Informatics in Medicine Unlocked 30 (2022) 100908

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Predicting hospital readmission risk in patients with COVID-19: A machine learning approach

Mohammad Reza Afrash[a], Hadi Kazemi-Arpanahi[b,c], Mostafa Shanbehzadeh[d,*], Raoof Nopour[e], Esmat Mirbagheri[f]

[a] Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[b] Department of Health Information Technology, Abadan Faculty of Medical Sciences, Abadan, Iran
[c] Student Research Committee, Abadan Faculty of Medical Sciences, Abadan, Iran
[d] Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran
[e] Department of Health Information Management, Student Research Committee, School of Health Management and Information Sciences Branch, Iran University of Medical Sciences, Tehran, Iran
[f] Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Keywords:
COVID-19
Coronavirus
Artificial intelligent
Machine learning
Readmission

ABSTRACT

Introduction: The Coronavirus 2019 (COVID-19) epidemic stunned the health systems with severe scarcities in hospital resources. In this critical situation, decreasing COVID-19 readmissions could potentially sustain hospital capacity. This study aimed to select the most affecting features of COVID-19 readmission and compare the capability of Machine Learning (ML) algorithms to predict COVID-19 readmission based on the selected features.
Material and methods: The data of 5791 hospitalized patients with COVID-19 were retrospectively recruited from a hospital registry system. The LASSO feature selection algorithm was used to select the most important features related to COVID-19 readmission. HistGradientBoosting classifier (HGB), Bagging classifier, Multi-Layered Perceptron (MLP), Support Vector Machine ((SVM) kernel = linear), SVM (kernel = RBF), and Extreme Gradient Boosting (XGBoost) classifiers were used for prediction. We evaluated the performance of ML algorithms with a 10-fold cross-validation method using six performance evaluation metrics.
Results: Out of the 42 features, 14 were identified as the most relevant predictors. The XGBoost classifier outperformed the other six ML models with an average accuracy of 91.7%, specificity of 91.3%, the sensitivity of 91.6%, F-measure of 91.8%, and AUC of 0.91%.
Conclusion: The experimental results prove that ML models can satisfactorily predict COVID-19 readmission. Besides considering the risk factors prioritized in this work, categorizing cases with a high risk of reinfection can make the patient triaging procedure and hospital resource utilization more effective.

1. Introduction

Hospital readmission is a well-accepted metric of hospital care quality [1]. It is defined as the new hospitalization in the same hospital within a specified time between 30 and 60 days after initial hospital discharge [2–4]. The high readmission rates are most probably related to the quality of care delivered by hospitals and other health centers during or after the former admission [5,6]. Because of the high costs that readmission imposes on hospitals and patients, it has gained substantial attention as one of the most important criteria for evaluating the quality of care and discharge procedures. Estimates show that 60% of patient readmission can be prevented [7,8].

As the prevalence of the COVID-19, the health care systems of many countries were collapsed and could not meet the growing needs of patients to diagnose, treatment, and care services [9,10]. Many patients in such conditions were discharged after admission with partial recovery

Abbreviations: COVID-19, Coronavirus disease 2019; CDSS, Clinical Decision Support Systems; ML, Machine learning; CRISP, Cross-Industry Standard Process; LASSO, Least Absolute Shrinkage and Selection Operator; AUC, Area under the curve; HGB, Hist Gradient Boosting; MLP, Multi-Layered Perceptron; SVM, Support Vector Machine; XGBoost, Extreme Gradient Boosting.

* Corresponding author. Department of Health Information Technology, Ilam University of Medical Sciences, Ilam, Iran.
E-mail addresses: mmdrza.lp19@gmail.com (M.R. Afrash), hadi.kazemi67@yahoo.com (H. Kazemi-Arpanahi), shanbehzadeh-m@medilam.ac.ir (M. Shanbehzadeh), raoof.n1370@gmail.com (R. Nopour), Mirbagheri.e@iums.ac.ir (E. Mirbagheri).

https://doi.org/10.1016/j.imu.2022.100908
Received 18 January 2022; Received in revised form 18 February 2022; Accepted 6 March 2022
Available online 8 March 2022

[11]. Meanwhile, due to the unknown and aggressive nature of the disease, the readmission rate of patients increased [12]. Readmission imposes additional costs on care organizations and patients. In addition, it will reduce the quality indicators of service delivery; increase the rate of serious complications and deaths during the pandemic [13]. According to the formal reports, about 5% of COVID-19 confirmed patients necessitate hospitalization care services, and the tolls of readmission from this disease report vary from 2 to 10% [14,15].

In this situation, enhancing the capability of the healthcare system against the pandemic requires attention to technological and intelligent-based solutions such as Clinical Decision Support Systems (CDSSs) [16,17]. CDSSs attracted increasing interest because of the growing availability of a large amount of patient-level data [18,19]. CDSSs using available patient data at the time of admission may provide caregivers with valuable information regarding the likelihood risk of COVID-19 readmission [20,21]. Machine learning (ML) algorithms are complex and flexible classification modeling that leverage big datasets to reveal new and practical patterns [18,22]. ML algorithms will reduce uncertainties and ambiguities related to new diseases such as COVID-19 by providing diagnostic and predictive models based on valid and scientific evidence to assess risks, screening, forecasting, and health planning [23,24]. Recently, published works have shown that several ML methods are more accurate than conventional statistics models for predicting clinical outcomes in COVID-19 hospitalized patients. They are such as predicting the Length of Stay (LOS), hospital bed occupancy and turnover, Intensive Care Unit (ICU) admission, and respiratory intubation [25–27].

Due to the high prevalence of the disease in our country and the existence of some limitations and lack of healthcare resources [28], therefore, the purpose of this study is to develop an effective and efficient diagnostic model based on comparing the performance of ML algorithms for COVID-19 readmission prediction. Therefore, the present study seeks to answer two questions. What are the most important predictor variables affecting readmission and worsening of patients after receiving first hospitalization services? And which ML model is more effective for predicting readmission?

2. Material and methods

2.1. Study roadmap and experiment environment

The present study was conducted in the form of a retrospective and single-center study in 2022 to predict readmission in patients with confirmed COVID-19 based on one of the most popular ML methods called the Cross-Industry Standard Process (CRISP). It was carried out through five main steps including, 1- Data understanding, 2- Data pre-processing, 3- Feature selection, 4- Classifier, and 5-Evaluation. Fig. 1 shows the proposed models of study steps and sub-steps based on CRISP. This study used Python programming language to run all experiments on the data mining algorithms to predict readmission in patients with confirmed COVID-19 (see Fig. 2).

2.2. Data set description

The included cases are defined based on 42 variables in three main classes, including patient's demographics (three variables), hospitalization (eight variables), and clinical (31 variables) (see Table 1). After reviewing the demographical, clinical and hospitalization information of the patients with confirmed COVID-19, statically analysis was performed to describe the differences in the patients with confirmed COVID-19 data, were readmitted or not. For this purpose, the differences in demographical and hospitalization information of patient were described based on whether the patients were readmitted or not, and the relationship of each feature with readmission was checked by the Chi-square test.

Of 5791 COVID-19 hospitalized patients, 3071 (53.04%) were male,

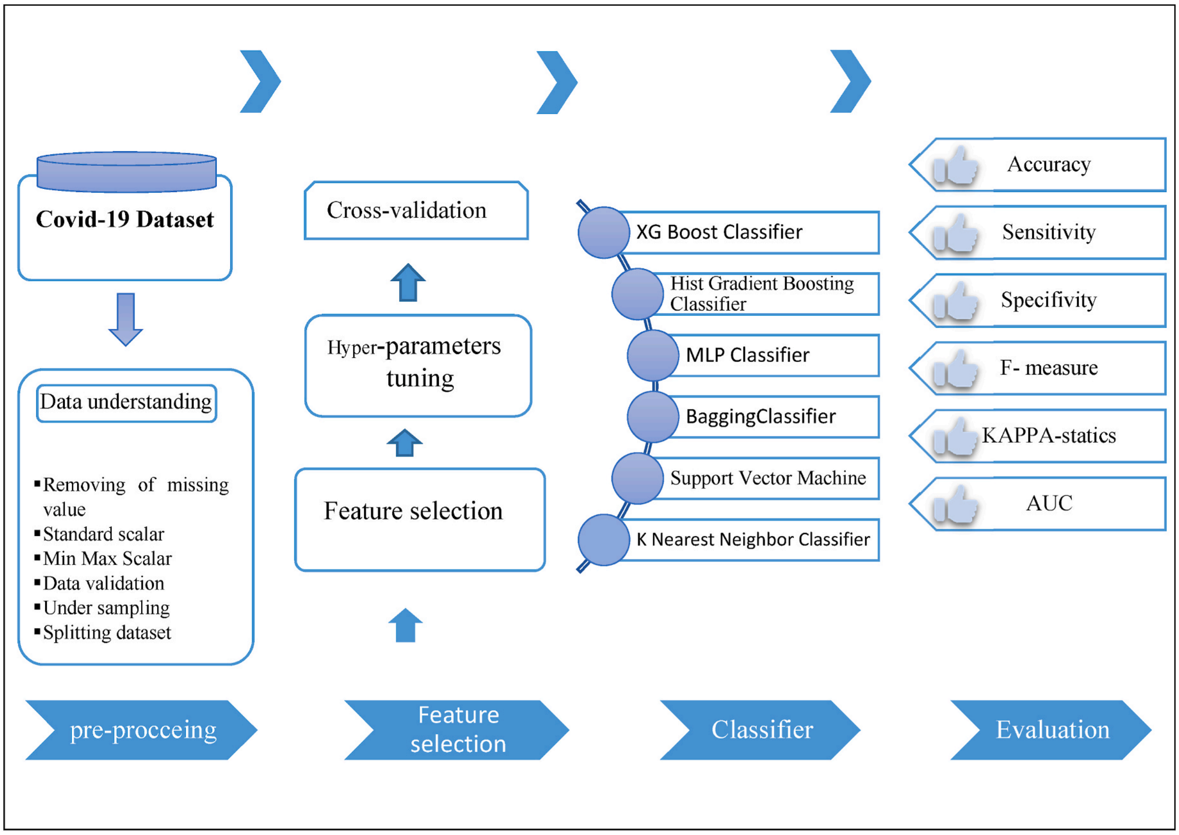

Fig. 1. The roadmap of the proposed system for prediction of readmission based on the CRISP method.

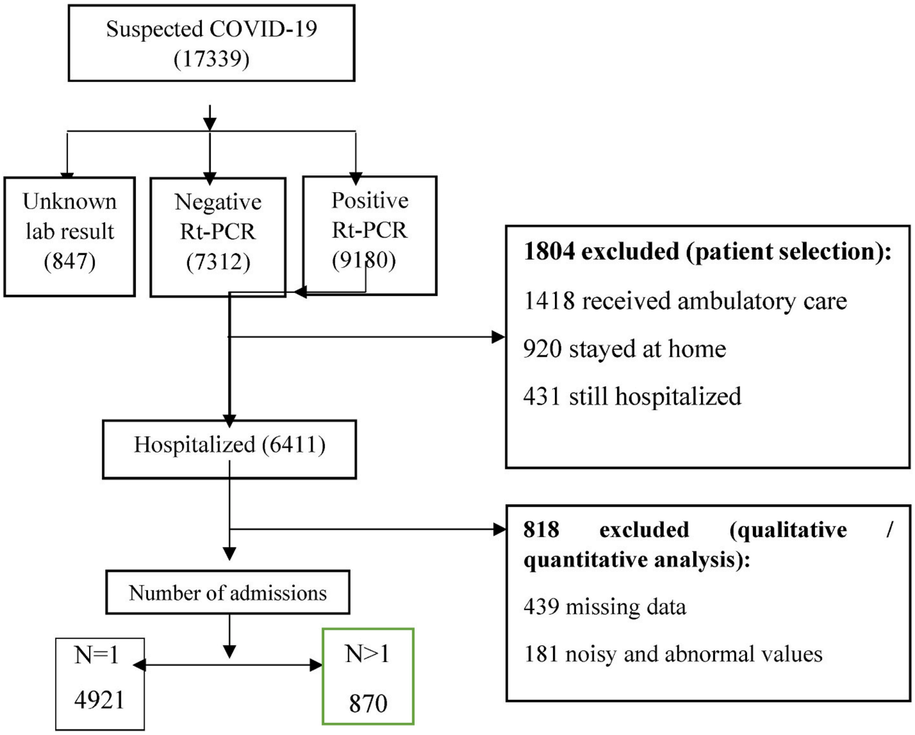

Fig. 2. Flow chart describing patient selection.

2720 (46.96%) were women, and the median age of participants was 57.25 (interquartile 00–100). 528 (13.87%) were hospitalized in ICU, and 2075 (86.13%) were hospitalized in general wards. Out of 5791 included patients, 870 (15.02%) patients were readmitted within 30 days after initial discharge.

2.3. *Ethical consideration*

The ethical committee board approved the study of Ilam University of Medical Sciences (Ethics code: IR.MEDILAM.REC.1399.294). To protect the privacy and confidentiality of patients, we concealed the unique identification information of all patients in the process of data collection and presentation.

2.4. *Preprocessing step*

Preprocessing on the dataset was applied before the training of the proposed model. Several preprocessing steps were examined on the dataset, including removing missing values (rows with missing values greater than 70% were removed.), Standard scalar, Min-Max Scalar, Data validation under sampling for correct use of data in the machine learning algorithms. The noisy and abnormal values, duplicates, and meaningless data impacted ML models' results and were examined and removed by two authors: (M: A and M: SH).

2.5. *Patient selection criteria*

After applying the exclusion criteria, out of 9180 confirmed COVID-19 patients, 6411 hospitalized cases were included in the study. In the preprocessing steps, 818 patient record values were removed, and after deleting these values, the number of patient records was reduced to 5791 cases. Among them, 870 (15.02%) cases were readmitted after a 30-day of the first hospitalization.

2.6. *Feature selection*

Feature selection or variable selection is needed before feeding data into the ML algorithms since outside dimensions affect the classification performance and precision and decrease run time [29]. To select the most important feature to predict readmission, we used Least Absolute Shrinkage and Selection Operator Features Selection Algorithm (LASSO) in this study. The LASSO selects the most important and relevant features for predicting readmission in COVID-19 patients according to updating the absolute value of the variables' coefficient. If the coefficients value of variables is equal to zero, these zero Values for features eliminated that from features subset, and if any variables obtained high values for coefficients. Hence, the feature included in selected variables subsets.

2.6.1. *Machine learning methods*

In this study, to predict the readmission in the patient with confirmed COVID-19, we used seven ML classification algorithms, including Hist Gradient Boosting (HGB) classifier, Bagging classifier, Multi-Layered Perceptron (MLP) classifier, Support Vector Machine ((SVM) kernel = linear), SVM (kernel = RBF), and Extreme Gradient Boosting (XGBoost) classifier.

2.7. *Performance metrics*

To evaluate the performance of applied algorithms and verify the quality of the algorithms in this study, we used the k-fold cross-validation method. Cross-validation is a resampling method used to assess ML models in an unseen data sample. This method has one parameter named k that refers to the number of parts that the dataset should be split. In this study, we use 10 -fold cross validation method. In 10-fold cross-validation methods, the algorithms are trained and tested 10-time times, and then the mean evaluation metrics. Accuracy, specificity, sensitivity, KAPA statistic, Area under the curve (AUC) are measured at the end of the process curve (Equations (1)–(6)).

Table 1
Patient characteristics variable data.

Patient Characteristics	Variables	Total		Readmission	Non-Readmission	P-value
				N	N	
Demographical	Sex	Female	2720	412	2308	<0.002**
		Male	3071	332	2739	
	Marital status	single,	1219	631	588	<0.004**
		married	4572	239	4333	
	Age	0–30	1363	152	1211	<0.001**
		30–60	1836	146	1690	
		60–90	2952	572	2380	
Hospitalization	Number of admissions	1	4921	0	4921	
		2–4	780	780	0	<0.002**
		>4	90	90	0	
	Type of admission	Inpatient care	2075	524	1551	<0.001**
		Outpatient care	3716	346	3370	
	ICU admission	Yes	528	462	66	<0.002**
		No	5263	408	4855	
	Oxygen therapy	Yes	720	543	177	<0.161
		No	5071	327	4744	
	CRP on admission	Yes	380	329	51	<0.039**
		No	5411	541	4870	
	Duration of hospitalization	<24 h	3917	43	3874	<0.497**
		1–7 days	1465	519	946	
		>7days	409	308	101	
	Patient status on discharge	Partial recovery-	1430	774	656	<0.041**
		Complete recovery	3970	62	3908	
		dead	391	34	357	
	Time to readmission	<30 days	1300	257	1043	<0.052
		>30days	4491	613	3878	
	COVID status	Critical	520	14	506	<0.001**
		Severe	1034	142	892	
		Moderate	2300	540	1760	
		Mild	1540	98	1442	
		Recovered	397	14	383	
	Severe kidney disease	Yes	240	49	191	<0.630
		No	5551	821	4730	
	Solid organ transplantation	Yes	182	94	88	<0.951
		No	5609	776	4833	
	Lymphocytes on discharge	Yes	746	297	449	<0.832
		No	5045	573	4472	
	Coronary artery disease	Yes	570	381	189	<0.267
		No	5221	489	4732	
	Cancer	Yes	168	119	49	<0.574
		No	5623	751	4872	
	History of CT result	Normal	3321	540	2781	<0.059
		Unmoral	2470	330	2140	
	Pregnancy	Yes	94	23	71	<0.720
		No	5697	847	4850	
	Congestive heart failure	Yes	350	180	170	<0.968
		No	5441	690	4751	
	Cerebrovascular disease	Yes	49	8	41	<0.602
		No	5742	862	4880	
	C reactive protein on admission	Yes	5308	710	4598	<0.057
		No	753	160	593	
	Congestive heart failure	Yes	135	94	41	<0.619
		No	5656	776	4880	
	Asthma	Yes	74	41	33	<0.570
		No	5717	829	4888	
	Metastatic solid tumor	Yes	14	3	11	<0.924
		No	5776	867	4909	
	Diabetes mellitus	Yes	364	79	285	<0.738
		No	5427	791	4636	
	D-dimer	Yes	4680	361	4319	<0.042**
		No	1111	509	602	
	Dyspnea	Yes	1640	490	1150	<0.069
		No	4151	380	3771	
	Underlying diseases	Yes	839	538	301	<0.073
		No	4952	468	4484	
	Headache	Yes	4981	681	4300	<0.075
		No	810	189	621	
	Weakness and lethargy	Yes	5134	526	4608	<0.052
		No	657	344	313	
	Body pain	Yes	4391	617	3774	<0.061
		No	1400	253	1147	
	Pain or pressure in the chest	Yes	2670	594	2076	<0.068
		No	3121	276	2845	

(continued on next page)

Table 1 (*continued*)

Patient Characteristics	Variables	Total		Readmission	Non-Readmission	P-value
				N	N	
	High fever	Yes	4621	713	3908	<0.072
		No	1170	157	1013	
	Nausea & Vomiting		Yes	3910	672	3238
	No	<0.067	1881	198	1683	
	Cough	Yes	4627	593	4034	<0.0512
		No	1164	277	887	
	Gastrointestinal symptoms	Yes	234	56	178	<0.102
		No	5557	814	4743	
	Chronic pulmonary	Yes	261	73	188	<0.284
		No	5530	797	4733	
	Hypertension	Yes	840	142	698	<0.043**
		No	4951	728	4223	
	Consolidation	Yes	461	59	402	<0.0497**
		No	5330	811	4519	
	Pleural fluid	Yes	571	137	434	<0.0581
		No	5220	733	4487	
	Hypersensitive troponin	Yes	892	261	568	<0.042*
No		4899	609	4290		

$$\text{classification accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{classification sensitivity} = \frac{Tp}{TP + FN} \tag{2}$$

$$\text{c lassification specificity} = \frac{TN}{TN + FP} \tag{3}$$

$$\text{classification error} = \frac{FP + FN}{TP + TN + FP + FN} \tag{4}$$

$$f - \text{measure} = 2\frac{\text{precision}*\text{sensitivity}}{\text{precision} + \text{sensitivity}} \tag{5}$$

3. Results

3.1. Patient characteristics

The mean age of patients who were readmitted to the hospital was 59 ± 9 years old. The mean age of patients who were not readmitted to the hospital was 51 ± 6 years old ($p < 0.002$). Table 1 indicated that there was a significant association between some features of patients who readmitted or not: features with p-value < 0.005 that showed in Table 1 with (** symbol) have a significant difference in patients who readmitted d or not class. For example, the results showed that there was a significant relationship between ICU admission and COVID status with readmission (p-value < 0.002) and (p-value-<0.001), respectively.

3.2. Feature selection

The LASSO feature selection method selects the most important and relevant features for predicting readmission according to updating the absolute value of the variables' coefficient. The LASSO feature selection ranks the relevant variables. After feature selection, out of 42, 28 variables have not been selected to predict readmission and have been deleted from the dataset. The top 14 selected important variables by the LASSO feature selection method and their scores are represented in Table 2.

Based on Table 2, COVID-19 status, ICU admission, and oxygen therapy obtain the highest score for the prediction of readmission in a patient with COVID-19. Moreover, age and solid metastatic tumor have a low score in relevant variables scores, so it means that age and solid metastatic tumor have a low impact on the prediction of readmission in confirmed COVID-19 patients.

Table 2
Important variables selected by the LASSO algorithm.

Order	Feature name	Score	P-Value
1	COVID status	3.78	0/015
2	ICU admission	3.50	0/035
3	Oxygen therapy	3.31	0/012
4	CRP on admission	3.19	0/047
5	Duration of hospitalization	3.08	0/032
6	Solid organ transplantation	2.94	<0/001
7	Lymphocytes on discharge	2.71	0/001
8	Coronary artery disease	2.64	0/023
9	Cerebrovascular disease	2.47	0/027
10	C reactive protein on admission	2.39	0/012
11	Congestive heart failure	2.15	0/017
12	Asthma	2.09	0/021
13	Metastatic solid tumor	2.03	0/006
14	Age	1.74	0/045

3.3. Results of hyper-parameters tuning

The performance of ML algorithms is highly dependent on the selection of their hyper-parameters. Hyper-parameters are applied to ML algorithms to produce the best model on a given dataset. After the preprocessing step, several ML modeling was performed by adjusting and optimizing hyper-parameters. The best hyper-parameters needed to build models with the highest F-criteria score were identified during this step. In the present study, to select the most precise and powerful models, the Randomized Search CV method was used for parameter adjustment and optimization algorithms, including HGB classifier, Bagging classifier, MLP classifier, SVM (kernel = linear), SVM (kernel = RBF), and XGBoost classifier. Table 3 represents the best Hyper-parameters for ML algorithm modeling for predicting readmission.

3.4. K-fold cross-validation

Selected features by the LASSO feature selection method were tested on seven ML algorithms with a 10-fold cross-validation method. 10-fold cross-validation splits our selected data set into ten subsets and performs the holdout method ten times. 90% of data was used for training ML algorithms for each run, and 10% was fed into the algorithms to test models. To measure the performance of ML algorithms with a 95% confidence interval, we measured the mean of evaluation metrics. Table 4 shows the results of seven prediction models on the selected feature by the LASSO method with a 10-fold cross-validation method to predict the readmission in COVID-19 patients.

Table 4 shows the results of the ML models on the adopted features

Table 3
Best hyper-parameters for ML algorithm modeling in prediction of readmission.

Num	Algorithms	Hyper-parameters	f-score
1	HistGradientBoostingClassifier	‘verbose’ = 2, ‘random_state’ = 999, ‘max_leaf_nodes’ = 62, ‘max_iter’ = 150, ‘max_depht’ = 7, ‘learning rate’ = 0.1	93.7
2	BaggingClassifier	‘verbose’ = 2, ‘random_state’ = 999, ‘n_estimation’ = 12, ‘max-samples’ = 0.5, ‘bootstrap’ = ‘true’	91.28
3	MLP Classifier	‘Learning rate’ = ‘constant’, hidden_layer_size’ = (100,100,100), ‘alpha’ = 0.05, ‘activation’ = ‘rulo’	91.07
4	SVM (kernel = linear)	C = 100,G = 0.0001	90.09
5	SVM (kernel = RBF)	C = 10, G = 0.001	89.24
6	XG Boost Classifier	‘min_chid_weigh’ = 1′max_depht’ = 12,‘learning_rate’ = 0.1, ‘gamma’ = 0.4, ‘colsample_bytree’ = 0.3	89.01
7	K Nearest Neighbor Classifier	K = 3, ‘n_jobs’ = −1, ‘algorithm’ = ‘auto’	87.00

by the LASSO feature selection method in ten independent runs. The results show that the HGB classifier gave a mean accuracy of 88.6%, a mean sensitivity of 88.4%, a mean specificity of 88.9.55%, mean F-measure of 88.1%, a mean for Kappa statistic of 88.6%, and AUC of 88.2% when selected risk factors were used. Bagging classifier obtained a mean accuracy of 84.7%, a mean sensitivity of 84.7%, a mean specificity of 84.1%, a mean F-measure of 84.5%, a mean for Kappa statistic of 84.36.6%, and AUC of 84.3% when the LASSO feature selection method was included in the classifier. Based on Table 3, the MLP classifier shows good performance that has a mean accuracy of 88.6%, 88.9% for a mean of specificity, 88.4% for a mean sensitivity of 88.1%, a Mean F-measure, 88.6% a mean of Kappa Statistic, and 88.2% for a mean of AUC metrics. The performance of the XGBoost classifier was excellent, as shown in Table 3. The XGBoost classifier achieved 91.7% for a mean accuracy, 91.3% specificity, 91.6% mean of sensitivity, 91.8% mean F-measure, 91.37% a mean of Kappa Statistic 91.4% for a mean of AUC per ten independent runs.

The SVM (kernel = linear) was the second-best classifier that has a mean of accuracy 88.9%, 87.3% for a mean of specificity, 91.2% for a mean of sensitivity, 89.2% mean F- measure, 88.7% a mean of Kappa Statistic and 89.2% obtained as a mean of AUC. The SVM (kernel = RBF) has a mean accuracy of 85.7%, a mean sensitivity of 86.1%, a mean specificity of 85.0%, Mean F-measure of 85.9%, a mean for Kappa rate of 86.7%, and AUC of 86.3% when LASSO feature selection method was included in the classifier. The KNN classifier with mean classification accuracy 88.3%, specificity 87.8%, sensitivity 89.2%, F- measure 89.37%, Kappa statistic 88.3%, and AUC 88.6% achieved nearly acceptable performance.

As shown in Fig. 3, the performance of the XGBoost classifier outperformed the other six ML models with 91.7% mean accuracy, 91.3% mean specificity, 91.6% mean sensitivity, 91.8% mean F-measure, and 0.9145 AUC. The second important model was SVM with the linear kernel (ACU = 0.892), and the worst performance was observed for the HGB classifier out of six other ML algorithms (AUC = 0.8233). The classification report and ROC curve of the XGBoost classifier as the best classification algorithm in the present study in terms of the highest evaluation metrics are displayed in Fig. 4.

4. Discussion

Given the unknown nature of COVID-19 with a wide range of symptoms and complications, it is important to implement intelligent-based models for estimating the possibility of its reinfection and recurrence [30,31]. Readmission and disease recurrence prediction is complex and challenging, especially in new and ambiguous diseases such as COVID-19 [32,33]. Based on our knowledge, this work is one of the few studies that applied ML algorithms for predicting the readmission risk of patients with COVID-19.

So far, most previous ML-based studies have focused on predicting readmission of chronic conditions such as cardiovascular [1,34–39], stroke [40–44], and COPD [5,6,45–47]. Till now, few studies have been conducted about COVID-19 readmission. In Rodriguez's study (2021), a predictive model for readmission in COVID-19 patients was presented based on an ML classifier. They concluded that ML and data mining-based approaches have seemed fruitful for readmission prediction [20]. Koteswari (2020) proposed an intelligent model to predict the readmission probability of various COVID-19 cases using ML techniques. The experimental results demonstrate ML-based predictive models can reduce COVID-19 readmission [30]. Raftarai (2021) compared the performance of four ML algorithms for predicting readmission in patients with COVID-19. The AdaBoost ensemble classifier yielded the best performance (accuracy 91.61%) [33]. Similarly, Jia (2021) assessed the performance of some ML algorithms to predict future deterioration

Table 4
10-fold CV Classification performance of different classifiers on selected features.

Classifier		Mean Accuracy	Mean Specificity (%)	Mean Sensitivity	Mean F- measure	Kappa Statistic (KS)	AUC
HGB Classifier	Mean	0.8176	0.814	0.8296	0.8201	82.4%	0.8233
	95% CI	(0.81, 0.83)	(0.8, 0.82)	(0.81, 0.85)	(0.81, 0.83)	(0.82, 0.86)	(0.81, 0.83)
	STD	0.0154	0.0127	0.0296	0.0148	0.0257	0.0157
Bagging Classifier	Mean	0.847	0.841	0.847	0.845	84.36%	0.843
	95% CI	(0.84, 0.85)	(0.84, 0.85)	(0.84, 0.85)	(0.85, 0.85)	(0.84, 0.85)	(0.84, 0.85)
	STD	0.0172	0.0116	0.00128	0.0194	0.0127	0.0182
MLP Classifier	Mean	0.886	0.889	0.884	0.881	88.6%	0.882
	95% CI	(0.88, 0.89)	(0.88, 0.89)	(0.88, 0.89)	(0.88, 0.89)	(0.88, 0.89)	(0.88, 0.89)
	STD	0.0027	0.0112	0.0134	0.00140	0.010	0.0129
XGBoost Classifier	Mean	0.917	0.913	0.916	0.918	91.37%	0.9145
	95% CI	(0.91, 0.92)	(0.91, 0.92)	(0.91, 0.92)	(0.91, 0.92)	(0.91, 0.92)	(0.91, 0.92)
	STD	0.0146	0.0138	0.0147	0.0175	0.01924	0.0126
SVM (kernel = linear)	Mean	0.8896	0.8733	0.912	0.892	88.7%	0.892
	95% CI	(0.87, 0.90)	(0.66, 0.88)	(0.90, 0.93)	(0.88, 0.90)	(0.88, 0.89)	(0.88, 0.90)
	STD	0.0174	0.0167	0.0129	0.0182	0.0140	0.01864
SVM (kernel = RBF)	Mean	0.857	0.850	0.861	0.859	86.7%	0.863
	95% CI	(0.85, 0.86)	(0.84, 0.86)	(0.85, 0.87)	(0.85, 0.87)	(0.86, 0.87)	(0.86, 0.87)
	STD	0.0127	0.01734	0.0129	0.0134	0.0118	0.01727
K Nearest Neighbor Classifier	Mean	0.8835	0.8785	0.892	0.8937	88.3%	0.886
	95% CI	(0.88, 0.89)	(0.87, 0.89)	(0.89, 0.90)	(0.89, 0.90)	(0.88, 0.89)	(0.88, 0.89)
	STD	0.0014	0.0174	0.018	0.0162	0.0183	0.0163

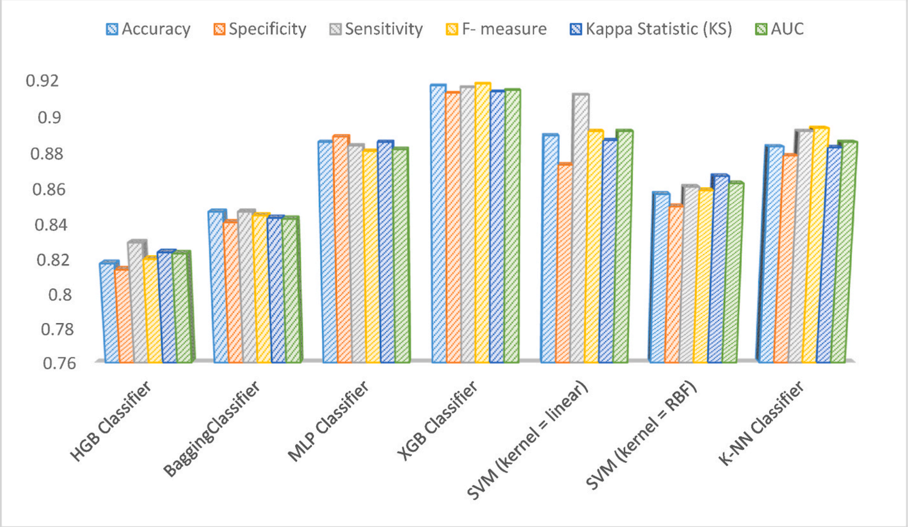

Fig. 3. Comparison of classification models performance on selected features.

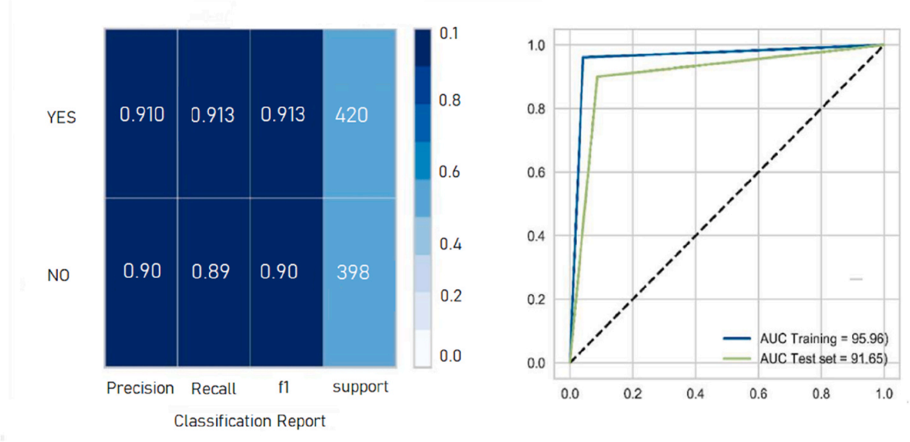

Fig. 4. Classification report and AUC curve of the XGBoost classifier.

among discharged patients with COVID-19. Finally, the best performance was yielded by XGBoost with a mean accuracy of 91.7%, mean specificity of 91.3%, mean sensitivity of 91.6%, mean F-measure of 91.8%, and AUC of 91.45%. Ryu (2021) [48] showed Gradient Boosting Machine (GBM) and Lo (2021) [49] concluded Categorical boosting (Catboost) had the highest AUC performance (= %75.1 and %75.15 respectively) in prediction readmission. Besides in recent studies (performed in 2021) by Zhao [50], Darabi [51], Chen [52], Shah [53], the results showed Boosting algorithms gained better performance in predicting patient readmission.

Boosting like Adaptive Boosting (Ada Boost), XGBoost, HGB, Catboost and GBM is a set of powerful and most widely used ML algorithms. Boosting classifiers improve the classification accuracy by combining of the outputs from a sequence of weak learner and developing a robust predictive model [54,55]. The results of previous studies showed that the performance of these algorithms was optimum in predicting hospital readmission risk in patients with COVID-19. In the present study, due to the optimization of prediction variables through performing feature selection and data preprocessing before using them as inputs for modeling, the performance of the implemented models has been improved. Similarly in the current work the XGBoost model outperformed the other six techniques (0.91% AUC, 0.91–0.92 CI and 0.0146 STD).

Since the COVID-19 pandemic began, several studies selected clinically important predictors for post-discharge COVID-19. For example, Rodriguez's study (2021) indicated underline chronic disease, hypoxia (oxygen saturation $\leq$94%), increased LDH, CRP, and ESR as the most effective factors on hospital readmission [20]. In another study performed by Mendito (2021), several clinical features such as age, neutrophilia count, sequential organ failure assessment (SOFA), LDH, CRP, and D-dimer are recognized as highly contributing factors to the readmission of COVID-19 patients [31]. But, Duarte's research (2021) detected polypharmacy, living in residential care or nursing homes, general illness, chest pain, psychological symptoms, syncope, and superinfection as the most relevant factors on COVID-19 hospital readmission [56]. Accordingly, in Nematshahi et al.'s (2021) study, the period between discharge to readmission, age, gender, underline disease, creatinine level, and pulmonary involvement were renowned as influencing factors in predicting COVID-19 readmission [57].

Similarly, in Jeon's (2020) research, age and sex variables and the presence of underlying disease are effective in increasing the risk of readmission of COVID-19 patients [58]. The presence of comorbidities, high BMI, adult age, laboratory indicators such as CRP, creatinine, and ALT/ASP rate was introduced as one of the most important underlying factors for readmission in COVID-19 patients in the Verna study [59]. In a systematic review study conducted by Akbari et al. (2021), they concluded that male sex, white ethnicity, comorbid diseases, and old age are affecting variables on COVID-19 readmission [60]. Fukushima's study (2021) also showed that certain comorbidities such as diabetes, hypertension, and cardiovascular diseases have a higher capability in predicting the readmission risk among COVID-19 patients [61]. Age over 60 years, underlying diseases, especially diabetes, high creatinine level, and lung involvement were the essential predictors of readmission in the patients with COVID-19 (et al. [32]). The most important variables in the Green (2021) study for readmission prediction were age, LOS, ICU admission, oxygen saturation, D-dimer, and cardiovascular diseases [62].

Similarly, we identified 14 highly correlated variables with the output class. Major risk factors for readmission in the current study include COVID-19 status, ICU admission, Oxygen therapy, CRP on admission, duration of hospitalization, Solid-organ transplantation, Lymphocytes on discharge, Coronary artery disease, Cerebrovascular disease, CRP on admission, congestive heart failure, asthma, metastatic solid tumor, and age most of which are non-modifiable.

It should be noted that the identified variables in the present study are consistent with the previous researches. In the reviewed studies, baseline variables (e.g. age and sex), laboratory indicators, underlying diseases (comorbidities) and resource utilization variables such as LOS, ICU admission, and oxygen therapy play a pivotal role in predicting the readmission of patients with COVID-19. However in these studies, the importance of radiological data for readmission risk prediction among COVID-19 patients, has been neglected. Similarity, in the present study, after doing feature selection, the selected data set lacks radiological variables. Therefore, more studies are needed in this regard.

In addition, several models for predicting the risk of readmission among COVID-19 patients have been developed, one of which gained reasonable performance in the evaluation phase. Interestingly, the selected ML algorithm (XGBoost) can predict the 30-day readmission risk of patients with high accuracy. The proposed model of the present study can help healthcare providers timely detect patient deterioration and reduce the severe complications and the resulting mortalities. This study is a retrospective-single-center study including a relatively small number of patient data. Therefore, the findings may not be generalizable to the wider population. In addition, the existence of some noisy data fields such as inconsistency, meaningless, missing, error-prone, and abnormal fields might impact the data mining accuracy.

Moreover, we used only eight ML algorithms for prediction analyses based on some clinical features. Our data set furthermore lacked clinically essential variables such as imaging indicators. Therefore, at first, to remove noisy data, the normal range of each variable is defined using the opinion of two infectious diseases specialists. Then, we specified all the values outside the defined range and completed them by referring them to the responsible doctor. In addition, the records with more than 70% of empty fields (=439 as shown in Fig. 1) were removed. The missing fields in the records with less than 70% missing are imputed by mean and mode values substitution for continuous and discrete variables, respectively. Additional external validation methods should be used to prove the results of the present study and further verify the generalizability of our results. Finally, the selected dataset lacks some clinical variables such as radiological indicators. As practical solutions, the accuracy and generalizability of our models will be enhanced if we test more ML techniques at the larger, multicenter, and prospective datasets.

5. Conclusion

We implement and validate several predictive models stratifying readmission risk for COVID-19 patients. In particular, it has been observed that the XGBoost model performed best on classification accuracy better than the other ML algorithms. This method can provide caregivers and hospital administrators with an effective instrument to allocate limited hospital resources best. These models also may be an advantage in better and customized care delivery, lessen clinician workload, and diminish severe complication and death in the COVID-19 patients. In future work, the proposed method is expected to be applied to other hospital resource utilization domains such as ICU bed turnover, LOS, and respiratory ventilator.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This article was extracted from the project with the Ethics code: IR. MEDILAM.REC.1399.294. Also, we thank from research committee of Ilam University of Medical Sciences for financially supporting this project.

References

[1] Mahajan SM, Ghani R. Using ensemble machine learning methods for predicting risk of readmission for heart failure. Stud Health Technol Inf 2019:243–7.

[2] Baillie CA, VanZandbergen C, Tait G, Hanish A, Leas B, French B, et al. The readmission risk flag: using the electronic health record to automatically identify patients at risk for 30-day readmission. J Hosp Med 2013;8(12):689–95.

[3] Jamei M, Nisnevich A, Wetchler E, Sudat S, Liu E. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PLoS One 2017;12(7): e0181173.

[4] Tavares MG, Tedesco-Silva H, Pestana JOM. Early Hospital Readmission (EHR) in kidney transplantation: a review article. Braz J Nephrol 2020;42:231–7.

[5] Goto T, Jo T, Matsui H, Fushimi K, Hayashi H, Yasunaga H. Machine learning-based prediction models for 30-day readmission after hospitalization for chronic obstructive pulmonary disease. COPD 2019;16(5–6):338–43.

[6] Hemmrich M, Kaskovich S, Venable L, Carey K, Churpek M, Press V. Accuracy comparison of a machine learning readmission prediction model with HOSPITAL and PEARL scores for chronic obstructive pulmonary disease (COPD) inpatients. D102 OPTIMIZING OUTCOMES IN COPD: American Thoracic Society; 2019. p. A7118 [A].

[7] Wallmann R, Llorca J, Gómez-Acebo I, Ortega ÁC, Roldan FR, Dierssen-Sotos T. Prediction of 30-day cardiac-related-emergency-readmissions using simple administrative hospital data. Int J Cardiol 2013;164(2):193–200.

[8] Dharmarajan K, Hsieh AF, Lin Z, Bueno H, Ross JS, Horwitz LI, et al. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. JAMA 2013;309(4):355–63.

[9] Navik U, Bhatti J, Sheth V, Jawalekar S, Bhatti G, Kalra S. Multi-organ failure in COVID-19 patients: a possible mechanistic approach. 1st12. Netherland: Chandigarh University; 2020. p. 16–8.

[10] Hu Y, Deng H, Huang L, Xia L, Zhou X. Analysis of characteristics in death patients with COVID-19 pneumonia without underlying diseases. Acad Radiol 2020;27(5): 752.

[11] Donnelly JP, Wang XQ, Iwashyna TJ, Prescott HC. Readmission and death after initial hospital discharge among patients with COVID-19 in a large multihospital system. JAMA 2021;325(3):304–6.

[12] Chaudhry Z, Shawe-Taylor M, Rampling T, Cutfield T, Bidwell G, Chan XHS, et al. Short durations of corticosteroids for hospitalised COVID-19 patients are associated with a high readmission rate. J Infect 2021;82(6):276–316.

[13] C.0+ VV, 35, +, +3\63Ulversoy KA, Murrow JR. Impact of the COVID-19 pandemic on cost and readmission rates of heart failure patients at Piedmont Athens regional.

[14] Naghavi S, Kavosh A, Adibi I, Shaygannejad V, Arabi S, Rahimi M, et al. COVID-19 infection and hospitalization rate in Iranian multiple sclerosis patients: what we know by May 2021. Multiple Sclerosis Relat Disorders 2021:103335.

[15] Szente Fonseca SN, de Queiroz Sousa A, Wolkoff AG, Moreira MS, Pinto BC, Valente Takeda CF, et al. Risk of hospitalization for Covid-19 outpatients treated with various drug regimens in Brazil: comparative analysis. Trav Med Infect Dis 2020;38:101906.

[16] Shanbehzadeh M, Nopour R. Determination of the most important diagnostic criteria for COVID-19: a step forward to design an intelligent clinical decision support system. J Adv Med Biomed Res 2021;29(134):176–82.

[17] Moulaei K, Shanbehzadeh M, Mohammadi-Taghiabad Z, Kazemi-Arpanahi H. Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med Inf Decis Making 2022;22(1):1–12.

[18] Rojas JC, Carey KA, Edelson DP, Venable LR, Howell MD, Churpek MM. Predicting intensive care unit readmission with machine learning using electronic health record data. Ann Am Thorac Soc 2018;15(7):846–53.

[19] Nopour R, Shanbehzadeh M, Kazemi-Arpanahi H. Developing a clinical decision support system based on the fuzzy logic and decision tree to predict colorectal cancer. Med J Islam Repub Iran 2021;35(1):341–8.

[20] Rodriguez VA, Bhave S, Chen R, Pang C, Hripcsak G, Sengupta S, et al. Development and validation of prediction models for mechanical ventilation, renal

replacement therapy, and readmission in COVID-19 patients. J Am Med Inf Assoc 2021;28(7):1480–8.

[21] Huang CD, Goo J, Behara RS, Agarwal A. Clinical decision support system for managing copd-related readmission risk. Inf Syst Front 2020;22(3):735–47.

[22] Kalagara S, Eltorai AE, Durand WM, DePasse JM, Daniels AH. Machine learning modeling for predicting hospital readmission following lumbar laminectomy. J Neurosurg Spine 2018;30(3):344–52.

[23] Chaurasia V, Pal S. Application of machine learning time series analysis for prediction COVID-19 pandemic. Res Biomed Eng 2020:1–13.

[24] Ghafouri-Fard S, Mohammad-Rahimi H, Motie P, Minabi MA, Taheri M, Nateghinia S. Application of machine learning in the prediction of COVID-19 daily new cases: a scoping review. Heliyon 2021;7(10):e08143.

[25] Shanbehzadeh M, Orooji A, Kazemi-Arpanahi H. Comparing of data mining techniques for predicting in-hospital mortality among patients with covid-19. J Biostat Epidemiol 2021;7(2):154–73.

[26] Machine learning to predict ICU admission, ICU mortality and survivors' Length of Stay among COVID-19 patients: toward optimal allocation of ICU resources. In: Dan T, Li Y, Zhu Z, Chen X, Quan W, Hu Y, et al., editors. IEEE International Conference on bioinformatics and biomedicine (BIBM); 2020. IEEE; 2020.

[27] Lorenzen SS, Nielsen M, Jimenez-Solem E, Petersen TS, Perner A, Thorsen-Meyer H-C, et al. Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark. Sci Rep 2021;11(1):1–10.

[28] SoleimanvandiAzar N, Irandoost SF, Ahmadi S, Xosravi T, Ranjbar H, Mansourian M, et al. Explaining the reasons for not maintaining the health guidelines to prevent COVID-19 in high-risk jobs: a qualitative study in Iran. BMC Publ Health 2021;21(1):1–15.

[29] Ibrahim S, Nazir S, Velastin SA. Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. J Imag 2021;7(11).

[30] Koteswari MJL, Balaji M, Sainadh K, Kavya KCS, Ch K. Reducing Covid-19 readmissions using machine learning. Turk J Physiother Rehabil 2021;32:2.

[31] Menditto VG, Fulgenzi F, Bonifazi M, Gnudi U, Gennarini S, Mei F, et al. Predictors of readmission requiring hospitalization after discharge from emergency departments in patients with COVID-19. Am J Emerg Med 2021;46:146–9.

[32] Drewett GP, Chan RK, Jones N, Wimaleswaran H, Howard ME, McDonald CF, et al. Risk factors for readmission following inpatient management of COVID-19 in a low-prevalence setting. Intern Med J 2021;51(5):821–3.

[33] Raftarai A, Mahounaki RR, Harouni M, Karimi M, Olghoran SK. Predictive models of hospital readmission rate using the improved AdaBoost in COVID-19. Intelligent computing applications for COVID-19. CRC Press; 2021. p. 67–86.

[34] Predicting 30-day readmission in heart failure using machine learning techniques. In: Kerexeta J, Artetxe A, Escolar V, Lozano A, Larburu N, editors. HEALTHINF 2018 - 11th International Conference on health Informatics, Proceedings; part of 11th International Joint Conference on biomedical Engineering systems and Technologies, BIOSTEC 2018; 2018.

[35] Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC Heart Fail 2019;6(2):428–35.

[36] Najafi-Vosough R, Faradmal J, Hosseini SK, Moghimbeigi A, Mahjub H. Predicting hospital readmission in heart failure patients in Iran: a comparison of various machine learning methods. Healthcare Informat Res 2021;27(4):307–14.

[37] Sampedro-Gómez J, Higuero-Saavedra A, Lorenzo-Martín ÁL, Ramírez-Hernández P, Valenzuela-Serrano M, Sánchez PL. Prediction of in-hospital mortality and 30-day readmission in heart failure using machine learning. REC (Rev Esp Cardiol): CardioClinics 2021;53(4):26–33.

[38] Sarijaloo F, Park J, Zhong X, Wokhlu A. Predicting 90 day acute heart failure readmission and death using machine learning-supported decision analysis. Clin Cardiol 2021;44(2):230–7.

[39] Shin S, Austin PC, Ross HJ, Abdel-Qadir H, Freitas C, Tomlinson G, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. ESC Heart Fail 2021;8(1):106–15.

[40] Hung LC, Sung SF, Hu YH. A machine learning approach to predicting readmission or mortality in patients hospitalized for stroke or transient ischemic attack. Appl Sci 2020;10(18).

[41] Darabi N, Hosseinichimeh N, Noto A, Zand R, Abedi V. Machine learning-enabled 30-day readmission model for stroke patients. Front Neurol 2021;12.

[42] Lineback CM, Garg R, Oh E, Naidech AM, Holl JL, Prabhakaran S. Prediction of 30-day readmission after stroke using machine learning and natural language processing. Front Neurol 2021;12.

[43] Chen Y-C, Chung J-H, Yeh Y-J, Lin H-F, Lin C-H, Hsien H-H, et al. Machine learning algorithms to predict 30-day readmission in patients with stroke: a prospective cohort study. 2020.

[44] Kommina L, Theerthagiri P, Payyavula Y, Vemula PS, Reddy GD. Post-stroke readmission prediction model using machine learning algorithms. Emerging trends in data driven computing and communications. Springer; 2021. p. 53–65.

[45] Min X, Yu B, Wang F. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. Sci Rep 2019;9(1).

[46] Verma VK, Lin WY. A Machine Learning-Based Predictive Model for 30-Day Hospital Readmission Prediction for COPD Patients. In: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics; 2020.

[47] Kaskovich S, Hemmrich M, Venable L, Carey K, Churpek M, Press V. Matching patients with chronic obstructive pulmonary disease (COPD) to personalized care: a novel machine learning tool to predict cause of 90-day readmission. D102 optimizing outcomes IN COPD. American Thoracic Society; 2019. p. A7119 [A].

[48] Ryu B, Yoo S, Kim S, Choi J. Thirty-day hospital readmission prediction model based on common data model with weather and air quality data. Sci Rep 2021;11(1):1–9.

[49] Lo Y-T, Liao JC-h, Chen M-H, Chang C-M, Li C-T. Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms. BMC Med Inf Decis Making 2021;21(1):1–11.

[50] Zhao P, Yoo I, Naqvi SH. Early prediction of unplanned 30-day hospital readmission: model development and retrospective data analysis. JMIR Med Informat 2021;9(3):e16306.

[51] Darabi N, Hosseinichimeh N, Noto A, Zand R, Abedi V. Machine learning-enabled 30-day readmission model for stroke patients. Front Neurol 2021;12:425.

[52] Chen L, Chen S. Prediction of readmission in patients with acute exacerbation of chronic obstructive pulmonary disease within one year after treatment and discharge. BMC Pulm Med 2021;21(1):1–17.

[53] Shah AA, Devana SK, Lee C, Bugarin A, Lord EL, Shamie AN, et al. Prediction of major complications and readmission after lumbar spinal fusion: a machine learning–driven approach. World Neurosurg 2021;152:e227–34.

[54] Industrial transfer learning: boosting machine learning in production. In: Tercan H, Guajardo A, Meisen T, editors. IEEE 17th International Conference on Industrial Informatics (INDIN); 2019. IEEE; 2019.

[55] Pedrazzani R, Pintus A, De Ventura R, Marchini M, Ceroni P, Silva López C, et al. Boosting gold (I) catalysis via weak interactions: new fine-tunable impy ligands. ACS Organ Inorgan Au 2022;20:35–42.

[56] Romero-Duarte Á, Rivera-Izquierdo M, Láinez-Ramos-Bossini AJ, Redruello-Guerrero P, Cárdenas-Cruz A. Factors associated with readmission to the Emergency Department in a cohort of COVID-19 hospitalized patients. 2021.

[57] Nematshahi M, Soroosh D, Neamatshahi M, Attarian F, Rahimi F. Factors predicting readmission in patients with COVID-19. BMC Res Notes 2021;14(1):1–6.

[58] Jeon W-H, Seon JY, Park S-Y, Oh I-H. Analysis of risk factors on readmission cases of COVID-19 in the Republic of Korea: using nationwide health claims data. Int J Environ Res Publ Health 2020;17(16):5844.

[59] Verna EC, Landis C, Brown Jr RS, Mospan AR, Crawford JM, Hildebrand JS, et al. Factors associated with readmission in the US following hospitalization with COVID-19. Clin Infect Dis: Off Publ Infect Dis Soc Am 2021;5(20):2021.

[60] Akbari A, Fathabadi A, Razmi M, Zarifian A, Amiri M, Ghodsi A, et al. Characteristics, risk factors, and outcomes associated with readmission in COVID-19 patients: a systematic review and meta-analysis. Am J Emerg Med 2021;52:166–73.

[61] 36. Clinical features of and risk factors for 30-day readmission after an initial hospitalization with COVID-19. In: Fukushima EA, Santos CV, Sharma M, Szpunar SM, Saravolatz L, Bhargava A, editors. Open forum infectious diseases. Oxford University Press US; 2021.

[62] Green H, Yahav D, Eliakim-Raz N, Karny-Epstein N, Kushnir S, Shochat T, et al. Risk-factors for re-admission and outcome of patients hospitalized with confirmed COVID-19. Sci Rep 2021;11(1):17416.