Frontiers in Health Informatics
2023; 12: 152
www.ijmi.ir
Open Access

Predictive modeling of COVID-19 hospitalization using twenty machine learning classification algorithms on cohort data

Zeynab Salehnasab[1], Ali Mousavizadeh[2], Ghasem Ghalamfarsa[3], Ali Garavand[4], Cirruse Salehnasab[5]*

[1]Medical Student, School of Medicine, Student Research Committee, Yasuj University of Medical Sciences, Yasuj, Iran
[2]Associate Professor, Department of Biostatistics and Epidemiology, School of Health, Social Determinants of Health Research Center, Yasuj University of Medical Sciences, Yasuj, Iran
[3]Assistant Professor, Department of Basic Sciences, School of Medicine, Medicinal Plants Research Center, Yasuj University of Medical Sciences, Yasuj, Iran
[4]Assistant Professor, School of Allied Medical Sciences, Lorestan University of Medical Sciences, Khorramabad, Iran
[5]Assistant Professor, Department of Biostatistics and Epidemiology, School of Health, Social Determinants of Health Research Center, Yasuj University of Medical Sciences, Yasuj, Iran

Article Info

Article type:
Research

Article History:
Received: 2023-07-15
Accepted: 2023-08-01
Published: 2023-08-12

**** Corresponding author:***
Cirruse Salehnasab

Assistant Professor, Department of Biostatistics and Epidemiology, School of Health, Social Determinants of Health Research Center, Yasuj University of Medical Sciences, Yasuj, Iran

Email:
cirruse.salehnasab@gmail.com

Keywords:
Machine Learning
COVID-19 Hospitalization
Cohort Data
Prediction

A B S T R A C T

Introduction: The global COVID-19 pandemic has led to a health crisis, emphasizing the need to identify high-risk patients for effective resource allocation and prioritized hospitalization. Previous studies have been limited in their use of algorithms and variables, while this research expands to include lifestyle factors and optimizes hyperparameters for twenty machine learning algorithms, enhancing prediction accuracy and identifying key predictors.

Material and Methods: In this cross-sectional study, we analyzed data from 207 COVID-19 patients. The Boruta algorithm was used to select the best features for twenty classification algorithms, and RandomizedSearchCV was utilized to optimize hyperparameters. The models were evaluated using performance metrics such as accuracy, f-measure, and area under the curve (AUC).

Results: The study identified eight key predictors of COVID-19 hospitalization, which include gamma-glutamyl transpeptidase, alkaline phosphatase, diagnosis by CT scan, mean platelet volume, mean corpuscular volume, fasting blood sugar, red blood cell count, and mean corpuscular hemoglobin concentration. By optimizing the hyperparameters of twenty machine learning algorithms, the accuracy and AUC were improved. With an outstanding AUC of 81.25, the XGBClassifier model exhibited superior performance.

Conclusion: The findings of this study can assist clinicians in allocating resources effectively and improving patient care. Additionally, this approach can aid healthcare researchers in leveraging artificial intelligence to manage diseases.

Cite this paper as:
Salehnasab Z, Mousavizadeh A, Ghalamfarsa G, Garavand A, Salehnasab C. Predictive modeling of COVID-19 hospitalization using machine learning classification algorithms on cohort data. Front Health Inform. 2023; 12: 152. DOI: 10.30699/fhi.v12i0.473

INTRODUCTION

On March 11, 2020, the World Health Organization (WHO) declared COVID-19 a public health emergency of international concern with high potential for global spread [1, 2]. The rapid spread of this pandemic has caused chaos and requires quick responses to reduce its impact. Since the beginning of the pandemic, hospitalization has been required for all COVID-19 positive cases, regardless of the severity of the disease. However, with the significant increase in cases worldwide, hospital beds have become completely occupied, leading to an excessive workload and pressure on healthcare staff [3]. Therefore, quick identification techniques for patients at high risk of severe and non-severe cases are crucial for prioritizing hospital admission [4]. Currently, the SARS-CoV-2 RNA test is used to

diagnose COVID-19, which is a qualitative test that determines whether the patient is infected with the virus or not [5].

Physicians employ multiple criteria to assess and manage the need for hospitalization in COVID-19 patients. Among these criteria, CT scans serve as an adjunctive tool for diagnosis and evaluation of disease severity. However, it is noteworthy that nearly 20% of COVID-19 patients exhibit no discernible lung imaging changes [6]. Although protein-based antibody tests and antigen tests with faster results are now available, concerns regarding their accuracy persist. Additionally, conventional laboratory methods, including complete blood cell count, blood biochemistry, and immunological tests, have been utilized to evaluate the clinical progression of the disease. Notably, several studies have indicated an increased prevalence of lymphopenia in COVID-19 patients [7, 8]. A study by Wynants et al. [9] found a significant relationship between Covid-19 severity and age, gender, lactate dehydrogenase (LDH), C-reactive protein (CRP), and number of lymphocytes. LDH, CRP, and lymphocytes were also identified as accurate predictors of Covid-19 mortality by a Chinese team [10].

In recent years, machine learning algorithms have emerged as a powerful tool for building predictive models that can help healthcare providers identify patients who are at high risk of hospitalization. These algorithms are designed to analyze vast amounts of patient data to identify patterns and risk factors that may indicate the likelihood of hospitalization. By leveraging this data, healthcare providers can develop more accurate and reliable models for predicting hospitalization and allocating resources more effectively. Machine learning is a subset of artificial intelligence that involves the development of algorithms and statistical models that enable computers to learn from data without being explicitly programmed. In other words, it is a way for computers to automatically improve their performance on a task by learning from experience [2, 11].

Machine learning packages are software libraries and tools that provide pre-built algorithms, functions, and data structures to help developers build machine learning models more efficiently. These packages allow data scientists to easily apply machine learning techniques to their data and develop predictive models. There are many machine learning packages available for various programming languages such as Python and R, and some popular packages are including Scikit-learn, AutoML, ZenML, MLBox, H20, TPOT, etc. [11, 12].

Scikit-learn is a widely used machine learning (ML) library in Python. It offers various classification algorithms to perform supervised learning tasks, which include predicting classes of samples based on labeled data [12].

The use of ML for Covid-19 severity classification and prognostic assessment has been explored in several studies [3, 9, 13-23], including those that used a variety of algorithms such as support vector machine (SVM), artificial neural networks (ANN), random forest (RF), decision tree, logistic regression, and K-nearest neighbor (KNN).

Pormahmayun and Shakibi reported that ANN demonstrated the best performance with an overall accuracy of 89.98% in predicting mortality rate [3]. Zhou et al. used genetic algorithms and SVM for disease severity progression forecasting [24]. Wungu et al. found that high levels of CK-MB, PCT, NT-proBNP, BNP, and d-dimer were predictive markers for Covid-19 severity [13]. Cai et al. utilized random forest models for severity classification and regression to predict clinical outcomes using CT-Scan images [14]. Yaşar et al. used deep learning, RF, and gradient-boosted trees to classify Covid-19 patients into three severity groups [15]. Banoei et al. employed the SIMPLS method to predict hospital mortality [16]. Bayat et al. compressed their dataset using pairwise correlations and employed the XGBoost model for prediction [17]. Yan et al. used XGBoost to predict mortality risk and achieved an accuracy of 93% [18]. Wang et al. utilized XGBoost to predict Covid-19 severity with an AUC of 83% [19]. Hu et al. compared various ML algorithms for mortality risk prediction, and the LR model was selected as the final model due to its simplicity and interpretability [20]. Age, hs-CRP level, number of lymphocytes, and D-dimer level were the most important predictors identified by the LR model. Yao et al. developed an SVM algorithm for Covid-19 severity prediction using blood or urine test results, with an accuracy of 81.48% [21].

After conducting an extensive review of existing literature, it became apparent that many studies focused on predicting the risk and severity of COVID-19 in hospitalized patients have relied solely on hospital laboratory variables. Furthermore, several studies only employed one or two algorithms without optimizing their hyperparameters to enhance the accuracy of predictions. To address these limitations and improve upon previous research, our study aimed to predict COVID-19 patient hospitalization by utilizing twenty machine learning classification algorithms, each with optimized hyperparameters. To achieve this objective, we gathered data from the Dena Cohort Center located in Kohgiluyeh and Boyar Ahmad, Iran.

MATERIAL AND METHODS

Our cross-sectional study followed a methodology consisting of six phases, outlined in Fig 1.

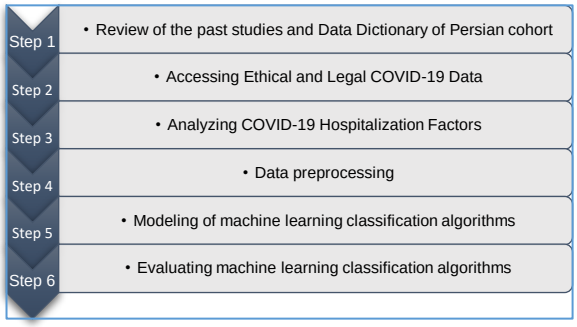

Fig 1: The methodology of study

To conduct a comprehensive cross-sectional study in applied settings, we drew upon multiple sources, including library resources, the latest research findings, and the Data Dictionary of the Persian cohort (see Appendix 1). By examining these sources, we identified critical variables and risk factors that influence the hospitalization of COVID-19 patients. Additionally, we conducted an extensive analysis to determine the most frequently utilized predictive machine learning models for this topic. Our thorough approach ensures that our study is based on the most up-to-date information and employs the most reliable and effective methods available.

The authors of this study approached the Persian Cohort Center to obtain the necessary permissions and investigate the data that was available in the center's database. This step was crucial to ensure that our study was conducted with ethical considerations and in compliance with relevant regulations. Through this process, we were able to access the necessary data to conduct a thorough analysis of variables and risk factors that influence the hospitalization of COVID-19 patients.

In order to conduct our study, we embarked on a thorough examination of the Dena Cohort Center database. We meticulously compared the variables in the data dictionary with existing research, ensuring a comprehensive and reliable investigation. From this rigorous evaluation, we successfully extracted data from 207 COVID-19 patients, aged between 35 and 70 years, who had contracted the virus between 20th March 2020 and 22nd November 2022. The data was accessed on 30th November 2022. Our analysis encompassed 92 independent variables, acquired through a meticulous methodology that prioritized the most up-to-date and relevant information. This approach laid a robust foundation for our subsequent examination of the variables and risk factors that influence the hospitalization of COVID-19 patients.

Table 1 lists the variables selected for the machine learning modeling process, which were identified through up-to-date information sources and analysis of the Persian Cohort database. These variables were chosen based on their potential predictive value in hospitalization risk for COVID-19 patients. By selecting these key variables, the study aimed to improve the accuracy and effectiveness of the machine learning algorithms used in the analysis.

Table 1: Features and specifications of the Dataset

Row	Variable	Description	Type	Role
1	Diagnosis_PCR	Diagnosis by PCR	Dichotomous	Input
2	Diagnosis_CTScan	Diagnosis by CT scan	Dichotomous	Input
3	Diagnosis_Symptoms	Diagnosis based on Symptoms	Dichotomous	Input
4	SmokeInHome	Whether or not the participant is/was exposed to smoke from a cigarette at home (passive/ second-hand smoking)	Dichotomous	Input
5	UseNonCigTobacco	Tobacco use defined as using Naas, Hookah, Pipe, or sticky once per week for at least six months	Dichotomous	Input
6	UseDrugs	Illicit drugs use defined as using illicit drugs once per week for at least six months	Dichotomous	Input
7	RightDBP1	Right arm first diastolic blood pressure	Discrete	Input
8	RightDBP2	Right arm second diastolic blood pressure	Discrete	Input
9	RightSBP1	Right arm first systolic blood pressure	Discrete	Input
10	RightSBP2	Right arm second systolic blood pressure	Discrete	Input
11	LeftDBP1	Left arm first diastolic blood pressure	Discrete	Input
12	LeftDBP2	Left arm second diastolic blood pressure	Discrete	Input
13	LeftSBP1	Left arm first systolic blood pressure	Discrete	Input
14	nearFarm	Whether or not the participant's house is in close proximity of farming areas	Dichotomous	Input
15	SupCalciumDValue	Specifies the number of calcium + vitamin D supplements taken over the time frame indicated	Discrete	Input
16	SupVitaminDPillInterval	Taking calcium supplements	Categorical*	Input
17	SupOmega3Interval	Taking omega-3 or fish oil supplements	Categorical	Input
18	SupIronInterval	Taking iron supplements	Categorical	Input
19	SupZincValue	Taking zinc supplements	Categorical	Input
20	SaltUseID	Adding salt to food at the table	Categorical	Input
21	GrilledFoodIntID	Frequency of eating barbecued foods and Kebabs (cooking on fire/grill)	Categorical	Input
22	FoodAlergy	List of any food allergies	Characteristic	Input
23	FoodSaltUsedID	Saltiness of food as perceived by participant	Categorical	Input

Row	Variable	Description	Type	Role
24	HasDiabetes	Diagnosed with diabetes	Categorical	Input
25	HasHypertension	Diagnosed with hypertension	Categorical	Input
26	HasCardiacDisease	Diagnosed with cardiac diseases	Dichotomous	Input
27	HasMI	Diagnosed with myocardial infarction (MI)	Dichotomous	Input
28	HasRenalFailure	Diagnosed with renal failure	Dichotomous	Input
29	HasHepatitisC	Diagnosed with hepatitis C	Dichotomous	Input
30	HasChronicLungDisease	Diagnosed with chronic lung disease	Dichotomous	Input
31	HasGallstone	Diagnosed with gallstones	Dichotomous	Input
32	HasStomachCancer	Diagnosed with stomach cancer	Dichotomous	Input
33	HasColorectalCancer	Diagnosed with colorectal cancer	Dichotomous	Input
34	HasBladderCancer	Diagnosed with bladder cancer	Dichotomous	Input
35	HasLungCancer	Diagnosed with lung cancer	Dichotomous	Input
36	HasEpilepsy	Diagnosed with epilepsy	Dichotomous	Input
37	HasChronicHeadaches	Diagnosed with chronic recurrent headaches	Dichotomous	Input
38	HasPsychiatricDisorder	Diagnosed with any kind of psychiatric disorder, other than depression	Dichotomous	Input
39	HasLearningDisability	Diagnosed with any kind of learning disability that has prevented the participant from educational achievements	Dichotomous	Input
40	HasAmnesia	Diagnosed with any kind of amnesia that has led to severe difficulty in daily functioning	Dichotomous	Input
41	HasTongueCancer	Diagnosed with tongue cancer	Dichotomous	Input
42	HasUterineCancer	Diagnosed with uterine cancer	Dichotomous	Input
43	HasMS	Diagnosed with multiple sclerosis (MS)	Dichotomous	Input
44	PLT	Platelet count	Continuous	Input
45	FBS	Fasting blood sugar	Continuous	Input
46	TG	Triglyceride	Continuous	Input
47	CHOL	Total cholesterol	Continuous	Input
48	SGOT	The aspartate aminotransferase (AST)	Continuous	Input
49	SGPT	The alanine aminotransferase (ALT)	Continuous	Input
50	ALP	Alkaline phosphatase	Continuous	Input
51	HDLC	High-density lipoprotein	Continuous	Input
52	Appearance	Urine appearance	Categorical	Input
53	SG	Specific gravity	Continuous	Input
54	PH	Urine pH level	Continuous	Input
55	Nitrite	Urine nitrite	Categorical	Input
56	Bilirubin	Urine bilirubin	Categorical	Input
57	Urobilinogen	Urine urobilinogen	Categorical	Input
58	Protein	Urine protein	Categorical	Input
59	Glucose	Urine glucose	Categorical	Input
60	Blood	Urine blood	Categorical	Input
61	Epithelial	Epithelial cells in urine	Categorical	Input
62	Bacteria	Bacteria in urine	Categorical	Input
63	Mucus	Mucus in urine	Categorical	Input
64	Cast	Cast in urine	Categorical	Input
65	Alb	Albumin in urine	Categorical	Input
66	Creat	Creatinine in urine	Categorical	Input
67	AscorbicAcid	Urine ascorbic acid level	Categorical	Input
68	KetoneBodies	Ketone bodies in urine	Categorical	Input
69	UrineWBC	Urine white blood cell count	Nominal	Input
70	UrineRBC	Urine red blood cell count	Nominal	Input
71	SleepDuration	The time between sleep and wakeup hour	Continuous	Input
72	HeightCm	Height	Continuous	Input
73	WeightKg	Weight	Continuous	Input
74	HipCircumference	Hip circumference	Continuous	Input
75	WristCircumference	Wrist circumference	Continuous	Input
76	BMI	Body mass index	Continuous	Input
77	WBC	White blood cell count	Continuous	Input
78	RBC	Red blood cell count	Continuous	Input
79	HGB	Hemoglobin	Continuous	Input
80	HCT	Hematocrit	Continuous	Input
81	MCV	Mean corpuscular volume	Continuous	Input
82	MCH	Mean corpuscular hemoglobin	Continuous	Input
83	MCHC	Mean corpuscular hemoglobin concentration	Continuous	Input
84	Lymphocytes	Lymphocytes	Continuous	Input
85	RDWCV	Red cell distribution width	Continuous	Input
86	MPV	Mean platelet volume	Continuous	Input
87	PDW	Platelet distribution width	Continuous	Input
88	BUN	Blood urea nitrogen	Continuous	Input
89	CERAT_Blood	Level of blood creatinine	Continuous	Input
90	GGT	Gamma-glutamyl transpeptidase	Continuous	Input

Row	Variable	Description	Type	Role
91	GenderID	Participant gender	Dichotomous	Input
92	Age	Age on the date of interview	Continuous	Input
93	Covid19Hosp	COVID-19 Hospitalization	Dichotomous	Target

*In this study, categorical variables are represented using one-hot encoding. This means that each category within the variable is mapped to a vector where the presence of a particular feature is denoted by 1 and absence by 0. This approach provides a numerical representation of categorical data, making it easier to analyze and process. See Appendix 1 for more information and coding of variables.

To prepare the data for modelling, we implemented a pre-processing phase that was essential to improve the quality of the raw data. As the results obtained from machine learning classification algorithms are highly dependent on the quality of the data, pre-processing was an essential step in our analysis. To achieve this, we followed a series of rigorous steps that included the following procedures:

Imputing missing value

During this step, we executed the following two tasks independently.

- We excluded any records and variables that had missing values exceeding 50% from our dataset, as these values could compromise the accuracy of our analysis.
- To address missing values in continuous and discrete variables, we replaced them with the mean and mode, respectively, within each class. By implementing these procedures, we were able to effectively address any missing values and ensure that our data were robust and reliable for subsequent machine learning classification analysis.

Under-sampling

To address the issue of imbalanced datasets, we implemented under-sampling methods aimed at normalizing the distribution of all classes by reducing the number of majority class records. An imbalanced class distribution typically features one or more classes with few samples (i.e., minority classes) and one or more classes with many samples (i.e., majority classes). In our study, we utilized the RandomUnderSampler method to decrease the number of majority class records effectively. This approach involves randomly selecting a subset of data for the targeted classes, making it a fast and straightforward method for balancing patient datasets. By using this technique, we were able to address the issue of imbalanced data and ensure that our machine learning classification algorithms could accurately identify patterns and insights within the dataset [12, 25, 26].

Data splitting

During this phase, we split the patient datasets into two separate sets: a training set and a testing set. The training set, which comprised 70% of the total data, was used to train our machine learning classification algorithms. Meanwhile, the testing set, consisting of 30% of the data, was used to evaluate the performance of our trained algorithms. This splitting of the dataset was necessary to ensure that our models could accurately generalize and predict outcomes for new, unseen data. By utilizing a split of 70-30%, we were able to optimize our algorithms for predictive accuracy while ensuring that our results were reliable and robust [12].

Feature scaling

During this phase, we applied a normalization technique (Equation 1) to separately scale the training and testing datasets. This technique ensures that all the numerical values in our datasets fall within a range of zero to one, allowing for fair comparisons and improved performance of our machine learning classification algorithms. By scaling the data in this manner, we were able to prevent any particular feature from dominating the model training and adversely affecting the results. Normalizing the data also makes it easier to interpret the model's output and helps to identify the most influential variables in predicting the hospitalization of COVID-19 patients [12, 26].

$$X_{normalized} = (X-X_{min})/(X_{max}-X_{min}) \quad (1)$$

Feature selection

In this phase, we utilized the Boruta algorithm, a type of wrapper method for feature selection, to identify the most critical predictors for predicting COVID-19 patient hospitalization. To do this, we employed the RandomForestClassifier algorithm to identify important features in the dataset that were both stable and unbiased. By using this method, we were able to select the most important features for our subsequent analysis, providing a more streamlined and effective approach to predicting hospitalization risk [12, 26, 27].

Hyperparameters are a type of parameters that have a significant impact on the performance and results of machine learning algorithms. They govern the learning process, but they are not the actual part of it. Adjusting these hyperparameters is essential for optimizing the machine learning model. This process involves an optimization problem and often requires a manual search. One common method used to search for optimal hyperparameters is the RandomizedSearchCV method, which uses k-fold cross-validation [11, 12]. In this study, the hyperparameters of twenty different machine

learning classification algorithms were optimized using the RandomizedSearchCV method. These algorithms include AdaBoostClassifier, BaggingClassifier, BernoulliNB, DecisionTreeClassifier, ExtraTreesClassifier, Gaussian Naive Bayes, GradientBoostingClassifier, HistGradientBoostingClassifier, KNeighborsClassifier, LinearDiscriminantAnalysis (LDA), LogisticRegression, LogisticRegressionCV, Multilayer Perceptron (MLP) Classifier, NuSVC, RandomForestClassifier, RidgeClassifier, RidgeClassifierCV, SGDClassifier, SVC, and XGBClassifier.

The performance of the machine learning models was evaluated by employing various metrics such as Accuracy, F-measure, and AUC (area under the curve) criteria, as defined in Equations 2 and 3. After developing the ML models, these metrics were used to assess the effectiveness and efficiency of the algorithms, providing valuable insights into their overall performance [28].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$F - measure = \frac{2*TP}{2*TP+FP+FN} \quad (3)$$

A Receiver Operating Characteristic (ROC) chart plots the relationship between false positive rate (FPR) and true positive rate (TPR), represented by the x and y axes, respectively. It illustrates the trade-offs between true positive (TP) and false-positive (FP) rates, helping to determine the optimal threshold for classification. By depicting these relative trade-offs, a ROC chart can be a useful tool for evaluating the effectiveness of a classification algorithm [28].

In the context of COVID-19 hospitalization predictions, true positive (TP) refers to the number of confirmed COVID-19 patients who were correctly hospitalized. True negative (TN) refers to the number of correctly predicted COVID-19 patients who did not require hospitalization. False positive (FP) is the number of COVID-19 patients who were incorrectly predicted to require hospitalization, while false negative (FN) refers to the number of COVID-19 patients who were incorrectly predicted not to require hospitalization [28]. These metrics provide valuable information on the accuracy and effectiveness of hospitalization prediction models for COVID-19 patients.

RESULTS

Preprocessing of patient data

After removing incomplete patient records, the initial patient dataset was reduced to 200 individuals, of which 65 were case-patients who were hospitalized and 135 were control patients who were not hospitalized. Due to under-sampling, the total number of normalized patient records was reduced to 130 individuals, consisting of 65 cases and 65 controls. From these records, 91 patients, which represented 70% of the dataset, were selected for training, while the remaining 39 cases were used for testing, representing 30% of the dataset.

By using Boruta algorithm, the feature selection analysis conducted on 92 variables revealed that gamma-glutamyl transpeptidase (GGT), alkaline phosphatase (ALP), diagnosis by CT scan, mean platelet volume (MPV), mean corpuscular volume (MCV), fasting blood sugar (FBS), red blood cell count (RBC), and mean corpuscular hemoglobin concentration (MCHC) were the top eight predictors of COVID-19 hospitalization, respectively. The importance of these variables is demonstrated in Fig 2, where gamma-glutamyl transpeptidase was found to have the highest level of importance among the eight predictors.

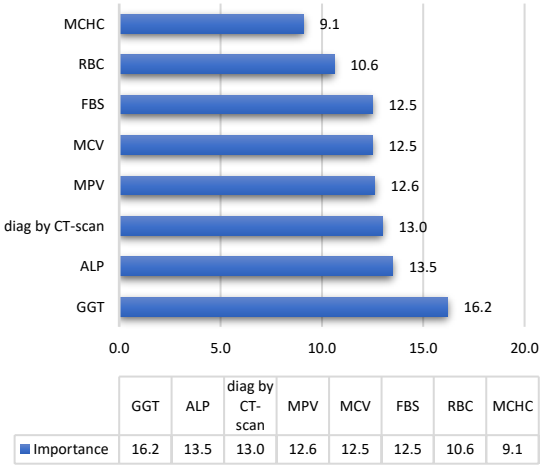

Fig 2: Key predictors of COVID-19 hospitalization using Boruta algorithm

Hyperparameter optimization for machine learning algorithms

In this study, the hyperparameter optimization process was conducted to identify the most effective combination of hyperparameters for each of the machine learning algorithms used. The performance of each algorithm was evaluated using the F-measure metric. By identifying the optimal hyperparameters for each algorithm, the models were able to achieve improved accuracy, precision, and generalization, making them more effective for the specific tasks they were designed to perform. The findings of this research provide valuable insights into the performance of machine learning algorithms and can be used to inform future research in this area, particularly in the optimization of hyperparameters to enhance the performance of machine learning models.

ML models' COVID-19 hospitalization prediction

performance

The results of evaluating the machine learning models on the test dataset are presented in Table 2 and Fig 3. The performance of each model was assessed using Accuracy, F-measure, and area under the curve (AUC) metrics. Among the algorithms tested, XGBClassifier (eXtreme Gradient Boosting Classifier) and SVC exhibited the highest performance based on these evaluation criteria. Notably, the XGBClassifier algorithm achieved the highest mean score of 80.60 across all evaluation metrics, as reported in Table 2.

Table 2: Evaluation results of machine learning models' performance

Row	Machine Learning Model	Accuracy	F-measure	AUC	Mean
1	XGBClassifier	80.56	80.00	81.25	80.60
2	SVC	77.78	80.00	80.00	79.26
3	ExtraTreesClassifier	77.78	73.33	76.88	76.00
4	RandomForestClassifier	75.00	75.68	76.25	75.64
5	AdaBoostClassifier	75.00	75.68	76.25	75.64
6	BernoulliNB	72.22	73.68	73.75	73.22
7	KNeighborsClassifier	72.22	72.22	73.13	72.52
8	GaussianNB	69.44	68.57	70.00	69.34
9	HistGradientBoostingClassifier	69.44	66.67	69.38	68.50
10	GradientBoostingClassifier	69.44	64.52	68.75	67.57
11	RidgeClassifier	66.67	68.42	68.13	67.74
12	NuSVC	67.36	66.46	67.97	67.26
13	DecisionTreeClassifier	66.67	64.71	66.88	66.09
14	MLPClassifier	66.67	62.50	66.25	65.14
15	LogisticRegression	61.11	61.11	61.88	61.37
16	LogisticRegressionCV	58.33	57.14	58.75	58.07
17	SGDClassifier	58.33	57.14	58.75	58.07
18	LinearDiscriminantAnalysis	58.33	57.14	58.75	58.07
19	RidgeClassifierCV	58.33	57.14	58.75	58.07
20	BaggingClassifier	55.56	61.90	58.13	58.53

AUC: Area under the curve, XGBClassifier: eXtreme Gradient Boosting classifier

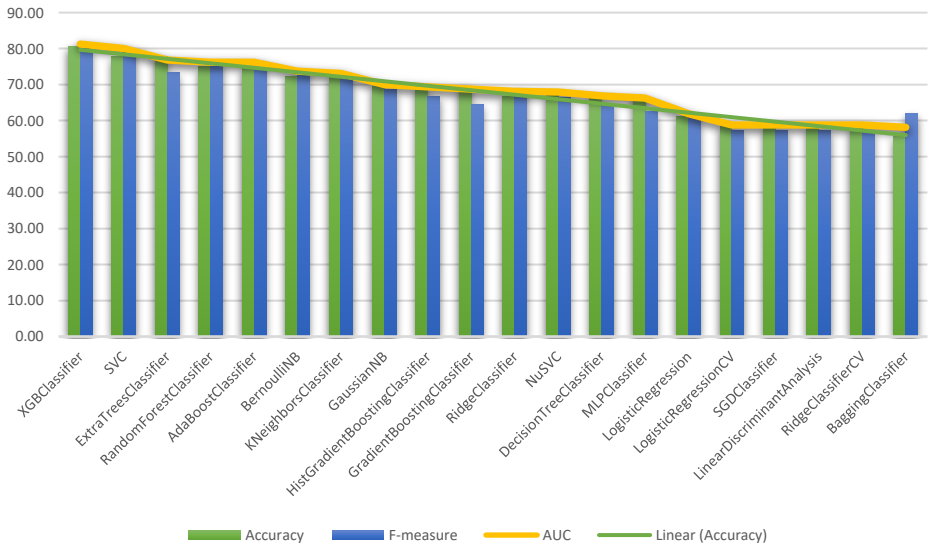

Fig 3: Evaluation Results chart of Machine Learning Models' Performance

Fig 4 presents a comprehensive performance evaluation of the XGBClassifier model, offering an extensive analysis of its performance. The evaluation comprises various measures such as the Classification Report, Confusion Matrix, Class Prediction Error, and AUC curve. These measures demonstrate the model's precision, recall, F-measure, and support for each class, as well as their average values. The AUC curve displays the model's performance across different thresholds, enabling the identification of the optimal threshold.

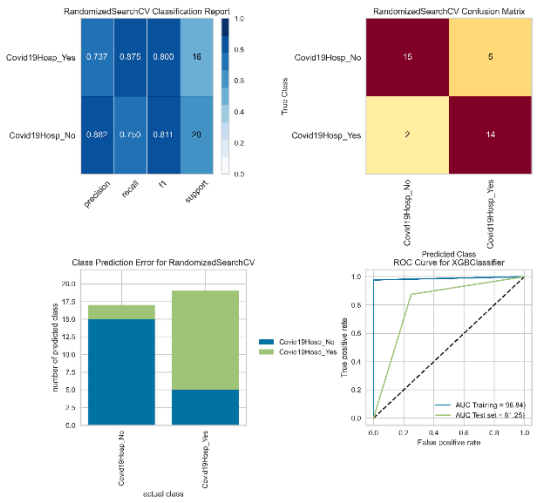

Fig 4: Comprehensive Performance Evaluation of XGBClassifier Model

DISCUSSION

Our study aimed to propose a model for predicting the hospitalization of COVID-19 patients using machine learning classification algorithms. To achieve this, we tested twenty different algorithms and optimized their hyperparameters. We focused on the most important predictors of hospitalization and selected the best-performing algorithms. In the following paragraphs, we discuss the key findings of our study.

Key factors for predicting COVID-19 patient hospitalization

Proper data preprocessing is critical in extracting valuable insights from data. In this study, we followed the latest scientific evidence to systematically preprocess the data. To ensure accuracy and reliability, we executed the Boruta algorithm 15,000 times to identify the most significant predictors that affect the outcome. After extracting the results, we discussed them with a panel of experts who confirmed the most important predictors for COVID-19 hospitalization identified in our study.

In managing the hospitalization of COVID-19 patients, effective allocation of healthcare resources is crucial. Machine learning can help identify the most efficient predictors among a wide range of variables. Our study identified gamma-glutamyl transpeptidase (GGT), alkaline phosphatase (ALP), CT scan diagnosis, mean platelet volume (MPV), mean corpuscular volume (MCV), fasting blood sugar (FBS), red blood cell count (RBC), and mean corpuscular hemoglobin concentration (MCHC) as the top eight predictors for COVID-19 hospitalization. These valuable insights can assist clinicians in allocating resources effectively and improving patient care.

Numerous studies [3, 9, 13-23] have examined risk factors that can predict the likelihood of hospitalization or mortality in COVID-19 patients. Our study confirms the significance of these risk factors as crucial predictors for hospitalization in COVID-19 patients. These factors, including liver factors such as GGT and ALP, fasting blood sugar level, and blood factors related to red blood cells, have been consistently reported as significant predictors in previous research. However, previous studies primarily relied on clinical laboratory data obtained during hospitalization. Our study is unique in that it includes lifestyle variables of patients, which is a significant strength of our approach. By incorporating a broader range of variables, including lifestyle factors, we identified key predictors of hospitalization that directly impact oxygen saturation percentage. Our findings highlight the importance of using a wider range of variables, including lifestyle factors, in predicting the risk of morbidity, mortality, and hospitalization in COVID-19 patients. Nonetheless, our study also indicates that clinical laboratory variables still play a crucial role in predicting patient outcomes. Overall, our study contributes to the development of effective strategies for the allocation of healthcare resources and the enhancement of patient care.

One of the other points to be discussed in this study is the examination of conditions and preventive measures aimed at identifying the disease's direction or progression towards hospitalization based on changes in clinical laboratory variables that may be influenced by environmental conditions. Additionally, investigating these factors can help prevent the occurrence of systemic complications and the progression of systemic disease that may worsen due to the virus's impact on the immune system. Improving immune function can significantly reduce the risk of infection for individuals.

Optimal ML models for predicting COVID-19 patient hospitalization

After identifying the most significant predictors, the algorithm was modeled by optimizing their hyperparameters. The performance of twenty models was evaluated based on criteria such as accuracy, AUC, F-measure, and their averages. The XGBClassifier model outperformed all other models and demonstrated the highest predictive accuracy in identifying COVID-19 patients who were likely to require hospitalization.

The present study employed a higher number of algorithms than previous studies, with a focus on optimizing their hyperparameters across a wider range. This constitutes a significant and valuable contribution of our research. The accurate selection of hyperparameters during the modeling process can enable researchers to apply these algorithms in related research areas. In future studies, incorporating these optimized hyperparameters can

also enhance the confidence and reliability of the models. Therefore, this approach has proved to be highly suitable and can serve as a benchmark for future studies.

Previous research using machine learning techniques on cohort data to predict COVID-19 hospitalization is limited. Reported AUC rates for machine learning models range from 80% to 83% for COVID-19 hospitalization prediction, while AUC rates of 80% to 90% have been reported for other outcomes such as mortality and risk of infection using only hospital data [3, 9, 13-23]. The XGBClassifier algorithm has been identified as one of the most effective models for predicting COVID-19 hospitalization, with an AUC of 83% reported in previous studies [19]. Our study also used the XGBClassifier algorithm and achieved an AUC rate of 81.25%, which is not significantly different from previous studies. However, it is essential to note that our study used data from the Cohort database. Thus, these results should be interpreted with caution.

The study's objectives were effectively achieved through the use of multiple algorithms and precise hyperparameter optimization, resulting in the identification of the most effective model for predicting hospitalization of COVID-19 patients. This methodology can aid researchers in the field of healthcare by leveraging artificial intelligence, particularly machine learning, to achieve goals such as prediction, diagnosis, providing treatment plans, determining drug dosages, and managing diseases. This approach increases the likelihood of acceptance of these models by healthcare teams.

CONCLUSION

In conclusion, the study aimed to propose a model for predicting the hospitalization of COVID-19 patients using machine learning classification algorithms. The study identified key predictors for COVID-19 hospitalization, including gamma-glutamyl transpeptidase (GGT), alkaline phosphatase (ALP), CT scan diagnosis, mean platelet volume (MPV), mean corpuscular volume (MCV), fasting blood sugar (FBS), red blood cell count (RBC), and mean corpuscular hemoglobin concentration (MCHC). The study also found that the XGBClassifier model outperformed all other models and demonstrated the highest predictive AUC in identifying COVID-19 patients who were likely to require hospitalization. The study's findings contribute to the development of effective strategies for the allocation of healthcare resources and the enhancement of patient care. The methodology used in this study can aid researchers in the field of healthcare by leveraging artificial intelligence, particularly machine learning, to achieve various goals.

ACKNOWLEDGMENT

We express our gratitude to our esteemed colleagues at the Dena Cohort Research Center, and we extend our heartfelt appreciation to Mr. Engineer Mohammad Pourebrahim for his invaluable cooperation.

AUTHOR'S CONTRIBUTION

All authors contributed to the literature review, design, data collection and analysis, drafting the manuscript, read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

ETHICS APPROVAL

This study was approved by Iran National Committee for Ethics in Biomedical Research with Approval ID IR.YUMS.REC.1401.112.

REFERENCES

1. Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K, et al. Diagnosis and treatment of adults with community-acquired pneumonia: An official clinical practice guideline of the American thoracic society and infectious diseases society of America. Am J Respir Crit Care Med. 2019; 200(7): e45-67. PMID: 31573350 DOI: 10.1164/rccm.201908-1581ST [PubMed]
2. Combi C, Pozzi G. Clinical information systems and artificial intelligence to support medicine in the COViD-19 pandemic. International Conference on Healthcare Informatics. IEEE; 2021.
3. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health (Amst). 2021; 20: 100178. PMID: 33521226 DOI: 10.1016/j.smhl.2020.100178 [PubMed]
4. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, et al. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials & Continua. 2020; 63(1): 537-51.
5. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395(10223): 497-506. PMID: 31986264 DOI: 10.1016/S0140-6736(20)30183-5 [PubMed]

6. Zhao W, Zhong Z, Xie X, Yu Q, Liu J. Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: A multicenter study. AJR Am J Roentgenol. 2020; 214(5): 1072-7. PMID: 32125873 DOI: 10.2214/AJR.20.22976 [PubMed]

7. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020; 382(18): 1708-20. PMID: 32109013 DOI: 10.1056/NEJMoa2002032 [PubMed]

8. Yang X, Yu Y, Xu J, Shu H, Xia JA, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: A single-centered, retrospective, observational study. Lancet Respir Med. 2020; 8(5): 475-81. PMID: 32105632 DOI: 10.1016/S2213-2600(20)30079-5 [PubMed]

9. 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: m1328. PMID: 32265220 DOI: 10.1136/bmj.m1328 [PubMed]

10. Wang L. C-reactive protein levels in the early stage of COVID-19. Med Mal Infect. 2020; 50(4): 332-4. PMID: 32243911 DOI: 10.1016/j.medmal.2020.03.007 [PubMed]

11. Hutter F, Kotthoff L, Vanschoren J. Automated machine learning: methods, systems, challenges. Springer Nature; 2019.

12. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011; 12: 2825-30.

13. Wungu CDK, Khaerunnisa S, Putri EAC, Hidayati HB, Qurnianingsih E, Lukitasari L, et al. Meta-analysis of cardiac markers for predictive factors on severity and mortality of COVID-19. Int J Infect Dis. 2021; 105: 551-9. PMID: 33711519 DOI: 10.1016/j.ijid.2021.03.008 [PubMed]

14. Cai W, Liu T, Xue X, Luo G, Wang X, Shen Y, et al. CT quantification and machine-learning models for assessment of disease severity and prognosis of COVID-19 patients. Acad Radiol. 2020; 27(12): 1665-78. PMID: 33046370 DOI: 10.1016/j.acra.2020.09.004 [PubMed]

15. Yaşar Ş, Çolak C, Yoloğlu S. Artificial intelligence-based prediction of Covid-19 severity on the results of protein profiling. Comput Methods Programs Biomed. 2021; 202: 105996. PMID: 33631640 DOI: 10.1016/j.cmpb.2021.105996 [PubMed]

16. Banoei MM, Dinparastisaleh R, Zadeh AV, Mirsaeidi M. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. Crit Care. 2021; 25(1): 328. PMID: 34496940 DOI: 10.1186/s13054-021-03749-5 [PubMed]

17. Bayat V, Phelps S, Ryono R, Lee C, Parekh H, Mewton J, et al. A severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) prediction model from standard laboratory tests. Clin Infect Dis. 2021; 73(9): e2901-7. PMID: 32785701 DOI: 10.1093/cid/ciaa1175 [PubMed]

18. Li Y, Hai-Tao Z, Jorge G, Yang X, Maolin W, Yuqi G, et al. A machine learning-based model for survival prediction in patients with severe COVID-19 infection. medRxiv. 2020.

19. Wang K, Zuo P, Liu Y, Zhang M, Zhao X, Xie S, et al. Clinical and laboratory predictors of in-hospital mortality in patients with coronavirus disease-2019: A cohort study in Wuhan, China. Clin Infect Dis. 2020; 71(16): 2079-88. PMID: 32361723 DOI: 10.1093/cid/ciaa538 [PubMed]

20. Hu C, Liu Z, Jiang Y, Shi O, Zhang X, Xu K, et al. Early prediction of mortality risk among patients with severe COVID-19, using machine learning. Int J Epidemiol. 2021; 49(6): 1918-29. PMID: 32997743 DOI: 10.1093/ije/dyaa171 [PubMed]

21. Yao H, Zhang N, Zhang R, Duan M, Xie T, Pan J, et al. Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. Front Cell Dev Biol. 2020; 8: 683. PMID: 32850809 DOI: 10.3389/fcell.2020.00683 [PubMed]

22. Liu Y, Yang Y, Zhang C, Huang F, Wang F, Yuan J, et al. Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury. Sci China Life Sci. 2020; 63(3): 364-74. PMID: 32048163 DOI: 10.1007/s11427-020-1643-8 [PubMed]

23. Petrilli CM, Jones SA, Yang J, Rajagopalan H, O'Donnell L, Chernyak Y, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: Prospective cohort study. BMJ. 2020; 369: m1966. PMID: 32444366 DOI: 10.1136/bmj.m1966 [PubMed]

24. Zhou K, Sun Y, Li L, Zang Z, Wang J, Li J, et al. Eleven routine clinical features predict COVID-19 severity uncovered by machine learning of longitudinal measurements. Comput Struct Biotechnol J. 2021; 19: 3640-9. PMID: 34188785 DOI: 10.1016/j.csbj.2021.06.022 [PubMed]

25. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research. 2017; 18: 1 - 5.

26. Salehnasab C, Hajifathali A, Asadi F, Parkhideh S, Kazemi A, Roshanpoor A, et al. An intelligent clinical decision support system for predicting acute graft-versus-host disease (aGvHD) following allogeneic hematopoietic stem cell transplantation. J Biomed Phys Eng. 2021; 11(3): 345-56. PMID: 34189123 DOI: 10.31661/jbpe.v0i0.2012-1244 [PubMed]

27. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. Journal of Statistical Software. 2010; 36(11): 1 - 13.

28. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In: Sattar A, Kang B (eds). AI 2006: Advances in Artificial Intelligence. Springer Link; 2006.