, 43, 44, 47) , D dimer (increased) (20, 40, 42, 47), ALT and/or AST (raised) (42, 43, 46, 48, 49), LDH (elevated) (36, 42, 46, 50), hypertension/ cardiovascular diseases (41-45, 47). On the other hand diarrhea(5, 21, 44, 48, 50), vomiting(36, 41, 42, 49), gender(5, 14, 36, 42), platelet count (low)(42, 46-49), blood type(20, 40, 42, 47), smoking history(5, 21, 44, 48, 50), and muscle pain (14, 42, 46, 47) has the least importance for predicting the COVID-19 mortality. In the current study after feature scoring, the absolute neutrophil and lymphocyte count and loss of taste or smell were determined as the top three predictors (average rank of importance: 6.83, 8.5, 10.83 and 11.66 respectively). The bottom ranking variables, that is, the items that were scored to be least essential included platelet count, magnesium rate, and headache (average rank of importance: 46.66, 46.83, and 47.16 respectively).

ML can be used with myriad applications for healthcare systems in tackling the COVID-19 pandemic. So far, Most ML-based prediction models have analyzed patients from different countries across the world. Gao (2014) proposed an ML predictive model based on the data of 2520 COVID-19 hospitalized patients for death anticipation. Finally, the most effective results are obtained by the Neural Network (NN) technique (AUC-ROC of 0.976%) (20). An (2020) conducted a retrospective analysis on data of 10237 patients to predict COVID-19 mortality. The results showed that the model developed with SVM with the sensitivity of 90.7%, specificity of 91.4%, and an ROC of 0.963% has the best performance(51). Agieb et al. (2020) have compared three ML classification successes in COVID-19 mortality prediction. Finally. the most successful results are obtained by using the SVM technique(50). Yadaw and their colleagues (2020) studied 3841 patient data to construct a prediction model through four ML algorithms to death anticipation. Finally. the XGBoost model with an ROC of 0.91% attained the best performance(43). Using MLP, Vaid et al. (2020) estimated the patient death prediction with an ROC of 0.822% outperformed all other models in this study (33). In another work by Zhao and et al. (2020) analyzed the data of 313 COVID-19 hospitalized patients showed that the ANN achieved the highest accuracy on prediction of mortality with an ROC of 0.75 (4). In this research, hospitalized COVID-19 patients from a large region in Khuzestan, Iran, were surveyed to determine the most important variables in mortality occurrence. Therefore, the investigations are not comprehensively generalizable and it is essential to study datasets from the whole of Iran to support the Iranian healthcare industry. The results showed that the Bayesian network algorithm with an accuracy of 89.31% and a sensitivity of 64.2% has a higher performance in predicting mortality.

The suggested models in this study can predict the death of patients with optimal evaluation metrics. These models may assist clinicians in enabling early detection, effective intervention, and possibly a decrease in death in COVID-19 patients. Designing a true and valid anticipative model may be improving the quality of care and increasing the survival rate of the patients. This led to decreasing ambiguity by offering quantitative, objective, and evidence-based models for risk stratification, prediction, and eventually episode of the care plan; thus guide clinical decision-making and hope to improve patient outcomes and quality of care in the limited medical resource organization (20, 54).

The power of our study lies in the real-world dataset with inclusive features, so analytical bias was constrained and the novel application of ML algorithms than classical analysis techniques. But there are two limitations that must to be addressed. First, only 1335 subjects were included, and thus, our sample was inadequate since ML algorithms demand a large-scale dataset to be include. Second, the developed models are based merely on publicly available parameters collected at the initial of hospitalization; the inclusion of other clinical and radiological features could contribute to increasing the accuracy of prediction models.

Conclusion

The results suggest that physicians can use such ML-based models to augment decision-making using routine clinical data. In this study, we build and evaluated some selected ML-based predictive models for mortality prediction in hospitalized patients using the most important clinical parameters collected on admission. This study recognized the most important clinical predictors that accurately predict COVID-19 mortality. In conclusion, using ML algorithms combining with qualitative and comprehensive hospital databases such as patient registries in this study can facilitate rapid and reliable COVID-19 mortality risk classification. In the future, the performance of our model will be enhanced if we test more ML techniques at large, multicenter, and qualitative datasets. Finally, it was anticipated that our findings would support the development of a Clinical Decision Support System (CDSS) for the prediction of COVID-19 related outcomes such as the need for ICU hospitalization or mechanical ventilator, mortality and, etc.

Acknowledgments

This article is extracted from a research project supported by the Abadan University of Medical Sciences (IR.ABADANUMS.REC.1400.222). We also thank the Research Deputy of the Abadan University of Medical Sciences for financially supporting this project. We also would like to thank all experts who participated in this study and played a role in the validation of the data elements.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References:

1. Peeri NC, Shrestha N, Rahman MS, Zaki R, Tan Z, Bibi S, et al. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? International journal of epidemiology. 2020.
2. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. The Lancet Infectious Diseases. 2020.
3. Albahri A, Hamid RA. Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. Journal of Medical Systems. 2020; 44(7):1-11. PMID: 32451808.
4. Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. PloS one. 2020;15(7):e0236618.
5. Hu H, Yao N, Qiu Y. Comparing Rapid Scoring Systems in Mortality Prediction of Critically Ill Patients With Novel Coronavirus Disease. Academic Emergency Medicine. 2020;27(6):461-8.
6. Jamshidi E, Asgary A, Tavakoli N, Zali A, Dastan F, Daaee A, et al. Symptom Prediction and Mortality Risk Calculation for

COVID-19 Using Machine Learning. medRxiv. 2021.

7. Liu Y, Wang Z, Ren J, Tian Y, Zhou M, Zhou T, et al. A COVID-19 Risk Assessment Decision Support System for General Practitioners: Design and Development Study. Journal of medical Internet research. 2020;22(6):e19786.

8. Alom MZ, Rahman M, Nasrin MS, Taha TM, Asari VK. COVID_MTNet: COVID-19 Detection with Multi-Task Deep Learning Approaches. arXiv preprint arXiv:200403747. 2020.

9. Bansal A, Padappayil RP, Garg C, Singal A, Gupta M, Klein A. Utility of Artificial Intelligence Amidst the COVID 19 Pandemic: A Review. Journal of Medical Systems. 2020;44(9).

10. Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. International Journal of Antimicrobial Agents. 2020;55(3):105924.

11. Hussain A, Bhowmik B, do Vale Moreira NC. COVID-19 and diabetes: Knowledge in progress. Diabetes Research and Clinical Practice. 2020;162.

12. Moujaess E, Kourie HR, Ghosn M. Cancer patients and research during COVID-19 pandemic: A systematic review of current evidence. Critical Reviews in Oncology/Hematology. 2020;150:102972.

13. Zheng Y, Zhu Y, Ji M, Wang R, Liu X, Zhang M, et al. A Learning-Based Model to Evaluate Hospitalization Priority in COVID-19 Pandemics. Patterns. 2020;1(6):100092.

14. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence. 2020:1-6.

15. Malki Z, Atlam E-S, Hassanien AE, Dagnew G, Elhosseini MA, Gad I. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, Solitons & Fractals. 2020;138:110137.

16. Shanbehzadeh M, Nopour R, Kazemi-Arpanahi H. Comparison of Four Data Mining Algorithms for Predicting Colorectal Cancer Risk. Journal of Advances in Medical and Biomedical Research.29(133):100-8.

17. Hernandez-Suarez DF, Ranka S, Kim Y, Latib A, Wiley J, Lopez-Candales A, et al. Machine-learning-based in-hospital mortality prediction for transcatheter mitral valve repair in the United States. Cardiovascular Revascularization Medicine. 2020.

18. Shanbehzadeh M, Nopour R, Kazemi-Arpanahi H. Comparison of Four Data Mining Algorithms for Predicting Colorectal Cancer Risk. Journal of Advances in Medical and Biomedical Research. 2021;29(133):100-8.

19. Coleman BC, Fodeh S, Lisi AJ, Goulet JL, Corcoran KL, Bathulapalli H, et al. Exploring supervised machine learning approaches to predicting Veterans Health Administration chiropractic service utilization. Chiropractic & manual therapies. 2020;28(1):47.

20. Gao Y, Cai G-Y, Fang W, Li H-Y, Wang S-Y, Chen L, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nature communications. 2020;11(1):1-10.

21. Ryan L, Lam C, Mataraso S, Allen A, Green-Saxena A, Pellegrini E, et al. Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: a retrospective study. Annals of Medicine and Surgery. 2020;59:207-16.

22. Sinha S. Reproducibility of parameter learning with missing observations in naive Wnt Bayesian network trained on colorectal cancer samples and doxycycline-treated cell

lines. Molecular bioSystems. 2015;11(7):1802-19.
23. Tian XW, Lim JS. Interactive Naive Bayesian network: A new approach of constructing gene-gene interaction network for cancer classification. Bio-medical materials and engineering. 2015;26 Suppl 1:S1929-36.
24. Golpour P, Ghayour-Mobarhan M, Saki A, Esmaily H, Taghipour A, Tajfard M, et al. Comparison of Support Vector Machine, Naïve Bayes and Logistic Regression for Assessing the Necessity for Coronary Angiography. International journal of environmental research and public health. 2020;17(18).
25. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer genomics & proteomics. 2018;15(1):41-51.
26. Lorencin I, Anđelić N, Španjol J, Car Z. Using multi-layer perceptron with Laplacian edge detector for bladder cancer diagnosis. Artificial intelligence in medicine. 2020;102:101746.
27. Talebi N, Nasrabadi AM, Mohammad-Rezazadeh I. Estimation of effective connectivity using multi-layer perceptron artificial neural network. Cognitive neurodynamics. 2018;12(1):21-42.
28. Li Q, Doi K. Analysis and minimization of overtraining effect in rule-based classifiers for computer-aided diagnosis. Medical physics. 2006;33(2):320-8.
29. Berhane TM, Lane CR, Wu Q, Autrey BC, Anenkhonov OA, Chepinoga VV, et al. Decision-Tree, Rule-Based, and Random Forest Classification of High-Resolution Multispectral Imagery for Wetland Mapping and Inventory. Remote sensing. 2018;10(4):580.
30. Feng M, Loy LY, Zhang F, Zhang Z, Vellaisamy K, Chin PL, et al. Go green! Reusing brain monitoring data containing missing values: a feasibility study with traumatic brain injury patients. Acta neurochirurgica Supplement. 2012;114:51-9.
31. Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh A. A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes. Journal of research in health sciences. 2018;18(2):e00412.
32. Pei D, Yang T, Zhang C. Estimation of Diabetes in a High-Risk Adult Chinese Population Using J48 Decision Tree Model. Diabetes, metabolic syndrome and obesity : targets and therapy. 2020;13:4621-30.
33. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach. JMIR medical informatics. 2021;9(1):e24207.
34. Aggarwal D, Bali V, Mittal S. An insight into machine learning techniques for Predictive Analysis and Feature Selection. International Journal of Innovative Technology and Exploring Engineering. 2019;8:342-9.
35. Brownlee J. Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python: Machine Learning Mastery; 2020.
36. Wu G, Yang P, Xie Y, Woodruff HC, Rao X, Guiot J, et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. European Respiratory Journal. 2020;56(2).
37. Hernandez-Suarez DF, Ranka S, Kim Y, Latib A, Wiley J, Lopez-Candales A, et al. Machine-learning-based in-hospital mortality prediction for transcatheter mitral valve repair in the United States. Cardiovascular Revascularization Medicine. 2021;22:22-8.

38. Subramani P, K S, B KR, R S, B DP. Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients. Personal and ubiquitous computing. 2021:1-14.
39. Sun L, Mo Z, Yan F, Xia L, Shan F, Ding Z, et al. Adaptive Feature Selection Guided Deep Forest for COVID-19 Classification With Chest CT. IEEE journal of biomedical and health informatics. 2020;24(10):2798-805.
40. Allenbach Y, Saadoun D, Maalouf G, Vieira M, Hellio A, Boddaert J, et al. Development of a multivariate prediction model of intensive care unit transfer or death: A French prospective cohort study of hospitalized COVID-19 patients. PloS one. 2020;15(10):e0240711.
41. Assaf D, Gutman Ya, Neuman Y, Segal G, Amit S, Gefen-Halevi S, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Internal and emergency medicine. 2020;15(8):1435-43.
42. Das AK, Mishra S, Gopalan SS. Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. PeerJ. 2020;8:e10083.
43. Yadaw AS, Li Y-c, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. The Lancet Digital Health. 2020;2(10):e516-e25.
44. Zhang Y, Xin Y, Li Q, Ma J, Li S, Lv X, et al. Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. Biomedical engineering online. 2017;16(1):125.
45. Zhou Y, He Y, Yang H, Yu H, Wang T, Chen Z, et al. Exploiting an early warning Nomogram for predicting the risk of ICU admission in patients with COVID-19: a multi-center study in China. Scandinavian journal of trauma, resuscitation and emergency medicine. 2020;28(1):1-13.
46. Booth AL, Abels E, McCaffrey P. Development of a prognostic model for mortality in COVID-19 infection using machine learning. Modern Pathology. 2020:1-10.
47. Pan P, Li Y, Xiao Y, Han B, Su L, Su M, et al. Prognostic Assessment of COVID-19 in the Intensive Care Unit by Machine Learning Methods: Model Development and Validation. Journal of medical Internet research, 2020, 22.11: e23128.
48. Chin V, Samia NI, Marchant R, Rosen O, Ioannidis JP, Tanner MA, et al. A case study in model failure? COVID-19 daily deaths and ICU bed utilisation predictions in New York State. European Journal of Epidemiology. 2020;35(8):733-42.
49. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj. 2020;369.
50. Agieb R. Machine learning models for the prediction the necessity of resorting to icu of covid-19 patients. International Journal of Advanced Trends in Computer Science and Engineering. 2020:6980-4.
51. East A, Ray S, Pope R, Cortina-Borja M, Sebire NJ. 45 Predicting long length of stay in a paediatric intensive care unit using machine learning. BMJ Publishing Group Ltd; 2020.
52. Bath C, Heger U, Petrovsky N, Senanayake S, Frydenberg J. New vaccine:: Australian scientists are beginning work on a vaccine that specifically targets the mutant strains of COVID-19 found to be more contagious and potentially deadlier than previous variants. 2021.
53. Poirier C, Luo W, Majumder MS, Liu D, Mandl KD, Mooring TA, et al. The role of environmental factors on transmission rates of the COVID-19 outbreak: an initial

assessment in two spatial scales. Scientific reports. 2020;10(1):1-11.

54. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. Journal of thoracic disease. 2019;11(Suppl 4):S574.