



Original Article

Predictive Modeling of Hospital Length of Stay in COVID-19 Patients Using Machine Learning Algorithms

Mohammad Reza Afrash¹, Hadi Kazemi-Arpanahi^{2,3}, Parvaneh Ranjbar⁴, Raof Nopour⁵, Morteza Amraei⁶, Mojgan Saki⁷, Mostafa Shanbehzadeh^{8,*}

¹Student Research Committee, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

²Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran

³Student Research Committee, Abadan University of Medical Sciences, Abadan, Iran

⁴Department of Scientometrics, Ilam University of Medical Sciences, Ilam, Iran

⁵Department of Health Information Technology and Management, School of Paramedical, Tehran University of Medical Sciences, Tehran, Iran

⁶Department of Health Information Technology, School of Allied Medical Sciences, Lorestan University of Medical Sciences, Khorramabad, Iran

⁷Department of Operating Room, Faculty of Allied Medical Sciences, Lorestan University of Medical Sciences, Khorramabad, Iran

⁸Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran

ARTICLE INFO

Article history

Received: 2021-07-04

Received in revised: 2021-08-07

Accepted: 2021-08-28

Manuscript ID: JMCS-2107-1209

Checked for Plagiarism: Yes

Language Editor:

Dr. Behrouz Jamalvandi

Editor who approved publication: Dr.

Asghar Mesbahi

DOI:10.26655/JMCHMSCI.2021.5.15

KEYWORDS

COVID-19

Coronavirus

Machine learning

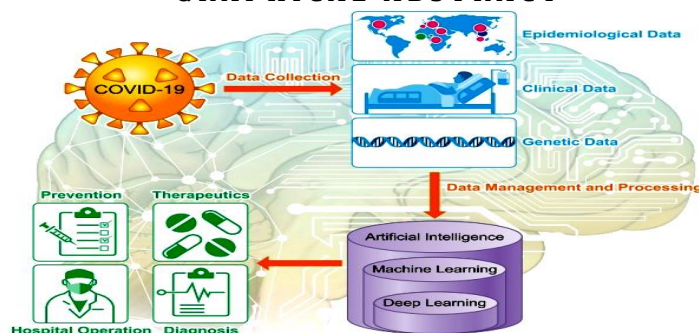
Data mining

Length of Stay

ABSTRACT

The rapid worldwide outbreak of COVID-19 has posed serious and unprecedented challenges to healthcare systems in predicting disease behavior, consequences and resource utilization. Therefore, predicting the Length of Stay (LOS) is necessary to ensure optimal allocate of scarce hospital resources. The purpose of this research was to construct a model for predicting COVID-19 patients' hospital LOS by multiple Machine Learning (ML) algorithms. Using a single-center registry, we studied the records of 1225 laboratory-confirmed COVID-19 hospitalized patients from February 9, 2020, to December 20, 2020. The most important clinical parameters in the COVID-19 LOS prediction were identified with a correlation coefficient at the P-value < 0.2. Then, the prediction models were developed based on seven ML techniques according to selected variables. Finally, to evaluate the performances of those models several standard quantitative measures includes accuracy, sensitivity, specificity and ROC curve were used to evaluate the proposed predictive models. After implementing feature selection, a total of 20 variables was identified as the most relevant predictors to build the prediction models. The results indicated that the best performance belonged to the Support Vector Machine (SVM) algorithm with the mean accuracy of 99.5%, mean specificity of 99.7%, mean sensitivity of 99.4%, and the standard deviation of 1.2. The SVM provided a reasonable level of accuracy and certainty in predicting the LOS in COVID-19 patients and potentially facilitates hospital bed management, turnover and optimized resource allocation.

GRAPHICAL ABSTRACT



* Corresponding author: Mostafa Shanbehzadeh

✉ E-mail: Mostafa.shanbehzadeh@gmail.com

© 2021 by SPC (Sami Publishing Company)

Introduction

The 2019 coronavirus disease (COVID-19) is a very contagious viral infection disease and thus far remains to be spread aggressively and has become a serious global health crisis [1-3]. So far, this disease has affected almost all countries with more than 2.5 million deaths worldwide [4]. The clinical manifestations of virus ranged from asymptomatic or mild flu like symptoms to severe complications including respiratory insufficiency and intensive care unit (ICU) hospitalization, where patients may be intubated for mechanical ventilation and ultimately death [5, 6].

The COVID-19 high transmission rates, unknown clinical patterns, lack of approved drug therapy or vaccines coupled with a long incubation period put a lot of pressure on healthcare organizations by increasing demands for medical services and the surge in hospitalization volumes [7, 8]. In response to this pandemic, the overwhelmed hospitals around the world, have endeavored to curb the outbreak by leverage predictive models for achieve proper decision-makings. COVID-19 has exposed the health systems to serious scarcity of hospital resources, i.e. beds, oxygen generator, personnel and etc. and overtiredness of care workers, which demands in advance accurate prediction models to efficiently triage patients and make best use of limited resources [9, 10].

It highlights the need for objective and evidence-based solutions for the effective use of medical resources available in hospitals, e.g. hospital beds, personnel, respiratory ventilators, etc. to prevent hospital overwhelming and optimal use of medical resources [11-13]. To reduce the pressure on hospitals and provide the best care for patients, especially in overwhelmed hospitals, it is necessary to effectively predict the Length of Stay (LOS). Therefore, an exact approximation of the patients' LOS would be of substantial worth for scientifically dealing with both medical resources and the distribution of caregivers [7, 14, 15]. Decreasing the LOS is significantly effective in managing of patient flows, enhancing resource utilization, improving patients' safety and reducing healthcare costs [16-20].

For this aim, early and accurate estimation of the hospital LOS metrics, would allow for optimal management of limited medical resources, hospital staffing, better patient scheduling and effective bed turnover process. In addition, health care organizations could help the design of well-organized clinical pathways and recognize the bottlenecks to improve resource utilization, resource allocation proactively, and better healthcare supply chain management [16, 17]. It is significant for hospital administrator, clinicians and health policymakers to make proper decisions for allocating of restricted medical resources particularly during current COVID-19 pandemic [17, 21]. Given this context, the ability to identify patients who are at risk for prolonged LOS during their hospitalization episode may be useful for identifying and prioritizing patient requirements, care planning, and optimizing service delivery [22]. Many healthcare systems across the world struggle to predict the prolonged LOS by leverage Machine Learning (ML) models for achieving proper decision-makings [23-27]. ML as a sub-branch of Artificial Intelligence (AI) has been extensively known as an efficient and promising analytical techniques for achieve proper decision-makings in healthcare by automatically extracting practical patterns from structured big dataset [28, 29]. These techniques are well-known tools for developing predictive models and can implicitly extract useful information from raw datasets [30]. In the previous studies, a large number of ML algorithms were trained for the forecast and classification of hospital LOS especially for cardiovascular [11, 31-33], malignancies [34-36], and orthopedics [37-39] conditions.

Towards this aim, we developed and compared several ML algorithms to predict COVID-19 LOS according to routine clinical data at admission time. More precisely, the study questions posed for the experiment were: Which prediction model presents better performance? Which prediction models are more efficient? And which models exploit a higher accuracy?

Material and methods

This is a retrospective, single-center, and cross-sectional study, that was conducted in 2021 for

predicting the COVID-19 patients LOS based on selected data-driven ML techniques. It was conducted in five stages, including, 1- data set description and participants, 2- data preprocessing, 3- feature reduction, 4- model development, 5- experiment evaluation and 6- ethical consideration.

Data set description and participants

In this study, a COVID-19 hospital-based registry database from Imam Khomeini hospital, Ilam city, West of Iran, was retrospectively reviewed. Only COVID-19 patients with positive real-time reverse-transcriptase PCR (RT-PCR) test

admitted from January 9, 2020, until January 20, 2021, met our inclusion criteria to be included in this study. During this period, a total of 12885 suspected cases with COVID-19 referred to Imam Khomeini hospital ambulatory and Emergency Departments (EDs). Of those, 3350 cases were introduced as confirmed COVID-19 by RT-PCR test. After applying the inclusion/ exclusion criteria, finally, 1225 records were fed in the study (Figure 1). In order 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.

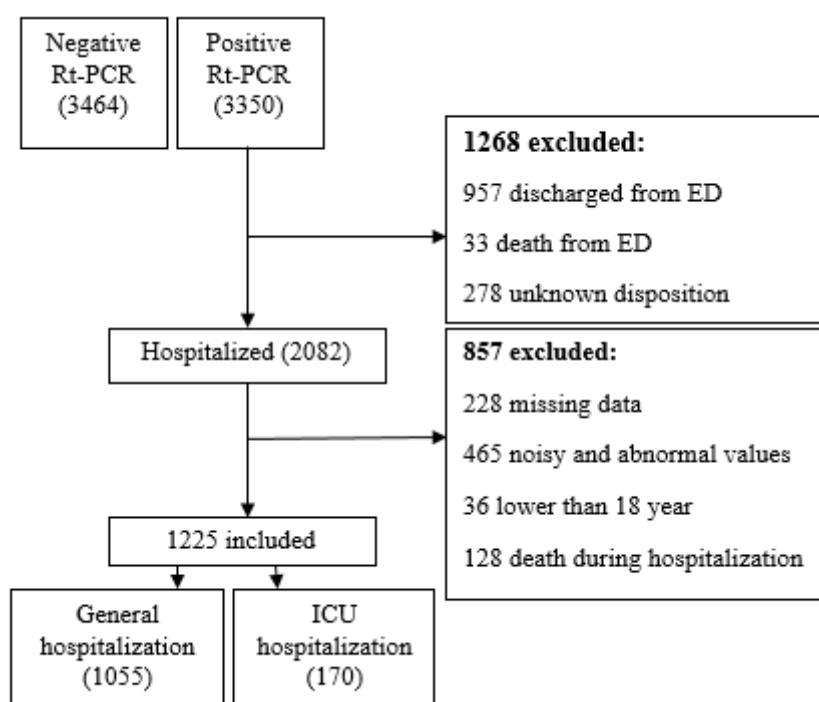


Figure 1: Flow chart describing patient selection

The exclusion criteria for patient selection included 1) Non-COVID-19 cases or non-hospitalized COVID-19 or patients with unknown disposition, 2) Patients who were less than 18 years of age, 3) Incomplete case records (missing more than 70%) and 4) Admission time before January 9, 2020, or after January 20, 2021.

The included cases were defined based on 53 risk factors in five main classes including patients' demographics (five variables), clinical manifestations (14 variables), comorbidities (seven variables), laboratory (26 variables), and treatment (one variable) (See Table 2).

Data preprocessing

Incomplete case records which had a lot of missing data (more than 70%) were excluded from the analysis. Also, the remaining missing values were imputed with the mean or mode of each variable. Noisy and abnormal values, errors, duplicates, and meaningless data were checked by researchers in collaboration with two infectious diseases specialists and hematologists. For different interpretations about data preprocessing, we contacted the corresponding physicians.

Feature selection

Feature selection or variable selection is an effective technique that is used to determine the most meaningful variables, and reduce the dimensions of the dataset and improve the efficiency of ML algorithms [40]. In this study, the variables with a correlation coefficient value less than 0.2 (P-value <0.2) were identified as effective risk factors in predicting the LOS of COVID-19 patients and included in the ML models.

Model development

To compare the performance of selected ML algorithms including Artificial Neural Network (ANN), Radial Basis Function (RBF), Support Vector Machine (SVM), Feedforward Neural Network (FNN), Probabilistic Neural Network (PNN), Pattern recognition network, and Decision Tree (DT), we carried out an experiment that concentrated on evaluating both the effectiveness and the efficiency of the models. The parameters of models used are shown in Table 1.

Table 1: Predictive model parameters

Model	Parameter
SVM	Kernel Function = Gaussian
RBF	Based on layers - 57-2 ; Spread = 150
ANN	57-10-5-2
FNN	Using python ToolBox; Create a FNN with FCM Clustering, specifying a sugeno FIS type; input: [1×57 struct]; output: [1×2 struct]; rule: [1×2 struct]; NumClusters = 2
PNN	Number of neuron 57-2 ; Spread = 0.1
Pattern recognition network	57-10-5-2
DT	---

All experiments were tested and implemented by using Python programming languages and the Scikit-learning library tools. Scikit-learning tools contain a set of ML algorithms for classification or prediction. ML techniques developed with this programming languages are used to a variety of real-world issues and it offers a well-defined framework for experimenters and developers to build and evaluate their models.

Experiment evaluation

In the present study, 10-fold cross validation method was applied to measure the unbiased estimate of prediction algorithm. To compare the performance of different algorithms in predicting LOS, several evaluation metrics including accuracy, sensitivity, specificity, and mean Area Under the Curve (AUC) was calculated. During the evaluation process, the confusion matrix was provided (for two classes).

Table 2: The calculation of performance metrics

Performance criteria	Calculation	
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$	
Sensitivity/ Recall	$\frac{TP}{TP + FN}$	
Specificity	$\frac{TN}{FP + TN}$	
ROC	$\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$	
Confusion matrix	Observed samples	Predicted samples
		True Positive (TP) False Positive (FP)
		False Negative (FN) True Negative (TN)

* FP=False Positive, TP=True Positive, FN=False Negative, TN=True Negative

Confusion matrix is a table that demonstrates a beneficial way to assess the performance of a classification model (or "classifier"). Each row in a confusion matrix shows an actual class while each column represents a predicted class (Table 2).

Ethical consideration

The study was approved by the ethical committee board of Ilam University of Medical Sciences (Ethic code: IR.MEDILAM.REC.1399.294). In order 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.

Result and Dissection

Demographic and clinical characteristics

After applying the exclusion criteria, a total of 1225 patients met eligibilities (Fig. 1). Of 1225 hospitalized COVID-19 patients, 664(54.20%) were males and 561 (45.80%) were females and the median age of participants was 57.25 (interquartile 18-100). 170 (13.87%) were hospitalized in ICU and 1055 (86.13%) hospitalized in general wards. Of these, 1136 (92.75%) were recovered and 89 (7.25%) were deceased. Descriptive statistics for the 1225 records in this dataset are shown in Table 3.

Table 3: The descriptive statistics of variables of the study after preprocessing

Qualitative variables	Values	Frequencies
Blood Type	A-, A+	17, 552
	B-, B+	13, 126
	O-, O+	29, 421
	AB-, AB+	6, 61
Gender	Male, Female	664, 561
Cough	Yes, No	958, 267
Contusion	Yes, No	409, 816
Nausea	Yes, No	401, 824
Vomit	Yes, No	346, 879
Headache	Yes, No	312, 913
Gastrointestinal symptoms	Yes, No	252, 973
Muscular pain	Yes, No	623, 602
Chill	Yes, No	591, 634
Fever	Yes, No	628, 597
Pneumonia	Yes, No	1044, 181
Oxygen therapy	Yes, No	1053, 172
Dyspnea	Yes, No	1078, 147
Loss of taste	Yes, No	272, 953
Loss of smell	Yes, No	305, 920
Runny Noise	Yes, No	437, 788
Sore throat	Yes, No	444, 781
Other underlined diseases	Yes, No	735, 490
Cardiac disease	Yes, No	306, 919
Hypertension	Yes, No	395, 830
Diabetes	Yes, No	268, 957
Smoking	Yes, No	41, 1184
alcohol addiction	Yes, No	11, 1214
C-reactive protein	Positive, Negative	1063, 162
Hypersensitive troponin	Positive, Negative	58, 1167
ICU admission (Outcome)	Yes, No	1055, 170
Quantitative variables	Range	Mean (SD)
Age (year)	18-100	57.25 (17.8)
Height	92-195	168.53 (8.5)
Weight	6.5-163	75.20 (13.0)
Creatinine	0.1-17.9	1.39 (1.4)
Red-cell count	1.38-13.1	4.56 (0.9)
White-cell count	1300-63000	8182.34 (4897.4)

Continue of Table 3

Qualitative variables	Values	Frequencies
Hematocrit	3.6-73.9	39.20 (6.7)
Hemoglobin	3.7-46	13.21 (2.4)
Platelet count	108000-691000	215493.66 (88380.1)
Absolute lymphocyte count	2-95	23.74 (11.8)
Absolute neutrophil count	8-98	74.52 (12.3)
Calcium	0.9-14.1	9.68 (0.8)
Phosphorus	2-12.4	3.50 (0.5)
Magnesium	1.14-19.1	2.16 (0.6)
Sodium	37-157	137.94 (5.3)
Potassium	2.5-14.2	3.98 (0.7)
Blood urea nitrogen	0.5-251	42.52 (31.7)
Total bilirubin	0.01-10	0.72 (0.7)
Aspartate aminotransferase	3.8-924	44.45 (53.5)
Alanine aminotransferase	2-672	38.29 (41.6)
Albumin	0.2-8.9	4.02 (0.5)
Glucose	18-994	136.09 (74.2)
Lactate dehydrogenase	4.6-6973	555.68 (339.0)
Activated partial thromboplastin time	1-120	28.56 (11.4)
Prothrombin time	0.9-46.8	12.82 (1.9)
Alkaline phosphatase	9.6-2846	213.12 (139.2)
Erythrocyte sedimentation rate	2-258	40.65 (28.8)

Variables included in the ML models

The results of feature selection for determining the most important diagnostic criteria affecting

COVID-19 hospital LOS based on the correlation coefficient at $P < 0.2$ are demonstrated in Table 4.

Table 4: The key diagnostic criteria at the $P < 0.2$

Variables	Pearson Correlation	P-value	Variables	Pearson Correlation	P-value
Age	-.045	*.119	Blood Type	-0.025128	.274
Height	-0.24	.409	Sex	-.107	.380
Weight	-0.25	.388	Cough	.299	*0.041417
Temperature	-0.32	.268	Contusion	-.122	.342
Creatinine	-.066	*.019	Hypertension	-.1744378	*0.0541
Red-cell count	.029	.315	Cardiovascular	.2746594	*0.125
White-cell count	-.054	*.057	Alcohol consumption	.7923469	0.218
Hematocrit	-.017	.562	Smoking	.3123146	0.354
Hemoglobin	-0.10	.724	Diabetes	.0980716	*0.104
Platelet count	.018	.532	Other underline disorders	.0904762	0.465
Absolute lymphocyte count	-.057	*.044	Sore throat	.0591151	0.640
Absolute neutrophil count	.061	*.033	Runny nose	-.1446846	0.253
Calcium	-.055	*.055	Loss of smell	.0335175	0.811
Phosphorus	-0.020	.476	Loss of taste	-.1192558	0.414
Magnesium	-.033	.243	Dyspnea	.4443414	*0.017
Sodium	-.015	.590	Oxygen therapy	.460136	*0.008
Potassium	.015	.607	Pneumonia	.2690936	*0.115
BUN	-.059	*.038	Fever	.0241734	0.842

Continue of Table 4

Variables	Pearson Correlation	P-value	Variables	Pearson Correlation	P-value
Total bilirubin	-.003	.915	Chill	.0269847	0.824
ASP	.054	*.057	Muscular pain	.0885438	0.466
ALT	.047	*.097	GI complications	-.20181	*0.179
Albumin	.024	.394	Headache	-.0297449	0.831
Glucose	.017	.552	Vomit	.0525788	0.696
LDH	.056	*.049	Nausea	.0083105	0.949
Activated partial thromboplastin time	-.036	.213	ESR	.040	*.157
Prothrombin time	.010	.714	Hyper sensitive troponin	-.2146846	0.439
Alkaline phosphatase	-.003	.929	C reactive protein	.0258788	*0.196

After feature selection, the 20 diagnostic criteria were acquired for the determined correlation coefficient at $P < 0.2$. These variables including age, creatinine, white-cell count, lymphocyte/neutrophil count, BUN, ASP, ALT, LDH, activated partial thromboplastin time, cough, hypertension, cardiovascular disorders, diabetes, dyspnea, oxygen therapy, pneumonia, GI complications, ESR, and C-reactive protein, were introduced as the most significant features (predictors) to predict hospital LOS.

Performance evaluation of models

The 10-fold cross-validation method was applied for running and evaluating the models, respectively. After pre-processing, we attempted to analyze the model performance by evaluation criteria including accuracy, sensitivity, specificity, and AUC-ROC. In this section, we assessed the effectiveness and efficiency of all classifiers with respect to running time, accuracy of classified cases and incorrectly classified cases (Table 5).

Table 5: Performance assessment of the classifiers

Evaluation criteria	Classifier						
	ANN	RBF	SVM	FNN	PNN	Pattern recognition network	DT
Time to build a model (s)	1530	1432	570	1370	825	735	170
Accuracy (%)	92	89	99	86	88	94	87

In order to better measure the actual performance of classifiers, 10 independent iterations of models were run. Finally, we evaluated the actual performance of our classifier in terms of standard deviation of accuracy, mean accuracy, mean specificity, and mean sensitivity (Table 6). Once the classifier algorithm is run, we

can investigate how efficient (performance) it is. Therefore, we compared the accuracy measurement of the predictive model based on the aforementioned criteria for ANN, RBF, SVM, FNN, PNN, Pattern recognition network, and DT techniques.

Table 6: The performance of the classifier for 10 independent iterations

Evaluation criteria	Classifier						
	ANN	RBF	SVM	FNN	Patternnet	Pattern recognition	DT
Standard deviation of accuracy	4.8842e-01	4.3289e-01	9.2124e-02	1.8027e-01	5.4548e-01	1.9826e-01	4.6511e-01
Mean Accuracy	9.2278e+01	8.9233e+01	9.9581e+01	8.6473e+01	8.8769e+01	9.4042e+01	8.7397e+01
Mean Specificity	9.6864e+01	9.8532e+01	9.9743e+01	8.7589e+01	9.1276e+01	9.9676e+01	8.5442e+01
Mean Sensitivity	8.7693e+01	7.9934e+01	9.9418e+01	8.5357e+01	8.6263e+01	8.8409e+01	8.9353e+01
MIN Accuracy	9.1468e+01	8.8751e+01	9.9429e+01	8.6177e+01	8.7844e+01	9.3709e+01	8.6799e+01
Max Accuracy	9.2992e+01	9.0181e+01	9.9714e+01	8.6702e+01	8.9655e+01	9.4281e+01	8.8131e+01



Figure 2: Confusion matrix calculation for the selected ML algorithms

The results of comparing confusion matrix metrics and AUC-ROC of different classifiers are shown in Figure 2 and 3.

In the present study, the prediction model which has higher performance based on evaluation criteria such as accuracy, specificity, sensitivity, and running time, was chosen as the best algorithm and implemented in a Clinical Decision Support System (CDSS). Table 6 represents the different measures of performance for the ML models. According to the experimental results of

the evaluation of selected ML models in 10-iterations, the SVM algorithm with the mean accuracy of 99.5%, mean specificity of 99.7%, mean sensitivity of 99.4%, and the standard deviation of 1.2., gained higher performance than the other techniques. The AUC-ROC for SVM was 99.8%. Besides, we observed that the SVM algorithm takes unto 0.07 second to build its model as the fastest, and unlike ANN takes about 1530 (s) that was the slowest.

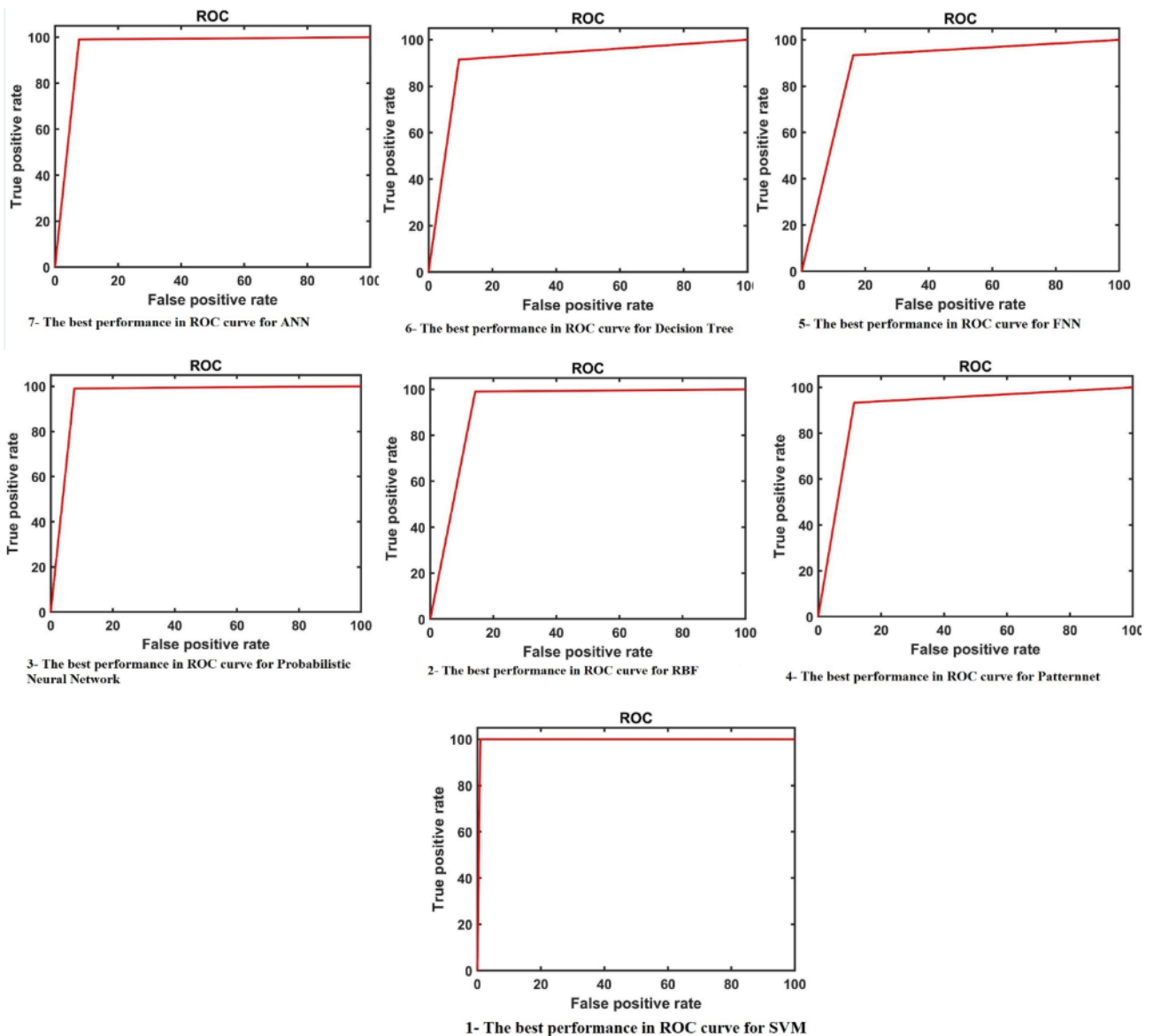


Figure 3: The ROC curve of the all ML algorithms

The results of comparing confusion matrix metrics and AUC-ROC of different classifiers are shown in Figure 2 and 3.

In the present study, the prediction model which has higher performance based on evaluation criteria such as accuracy, specificity, sensitivity, and running time, was chosen as the best algorithm and implemented in a Clinical Decision Support System (CDSS). Table 6 represents the different measures of performance for the ML models. According to the experimental results of the evaluation of selected ML models in 10-iterations, the SVM algorithm with the mean accuracy of 99.5%, mean specificity of 99.7%, mean sensitivity of 99.4%, and the standard

deviation of 1.2., gained higher performance than the other techniques. The AUC-ROC for SVM was 99.8%. Besides, we observed that the SVM algorithm takes unto 0.07 second to build its model as the fastest, and unlike ANN takes about 1530 (s) that was the slowest.

This study intended to compare the accuracy and efficiency of selected ML techniques for COVID-19 in-hospital LOS prediction. The need for this research derived from the increasing demand for ML capabilities, as a scientific and objective measures to predict the LOS. During this time, most healthcare settings have experienced with capacity reduction due to high referral volumes and bed occupancy as well as prolonged

hospitalization [41]. The exact prediction of LOS can support the bed administration and projecting future requirements for optimal medical resource allocation [17, 21]. Predicting hospital bed request (as well as associated medical resources) offer key evidence for hospital staffing and resource planning decisions. It is significant for clinicians and health policymakers to make proper decisions for allocating of restricted resources [21, 42]. Using ML based prediction models (intelligence system) is proven to be useful for optimum LOS estimation. This led to reducing uncertainty and ambiguity by offering systematic and evidence based system for hospital resource utilization and care planning [11, 43].

For this purpose, several ML methods, including ANN, RBF, SVM, FNN, PNN, Pattern recognition network, and DT were fed by using the optimized predictor variables. Feature selection is a significant step to prepare and customize the data before feeding it to the ML classifiers [44]. In this study, 53 primary features are reduced to 20 by using the correlation coefficient at the P -value < 0.2. In the bibliography, some studies have been undertaken to identify the key risk factors for COVID-19 hospital LOS [17, 21, 42]. The top clinical variables affecting longer LOS in reviewed studies included age (basic data), cardiovascular diseases and hypertension (underline diseases), fever and low oxygen saturation (manifestations), leukocytosis (immunological), pulmonary lesion (radiological), mechanical ventilation (oxygen therapy) and increased BUN. In general, high compliance was observed from the results of classifying and prioritizing variables in reviewed studies with the most common variables in the current study (Table 5).

So far, multiple studies have been conducted on the application of ML techniques to predict the LOS in hospitalized patients [11, 32, 34, 45-47]. It is proven that ML can be used with myriad applications for hospital LOS during the COVID-19 pandemic [42]. However to the best of our knowledge, limited studies have been done on the use of ML techniques in the prediction of COVID-19 LOS [17]. For example, Dan et al. (2020) in a retrospective study developed an SVM-based

model to predict the COVID-19 length of ICU stay. Finally the results showed good performance for predicting the LOS with AUC-ROC of 91% and Mean Absolute Error (MAE) of 0.723 [17]. Pei et al. [42] in their study assessed the performance of selected ML algorithms including K-nearest Neighbors(K-NN), Logistic Regression (LR) and Random Forest (RF) for prediction of patients' LOS at hospital during COVID-19 pandemic with the accuracy of 0.3442, 0.3524 and 0.3541, respectively [42]. Kabir [44] and Hijry (2020) developed a prediction model to anticipate the LOS and the results presented that Back Propagation(BP) Neural Network with an accuracy of 92.58% outperformed all other ML models examined [44, 48]. Mahboub [22] utilized DT classifier for predicting COVID-19 patients' hospital LOS. The experimental result showed that this algorithm with sensitivity of 96.5%, specificity of 87.8%, and accuracy of 96% has excellent performance [22]. Kulkarin (2021) designed a Neural Network-Multi Layered Perceptron (ANN-MLP) based model for predicting prolonged LOS of patients with an accuracy of 90.87% [49]. Sinha et al. (2021) also showed that ANN predictive performance for length of ICU stay in hospitalized patients with COVID-19 gained the best performance with Root Mean Square Error (RMSE) and MAE of 5.9451 and 4.6354, respectively [21]. East's [50] results showed that the model developed with ANN yielded the best performance to predict long LOS (AUC with 0.9760%) [50]. Chiari et al. [51] compared the performance of two RF and Extra Trees regression algorithms for COVID-19 LOS prediction. The experimental results showed good performance for the LOS prediction with the accuracy of 98% and 95% respectively [51].

In this study, multiple ML-based prediction models including ANN, RBF, SVM, FNN, PNN, Pattern recognition network, and DT were trained and evaluated to determine the most optimal algorithm for predicting the COVID-19 LOS. Unlike previous studies where ANN techniques has better performance for forecasting LOS, the obtained results in 10 iteration execution of the selected ML algorithms in the present work showed that the SVM

classifier with mean accuracy of 99.5%, mean specificity and sensitivity of 99.7% and 99.4%, respectively, having more predictive capabilities compared with other ML methods. The suggested model in this study can estimate the LOS of patients with optimal performance. It provides a better plan for hospital administrators, policy makers and clinicians in order to improve patient outcomes and quality of care especially in organizations with resource challenges. This led to decreasing ambiguity by offering scientific and evidence-based model for resource utilization and episode of care planning. But the model is inputted with a slight number of features (20 features), yet provides a precise calculation of the LOS. Moreover, the proposed model is simple, correct, and can be effortlessly implemented in clinical practice.

Limitations

This study had some limitations that necessary to be recognized. First, we dealt with a retrospective dataset that may suffer from imbalanced, noisy, duplicates, and meaningless values, which may cause of prediction bias. Second, this study was conducted at a single center and only based on 1225 data, so confined the generalizability of the predictive model and may have affected the performance metrics of the proposed models. Moreover, we only used seven ML algorithms for prediction analyses. Finally, the selected dataset lacked some important clinical variables such as radiological indicators. In the future, the performance accuracy of our computational model will be improved if we test more ML techniques, at larger, multicenter and prospective dataset equipped with more qualitative data regarding more diverse variables.

Conclusion

Estimating the LOS of hospitalized patients with COVID-19 by offering an objective and evidence-based approach is crucial for effective bed management, better patient scheduling, proper staffing and customized resource allocation. In this study, at first, a statistical based feature selection method was applied for predicting COVID-19 in-hospital LOS. Then, we developed and evaluated several ML models to predict the LOS of COVID-19 patients using routine clinical

data set. The evaluation of selected ML techniques performance demonstrated the suitability of these models, in particular the SVM model, for predicting in-hospital LOS. This model has the potential to augment informed decisions for effective management of COVID-19 patients. Besides, it can support the sharing of restricted hospital resources and enhance health care quality.

Acknowledgment

This article is extracted from a research project supported by the Ilam University of Medical Sciences (IR.MEDILAM.REC.1399.294). We also thank the Research Deputy of the Ilam University of Medical Sciences for financially supporting this project.

Funding

This research received financial support from Ilam University of Medical Sciences.

Authors' contributions

All authors contributed toward study designing, data mining, statistical analysis, reporting the results and agreed to be responsible for all the aspects of this work.

Conflict of Interest

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

References

- [1]. Dinicolantonio J.J., McCarty M., Barroso-Aranda J., *Open Heart*, 2021, **8**:1 [[Crossref](#)] , [[Google scholar](#)], [[Publisher](#)]
- [2]. Tan J.Y., Sim X.Y.J., Wee L.E., Chua Y.Y., Cherng B.P.Z., Ng I.M., *J. Med. Virol.*, 2021, **93**:1548 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [3]. Arian M., Valinejadi A., Vakilian F., *Iran. J. Publ. Health*, 2021, **50**:46 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [4]. Rehman A., Iqbal M.A., Xing H., Ahmed I., *Appl Sci. (Switzerland)*, 2021, **11**:3414 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [5]. da Rosa Mesquita R., Francelino Silva Junior L.C., Santos Santana F.M., Farias de Oliveira T., Campos Alcântara R., Monteiro Arnozo G.,

- Wien. Klin. Wochenschr., 2021, **133**:377 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [6]. Jahangir M., Nawaz M., Nanjiani D., Siddiqui M.S., *Hong Kong Med. J.*, 2021, **27**:35 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [7]. Hong Y., Wu X., Qu J., Gao Y., Chen H., Zhang Z., *Ann. Transl. Med.*, 2020, **8** [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [8]. Smati H., Cohen P.A., Nagda D.V., Saravanan Y., Kalugin P.N., Li C.Y., *J. Immigr. Minor. Health*, 2021 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [9]. Ferrell E., Drabiak K., Alfano-Torres M., Ahmed S., Ali A., Bjornstad B., *Clin. Ethic.*, 2021, 14777509211016287 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [10]. Liu S., Xi H.T., Zhu Q.Q., Ji M., Zhang H., Yang B.X., *PeerJ.*, 2021, **9**:e11154 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [11]. Daghistani T.A., Elshawi R., Sakr S., Ahmed A.M., Al-Thwayee A., Al-Mallah M.H., *Int. J. Cardiol.*, 2019, **288**:140 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [12]. Perillat L., Baigrie B.S., *J. Eval. Clin. Pract.*, 2021, **27**:708 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [13]. Valinejadi A., Sadoughi F., Salehi M., *Int. J. Prev. Med.*, 2016, **7**:33 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [14]. Wu S., Xue L., Legido-Quigley H., Khan M., Wu H., Peng X., *Plos one*, 2020, **15**:e0240959 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [15]. Ziari A., Ansari M., Valinejadi A., *Koomesh*, 2018, **20**:221 [[Google scholar](#)], [[Publisher](#)]
- [16]. Sreevalsan-Nair J., Vangimalla R.R., Ghogale P.R., *medRxiv*, 2020 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [17]. Dan T., Li Y., Zhu Z., Chen X., Quan W., Hu Y., *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, IEEE [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [18]. Yousefi M., Aliani S., Valinejadi A., Rezazadeh A., Khorsand A., Fazaeli S., *Koomesh*, 2018, **20**:403 [[Google scholar](#)], [[Publisher](#)]
- [19]. Shojaei S., Yousefi M., Ebrahimipour H., Valinejadi A., Tabesh H., Fazaeli S., *Koomesh*, 2018, **20**:283 [[Google scholar](#)], [[Publisher](#)]
- [20]. Fazaeli A.A., Fazaeli A.A., Hamidi Y., Moeini B., Valinejadi A., *Koomesh*, 2018, **20**:358 [[Google scholar](#)], [[Publisher](#)]
- [21]. Sinha S., Goel S., *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, IEEE [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [22]. Mahboub B., Al Bataineh M.T., Alshraideh H., Hamoudi R., Salameh L., Shamayleh A., *Front. Med.*, 2021, **8** [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [23]. Roimi M., Gutman R., Somer J., Ben Arie A., Calman I., Bar-Lavie Y., Gelbshtein U., Liverant-Taub S., Ziv A., Eytan D., Gorfine M., *medRxiv*, 2021 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [24]. Bacchi S., Tan Y., Oakden-Rayner L., Jannes J., Kleinig T., Koblar S., *Inter. Med. J.*, 2020 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [25]. Askari-Majdabadi H., Valinejadi A., Mohammadpour A., Bouraghi H., Abbasy Z., Alaei S., *Acta. Inform. Med.*, 2019, **27**:311 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [26]. Mohammadi A., Valinejadi A., Sakipour S., Hemmat M., Zarei J., Majdabadi H.A., *Int. J. Health Pol. Manag.*, 2018, **7**:336 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [27]. Alipour J., Safari Lafti S., Askari Majdabadi H., Yazdiyani A., Valinejadi A., *IIOAB J.*, 2016, 119 [[Google scholar](#)], [[Publisher](#)]
- [28]. Shanbehzadeh M., Nopour R., Kazemi-Arpanahi H., *J. Adv. Med. Biomed. Res.*, 2021, **29**:100 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [29]. Nassif A.B., Azzeh M., Banitaan S., Neagu D., *Springer*, 2016, **27**:2153 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [30]. Mei X., Lee H-C., Diao K-y., Huang M., Lin B., Liu C., *Nat. Med.*, 2020, **26**:1224 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [31]. Alshakhs F., Alharthi H., Aslam N., Khan I.U., Elasheri M., *Int. J. Gen. Med.*, 2020, **13**:751 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [32]. Alsinglawi B., Alnajjar F., Mubin O., Novoa M., Alorjani M., Karajeh O., *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, IEEE [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]

- [33]. Triana A.J., Vyas R., Shah A.S., Tiwari V., *J. Surg. Res.*, 2021, **264**:68 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [34]. Muhlestein W.E., Akagi D.S., Davies J.M., Chambless L.B., *Clin. Neurosurg.*, 2019, **85**:384 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [35]. Udelsman B.V., Jones P.S., Bababekov Y.J., Carter B.S., Chang D.C., *Clin. Neurosurg.*, 2019, **85**:E444 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [36]. Jo Y.Y., Han J., Woo Park H., Jung H., Lee J.D., Jung J., *JMIR Med. Inform.*, 2021, **9**:e23147 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [37]. Han C., Liu J., Wu Y., Chong Y., Chai X., Weng X., *Front. Surg.*, 2021, **8** [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [38]. Elbattah M., Molloy O., *Using machine learning to predict length of stay and discharge destination for hip-fracture patients*. Lecture Notes in Networks and Systems, Springer, Cham. 2016, 207 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [39]. Ramkumar P.N., Navarro S.M., Haeberle H.S., Karnuta J.M., Mont M.A., Iannotti J.P., *J. Arthroplasty*, 2019, **34**:632 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [40]. Karegowda A.G., Manjunath A., Jayaram M., *Int. J. Inform. Technol. Knowl. Manag.*, 2010, **2**:271 [[Google scholar](#)]
- [41]. Panocchia N., D'Ambrosio V., Corti S., Lo Presti E., Bertelli M., Scattoni M.L., *J. Med. Ethic.*, 2021, **47**:362 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [42]. Pei J., Lin X., Chen Q., *J. Phys: Conference Series*, 2021, IOP Publishing [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [43]. Goshtasbi K., Yasaka T.M., Zandi-Toghiani M., Djalilian H.R., Armstrong W.B., Tjoa T., *Head Neck*, 2020 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [44]. Kabir S., Farrokhvar L., *2019 18th IEEE International Conference On Machine Learning and Applications (ICMLA)*; 2019, IEEE [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [45]. Alsinglawi B., Alnajjar F., Mubin O., Novoa M., Alorjani M., Karajeh O., *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [46]. Alghatani K., Ammar N., Rezgui A., Shaban-Nejad A., *JMIR Med. Inform.*, 2021, **9** [[Crossref](#)]
- [47]. Staziaki P.V., Wu D., Rayan J.C., Santo I.D.O., Nan F., Maybury A., *Europ. Radiol.*, 2021, **31**:5434 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [48]. Hijry H., Olawoyin R., *2020 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2020, IEEE [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [49]. Kulkarni H., Thangam M., Amin A.P., *Europ. J. Clin. Invest.*, 2021, **51**:e13406 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]
- [50]. East A., Ray S., Pope R., Cortina-Borja M., Sebire N.J., *BMJ J.*, 2020, A15 [[Google scholar](#)], [[Publisher](#)]
- [51]. Chiari M., Gerevini A.E., Maroldi R., Olivato M., Putelli L., Serina I., *Pattern Recognition ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part I*; Springer International Publishing, 2021, 212 [[Crossref](#)], [[Google scholar](#)], [[Publisher](#)]

HOW TO CITE THIS ARTICLE

Mohammad Reza Afrash, Hadi Kazemi-Arpanahi, Parvaneh Ranjbar, Raoof Nopour, Morteza Amraei, Mojgan Saki, Mostafa Shanbehzadeh . Predictive Modeling of Hospital Length of Stay in COVID-19 Patients Using Machine Learning Algorithms, *J. Med. Chem. Sci.*, 2021, 4(5) 525-537
 DOI: 10.26655/JMCHMSCI.2021.5.15
 URL: http://www.jmchemsci.com/article_136288.html