Predicting intensive care unit admission of COVID-19 patients with open data: analysis of the first wave in Colombia

Ruben Acosta-Velasquez[1‡], William Fajardo-Moreno[1], Leonardo Espinosa-Leal[2*‡],

1 EAN University, Bogotá D.C., Colombia
2 Department of Business Management and Analytics, Arcada University of Applied Sciences, Jan-Magnus Jansson aukio 1, 00560, Helsinki, Finland

‡These authors contributed equally to this work.
* leonardo.espinosaleal@arcada.fi

Abstract

Optimizing intensive care resources using predicting modeling is paramount for fighting the COVID-19 pandemic. In this paper, we model the admission of COVID-19 patients in intensive care units (ICU) in Colombia using openly available data gathered from 18 March 2020 to 14 October 2020. After an intensive preprocessing of the data, we trained four different machine learning models using four different strategies for handling the imbalanced features. Our findings show that our best model (XGBoost) effectively predicts an Area Under the Curve (AUC-ROC) of 0.94, in line with the state-of-the-art results obtained in other predictive models obtained with medical data.

1 Introduction

In, January 2022, after more than two years since the appearance of COVID-19 and more than 5 million global deaths and more than 400 million global cases [1], the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has turned into one of the biggest global societal challenges of the 21st century due to its collateral consequences in mental health, economic and societal levels. From the perspective of the health care systems, compared to previous global pandemics, the main difference is the ratio of people infected in need of intensive care. COVID-19 virus has sent patients to Intensive Care Units (ICU) more often than previous related coronaviruses. These units contain artificial breathing systems necessary to keep the patients alive. The collapse of these units has been fundamental in augmenting the number of dead patients. To avoid an excess of casualties, it is necessary to create robust and accurate strategies to predict when patients need to be transferred to ICU.

Machine learning methods have become ubiquitous in all areas of knowledge and become a strategic tool for inference and prediction [2–4]. Healthcare is not an exception, and given the importance of finding strategies to fight against COVID-19, a thousand research papers have recently appeared in prognostic and diagnostic modeling. Given the importance of this research topic for improving the global conditions of the general population, exploring as many areas as possible is of paramount importance. However, some efforts to improve the clarity of the reported results and possible biases are needed [5]. More recently, some critics have pointed out that several machine learning models have been deployed in hospitals without proper testing [6]. Moreover, the consensus among specialists is that most of the developed models have not been helpful in the fight against COVID-19 [5, 7].

 doi:10.20944/preprints202212.0330.v1

Because of the synergies created due to the intense digitization of medical services [8] and the rapid reaction of several research groups around the world, the research literature has exploded, and many research products predicting the expansion of COVID-19 at different stages have appeared in the last year [9]. Moreover, despite the efforts to make available as much as possible the gathered data, it is essential to highlight that most of the models have been created using medical data records that, for privacy reasons, are difficult to share openly. Moreover, despite the high scores in the obtained models, it is clear that the lack of diversity in the fitted data makes difficult the broad adoption of most of the published results [10]. In this regard, one of the best use of machine learning methods should be in epidemiological modeling, for instance, by predicting the possible contagious mortality rate of ICU occupancy. The obtained models could help improve decision-making at different institutional levels.

In this work, to the best of our knowledge, we present the first machine learning models for predicting ICU admission of patients affected by COVID-19 fitted using openly available data, without including risks factor or comorbidities. This paper is organized as follows; in the first part, understandable state-of-the-art machine learning models used to predict ICU admission are presented and discussed. In the next part, the Research methodology shows the nature of the data used, including the preprocessing and the machine learning models employed. Then, in the Results and Discussion, we present the outcomes of the studied models and their comparison with the scientific literature. The last part concludes our research by presenting the learned lessons and future research.

2 state-of-the-art

The importance of finding effective strategies using numerical methods to fight COVID-19 at all levels has mobilized numerous academic resources and collaboration to create valuable models. The following is a comprehensive review of the literature in which research related to the use of machine learning methodologies in the prediction for the use of intensive care units for patients with positive cases of COVID-19 at the international level is presented. For each job, the size of the dataset, the algorithms used, and the performance measures with their associated results are presented.

In the prediction models of intensive care units, we find the first approximation from the statistical models within which we can find works such as that of Bonadia et al. [11]. The authors used data from patients from an Italian hospital to predict the mortality and the intensive care unit of COVID-19 patients through lung ultrasound data. In this work, the authors present an assessment scale used to estimate the severity identified in the ultrasound scans of each patient; this value allowed to significantly predict the final clinical result (death/survival) and the need for admission to the ICU.

In the work by Kottlors et al. [12], researchers used data from patients from two German hospitals and used a multivariate logistic regression analysis to study 58 individuals with confirmed COVID-19 infection. They started with a low-dose computed tomography (LDCT) scan of each patient, and estimates of body composition were made. As a result, it was possible to identify that the relationship between waist circumference and paravertebral muscle circumference (FMR), in addition to age, were significant predictors of the need for treatment in the ICU of mentioned patients

Salles-Neto et al. [13] developed an open application named Forecast UTI, which uses polynomial regression models adjusted to the fifth degree to forecast the number of ICU beds occupied by COVID-19 patients. They calculated the root-mean-square of forecast error (RMSE) for each model and chose the model with the lowest RMSE. The data used are artificial and are helpful to observe the performance of the information system. The Forecast UTI software is proposed as a resource to support predictions on the needs

 doi:10.20944/preprints202212.0330.v1

of intensive care units in health systems, especially the Brazilian healthcare system.

Another investigation of the statistical model line was done by Allenbach et al. [14]. Here, data from 152 patients of a French hospital to construct a predictive model. External validation was carried out with 132 patients from a different French hospital. For the modeling, logistic regression was used, and internal validation was carried out from re-samplings. It stands out that up to 35% of the patients with COVID-19 hospitalized in a medical ward were transferred to the ICU or died on day 14. The transfer rate to the ICU was also quantified at 11.6%.

Colombi and coauthors [15] proposed four models using logistic regression to evaluate the relationship between clinical parameters and chest computed tomography metrics versus patient outcomes in terms of whether they were admitted to the ICU. One of the models only considered clinical data. At the same time, the remaining three included a progressive addition of categories. For this research, they analyzed 236 patients from the emergency department of an Italian hospital. As a result, the model showed a performance for models that incorporated clinical aspects, and in addition to data from other categories, an Area Under the Curve (AUC-ROC) of 0.86 was obtained (See Table 1). Schalekam et al. [16] proposed a risk model to predict admission to the ICU or death, based on the analysis of data from 356 patients from two Dutch hospitals and using logistics regression. The data considered in the model included clinical data, chest radiographs, and laboratory results. The model had a performance measured in the AUC-ROC of 0.77.

Chao et al. [17] used data from 295 patients from three hospitals located in the United States, Iran, and Italy, related to demographic aspects, vital signs, and laboratory findings of patients with diagnostic images of computed tomography of the chest, to predict the need for intensive care units (ICU) by patients with positive COVID-19, using deep learning techniques; the results obtained show an outstanding performance in the prediction, obtaining an AUC-ROC of 0.884 and a sensitivity of 96.1%.

On the other hand, Schwab et al. [18] used data from 5644 patients from a hospital in Brazil to evaluate the performance of clinical prediction models, which used autonomous learning. The purpose was to predict the positive diagnosis of COVID-19 positive, or whether they will need hospitalization or intensive care. In this research, techniques such as logistic regression, neural networks, support vector machines, random forests, and increasing gradients were used. Specifically, relating to the prediction of intensive care units, this research managed to obtain an AUC-ROC performance of up to 0.98, the best performances are in the models that used logistic regression and support vector machines (See Table 1).

Another investigation carried out by Wollenstein-Betech et al. [19] using open data offered by the Mexican government on patients who presented tests for COVID-19, developed personalized models for the prediction of hospitalization, mortality, the need for an ICU, and the need for a ventilator for patients with a positive diagnosis for SARS-CoV-2. In this research data from 91,000 patients were used, and the analyzed information were demographic aspects, previous medical conditions, test results, hospitalization, mortality and whether a patient has developed pneumonia or not. For the development of the models, classification methods such as logistic regression and support vector machines were used, as well as random forests and decision trees powered by gradients.

Regarding the prediction of intensive care units, a precision of 80% with an AUC of 0.54 was obtained as a performance measure, in addition, the model also allowed us to identify that among the most important conditions of said prediction are the development of pneumonia (if available), cardiovascular disease, asthma, and COVID-19 test status. Having data on the development of pneumonia by the patient improves the

 doi:10.20944/preprints202212.0330.v1

performance of the model, increasing the precision to 82% and the AUC-ROC to 0.63 (See Table 1).

Another important contribution was made from the research of Mejía-Vilet [20], in this work, data were taken from 569 patients from a Mexican hospital divided into two cohorts, one for development and the other for validation, within said data aspects were considered demographics, medical history, and laboratory tests, in addition to a series of scores that offer clinical models on the behavior and evolution of the patient's health, based on these three prediction models were built using logistic regression analysis, the first it was called ABC-GOALSc which considered only clinical variables, the second ABC-GOALScl which considered clinical and laboratory variables and the third, ABC-GOALSclx, considered clinical variables, laboratory and x-ray images. Of the three previously mentioned models, the one that presented the best performance was the so-called ABC-GOALSclx, achieving an AUC-ROC of 0.86 (See Table 1).

Cohen et al. [21] also included in their research a prediction on intensive care units, in this work, a set of data on diagnostic chest images was used, which were obtained from different sources and contained information from different countries and continents. The training of the developed model was carried out through linear or logistic regression with predetermined parameters of Sci-kit learn and considered six characteristics, the performance of the model was measured with AUC-ROC and AUC-PRC (area under the recovery curve precision), and the results are presented in (See Table 1).

In another case, Zhao et al. [22] used data from 641 patients from a hospital in the United States, which contained information related to demographic medical records, comorbidities, and laboratory tests, based on these data a logistic regression was used for the construction of the model, the Performance precision was evaluated using AUC-ROC, in which a value of 0.74 ([95% CI % 0.63–0.85], p = 0.001) was obtained. Additionally, five significant variables were identified that predict admission to ICU, which are presented in Table 1. Vaid et al. [23] based on data from 4098 patients from five hospitals in the United States, developed a model for predicting mortality and the need for a critical care unit (ICU), using the XGBoost classifier, the predictions were established for 3, 5, 7, and 10 days. As a result, an AUC of 0.80 at 3 days, 0.79 at 5 days, 0.80 at 7 days, and 0.81 at 10 days was obtained.

In the research by Li et al. [24], the objective was to predict admission to the intensive care unit (ICU) and hospital mortality in patients with COVID-19 from a risk scoring system that used clinical variables, to achieve this they considered data from 5,766 patients suspected of being infected with COVID-19 in the period between February and May 2020 in the United States. Based on the aforementioned data, a deep neural network (DNN) prediction model was constructed, which showed the following performance data: AUC-ROC = 0.780 (95% CI [0.760–0.785]), sensitivity = 0.760, specificity = 0.709 and F1-score = 0.551 in the prediction of ICU admission.

Cheng et al. [25] used data from 1987 patients registered in a hospital in the United States, diagnosed with COVID-19 between February and April 2020, these data considered demographic information, clinical evaluations, and laboratory results. The main objective of the study was to predict admission to the ICU within 24 hours from the moment of prediction, for which a Random Forest (RF) model was used. As a result, an AUC-ROC of 0.95 was obtained.

Dan et al. [26] used data from 733 patients diagnosed with COVID-19, 909 variables in three categories: demographic, clinical and course examinations, and laboratory indicators, with this dataset, used a support vector machine (SVM) models to predict ICU admission. The results show an accuracy score was 0.83 and an AUC-ROC score was 0.84. Also analyzed, the model considered ten clinical variables separately and the AUC-ROC was between 0.95 (oxygen saturation) and 0.83 (lymphocyte absolute value).

Podder et al. [27] using models of random forest, XGBoost, logistic regression, Naïve

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2022 doi:10.20944/preprints202212.0330.v1

Bayes, Light Gradient Boosting Machine (LGBM), Multi-Layer Perceptron (MLP) and extra trees predicted ICU requirement. The authors used data from 5644 samples and 11 attributes. This dataset was imbalanced as 90.10% samples were for negative cases representing people without COVID-19. As a result, COVID-19 detection can be predicted with an AUC-ROC of 91% using a stacking ensemble with Naïve Bayes, LGBM, and logistic regression; another stacking and voting show AUC-ROC scores between 0.89 and 0.91; regarding individual models MLP, LGBM and Naïve Bayes achieved an AUC-ROC score of 0.9.

Subudhi et al. [28] compared the performance of 18 machine learning algorithms to predict ICU admission, the algorithms belonged to 9 broad categories: ensemble, Gaussian process, linear, Naive Bayes Machine-based, tree-based, nearest neighbor, support vector neural network models and discriminant analysis. The data set to train the model used in this investigation included 1,144 patients, and the validation data set included 334 patients. The results show that all ensemble-based models had a mean precision-recall area under the curve (AUC-PR) score of more than 0.77, the best score for AUC-ROC were for the AdaBoost Classifier, Random Forest Classifier, and Bagging Classifier, which got a value of 0.8.

Jamshidi et al. [29] used data from 263 adult patients with COVID-19 infection, who were admitted to ICUs at different hospitals to make an early prediction of mortality using machine learning. The models used were Logistic regression and random forest, as a result, the random forest model shows superior performance on both training and validation sets and predicts patient outcomes with a 0.7 sensitivity and 0.75 specificity.

Lorenzen et al. [30] used Random Forest models to predict ICU admission and ventilator use for all patients with a positive COVID-19 test. The sample of patients was 34012 patients who tested positive for COVID-19. The results show a prediction for a 5-day risk of ICU admission with an AUC-ROC of 0.986 and a 5-day risk of use of ventilation with an AUC-ROC of 0.995.

Finally, Heo et al. [31] who retrospectively analyzed the data of 5193 patients from one hundred Korean hospitals, said the data included seven variables and corresponded to the period between January and June 2020. In this work, two logistic regression models were used to predict admission to nursing units. intensive care, each of them, was differentiated by its predictors. The results showed that the two models had similar behavior, while the one that exclusively included clinical variables achieved an AUC-ROC of 0.884; the model that considered radiological and laboratory variables obtained an AUC-ROC of 0.880.

Table 1. Summary of the state-of-the-art.

Author	Algorithm and methodology	AUC-ROC	Dataset Size
Colombi et al. (2020) [15]	Logistic regression with clinical parameters and % V-WAL $< 73\%$	0.83 (0.78,0.88)	236 patients
	Logistic regression with clinical parameters and % S-WAL $< 71\%$ and AT area>262 cm^2	0.86 (0.80,0.90)	
	Logistic regression with clinical parameters and % VOL-WAL < 2.9 L% and AT area>262 cm^2	0.86 (0.81,0.90)	
Schwab et al. (2020) [18]	Support Vector Machines	0.98 (0.95,1)	5644 patients
	Logistic regression	0.98 (0.93,1)	
	Neural networks	0.97 (0.94,0.99)	
	Random Forest	0.97 (0.92,1)	
Continued on next page			

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2022 doi:10.20944/preprints202212.0330.v1

Table 1 – continued from previous page

Author	Algorithm and methodology	AUC-ROC	Dataset Size
	XGBoost	0.67 (0.53,0.98)	
Wollenstein-Betech et al. (2020) [19]	Support Vector Machines - Before knowing if the patient has developed pneumonia	0.538	91000 patients
	Logistic regression - Before knowing if the patient has developed pneumonia	0.548	
	Random Forest - Before knowing if the patient has developed pneumonia	0.541	
	XGBoost - Before knowing if the patient has developed pneumonia	0.554	
	Support Vector Machines - After knowing if a patient has developed pneumonia or not	0.623	
	Logistic regression - After knowing if a patient has developed pneumonia or not	0.633	
	Random Forest - After knowing if a patient has developed pneumonia or not	0.630	
	XGBoost - After knowing if a patient has developed pneumonia or not	0.639	
Mejía-Vilet et al. (2020) [20]	Logistic regression - ABC-GOALSc (Validation cohort)	0.77 (0.71, 0.86)	569 patients
	Logistic regression - ABC-GOALScl (Validation cohort)	0.87 (0.83, 0.92)	
	Logistic regression - ABC-GOALSclx (Validation cohort)	0.86 (0.81, 0.90)	
Cohen et al. (2020) [21]	Logistic regression	0.81 (0.70, 0.90)	369 patients
Zhao et al. (2020) [22]	Logistic regression	0.74	641 patients
Vaid et al. (2020) [23]	XGBoost at 3 days	0.8	4098 patients
	XGBoost at 5 days	0.79	
	XGBoost at 7 days	0.8	
	XGBoost at 10 days	0.81	
Dan et al. (2020) [32]	Support Vector Machines (SVM) Lasso models	0.95	733 patients
Cheng et al. (2020) [25]	Random Forest	0,74 ([95%, CI [0.63–0.85], p = 0.001)	1987 patients
Li et al. (2020) [24]	Deep Neural Network (DNN)	0,780 (95% CI [0,760–0,785])	5766 patients
Heo et al. (2021) [31]	Logistic regression models (Models with clinical variables)	0.884	5193 patients
	(Model with clinical, radiological and laboratory variables)	0.880	
	Random Forest (RF)	0.99	
	Multilayer perceptron (MLP)	0.95	
	Light Gradient Boosted Machine (LGBM)	0.88	

Continued on next page

 doi:10.20944/preprints202212.0330.v1

Table 1 – continued from previous page

Author	Algorithm and methodology	AUC-ROC	Dataset Size
Podder & Mondal (2020) [27]	Naive Bayes (NB)	0.64	5644 patients
	Extra Trees Classifier	0.91	
	Stacking 1 (RF, XGB, LR)	0.97	
	Stacking 2 (NB, LGBM, LR)	0.88	
	Voting 1 (Extra trees, RF, LGBM) Hard	0.99	
	Voting 1 (Extra trees, RF, LGBM) Soft	0.99	
	Voting 2 (MLP, NB, LGBM) Hard	0.85	
	Voting 2 (MLP, NB, LGBM) Soft	0.84	

Related works with models for prediction of ICU admission using different machine learning algorithms and strategies, AUC-ROC score, and size of the dataset.

3 Research Methodology

Public available data gathered from the National Institute of Health of Colombia[1] were used to train different machine learning models for the prediction of whether a patient is readmitted or not to ICU. The data with the information from all hospitals are stored and updated daily in two steps at the INS's website. First, the data corresponding to the patient's health situation on a given day k is saved in an individual file named according to that date. This file summarizes the patients' information, with their current status and locations. On the following day $(k+1)$, the file with the previous day's information is copied, and the new copy is named with the new date. The status and location of the patients are updated according to their current situation; for example, if a patient on day k was at the hospital, and the next day $(k+1)$ is moved to ICU, the status and location will change in the newly created file. Information regarding new patients is appended to this new file as well (See Fig 1). Therefore, the original dataset was a number of 210 files, corresponding to the number of days between 8 March 2020 to 14 October 2020. It is worth mentioning that patient data is available since March 14, 2020, however, the format and available features consigned in these files are different, so we don't include them in the present study.

The file's data are fetched from the web page of the INS and preprocessed to obtain a clean, unique dataset for training different machine learning models. We have used a random split of 70% for training and 30% for testing. Later, five-fold cross-validation with a random grid search was applied to optimize the parameters, considering a set of metrics for assessing the best model. The diagram in Fig 2 displays the general workflow used in this paper.

3.1 Data preprocessing

The data used in this paper was the set of historical files because not the last compiled. We used the data in this manner because in this way it is possible to track the change in status of each patient since it is added to the registry. The final dataset was composed of several CSV files with information from 18 March 2020 to 14 October 2020. It contains information from 930159 patients, but after a check of missing values and preprocessing, the final size was 843019 patients, where the final information for each one is presented in each row of data. The data preprocessing step included several

[1] Instituto Nacional de Salud (INS) (`https://www.ins.gov.co/`)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2022 doi:10.20944/preprints202212.0330.v1

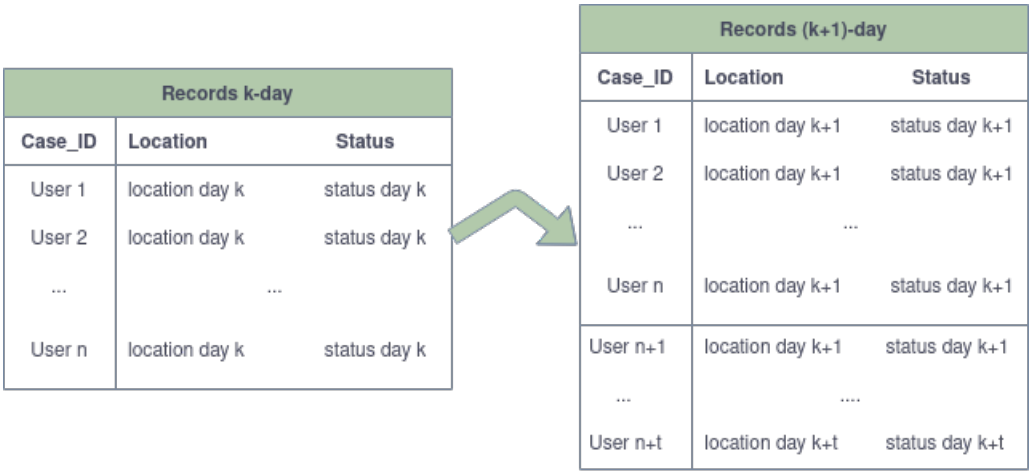

Fig 1. Scheme of updating data. General representation of the daily updating of the data in the public database. A new recorded dataset was generated every day upon updating the existing status of the patients and by appending the information about newly added patients.

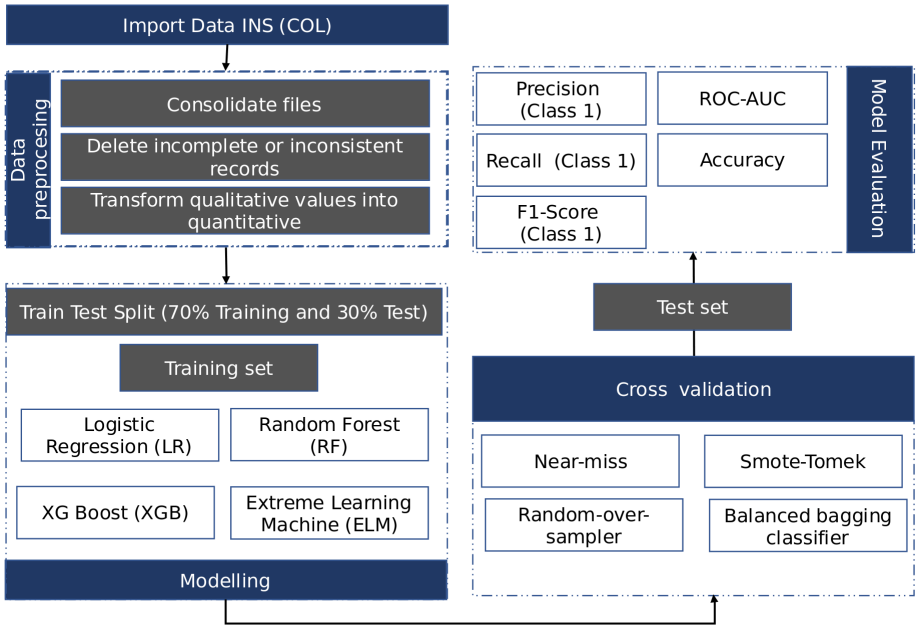

Fig 2. Machine learning pipeline. Diagram representing the complete steps performed in this work. First, the data is preprocessed and then divided for training and testing. Posteriorly, four different machine learning models in combination with four techniques for dealing with imbalanced data are adjusted, including a search of optimal parameters. The final performance of each model is evaluated using five different metrics.

operations, such as removing records without information or inconsistent values. The final dataset contains features such as the **case-ID** (*caso*), **notification date** (*fecha de notificación*), **department** (*departamento*)[2], **municipality** (*municipio*), **age** (*edad*),

[2]Political organization of Colombia.

 doi:10.20944/preprints202212.0330.v1

sex (*sexo*), **contagion source** (*fuente de contagio*), **symptom onset date** (*fecha de inicio de sintomas*), **date of death** (*fecha de muerte*), **date of diagnosis** (*fecha de diagnostico*), **date of recovery** (*fecha de recuperacion*), and last, several columns with information related to the **location** (*ubicación*) and the **status** (*estado*). These values might appear differently at each day following the changes in the condition of the patient. The feature **location** had values as recovered (*recuperado*), deceased (*fallecido*), hospital (*hospital*), hospital_uci (*hospital_uci*), *NaN* and home (*casa*), originally those values were different and there were additional values due to misspelling. For the feature **status**, a combination of similar terms was found; therefore we decided to join them into a single feature[3]. In general, bad spelling was found across several files, therefore manually constructed dictionaries were used to create a homogeneous dataset.

Given the foregoing, the dataset was preprocessed in every feature to obtain the count of days for every value in location and status, i.e., columns were introduced with numerical values informing the number of days in each condition then dictionaries were created to unify the values per feature. These features correspond to date values, then they were changed into the ordinal number of the day in the year. This allowed obtaining the number of days that a given patient was infected. The relative count was done with respect mainly to the notification date value. Symptom onset date, date of death, date of diagnosis, or date of recovery, were used (in that order) for some cases when the information about the notification date was not available. In other cases, the estimation of days that a person was infected, was obtained by taking the difference between the date of recovery or date of death with notification date or symptom onset date, or date of diagnosis.

Afterward, the preprocessing operations described above a final dataset was created, including a set of the original features, with other new features created with the values of location/status. These new features included the number of days a patient was at the house, at the hospital, and at the ICU. The final dataset contains four additional variables. The first one is a variable that includes the days a patient is infected but without clinical support (**days in-home care**), and the second one counts the days the patient is infected but with clinical support but no intensive care (**days in-hospital care**), the third one considers the days the patient is in ICU (**days in ICU**) and the last one is a feature indicating whether a patient was or not in the ICU (**ICU binary**), this is the feature to predict. Features such as municipality were omitted to owe to refer a geographical location that is covered by the department. Table 2 presents the features included in the original data along with the obtained via preprocessing. Groups 1, 2, 3, and 4 show the features that originally were in the data, group 5 show features obtained from features in groups 3 and 4, and finally, the feature in group 2 was not used due to this feature had only values: related, imported and in the study but, just a few cases were imported and most of them were related which means that most of the people were infected in contact with domestic people. Features in groups 1 and 5 were considered in the model.

3.2 Statistical description of the dataset

Foremost, class 0 represents people who have not been in ICU and class 1 people who have; hence there are 836123 people in class 0 which is 99.18%, and 6896 in class 1 which is 0.82% (See Table 3). Additionally, Fig. 3 shows the distribution of the quantitative features such as Age, Days in-home, Days in-hospital care, and Days in ICU, also the distribution of ICU cases in each department, which is to stand out as the general trend of the data.

[3]In Spanish, the term *estado* has two meanings: status and state. This was a source of confusion in the original dataset, and the reason why location and status appear as a combined feature.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2022 doi:10.20944/preprints202212.0330.v1

Table 2. Original and derived features of the dataset.

Group	Features	Type	Classes
1	**Age**	Quantitative	Integer (0-100)
	Department	Qualitative	36 Departments
	Sex	Qualitative	Male, Female
	Ethnic group	Qualitative	5 (Unknown, Indigenous, Black, Romani, and other)
	Asymptomatic	Qualitative	Yes, No
2	**Contagion sources**	Qualitative	Imported, Related, In study
3	**Symptom onset date**	Qualitative	Date (DD/MM/YY)
	Date of death	Qualitative	Date (DD/MM/YY)
	Date of diagnosis	Qualitative	Date (DD/MM/YY)
	Date of recovery	Qualitative	Date (DD/MM/YY)
4	**Location / Status**	Qualitative	Recovered, Deceased Hospital, Hospital_UCI, Home, NaN
5	**Days in-Home care**	Quantitative	Integer (0-101)
	Days in-Hospital care	Quantitative	Integer (0-126)
	Days in ICU	Objective	Integer (0-142)
	ICU Binary	Qualitative	Yes, No

The information shows the original characteristics of the data together with features obtained from those.

Table 3. Description of the dataset used in this work.

Parameter	Age	Days in-home	Days in-hospital care	Days in ICU
Count	843019 (**6896**)	843019 (**6896**)	843019 (**6896**)	843019 (**6896**)
Mean	39.5 (**60.19**)	25.1 (**14.8**)	1.07 (**5.85**)	0.12 (**14.43**)
STD	18 (**17.25**)	7.78 (**10.72**)	6.51 (**10.13**)	2.13 (**18.72**)
Min	0 (**0**)	0 (**0**)	0 (**0**)	0 (**1**)
25%	27 (**51**)	21 (**8**)	0 (**0**)	0 (**2**)
50%	37 (**62**)	25 (**13**)	0 (**0**)	0 (**8**)
75%	51 (**72**)	29 (**19**)	0 (**9**)	0 (**18**)
Max	115 (**100**)	211 (**101**)	149 (**126**)	142 (**142**)

Statistical description of the dataset used in this work. In parentheses and **bold**, the statistical parameters of the data corresponding to patients who required ICU.

Table 3 shows some descriptive parameters of the quantitative features of the dataset, though despite that, the feature days in ICU are considered among the descriptive analysis, it was not used to train the models. Moreover, the evident imbalance in data suggests that additional strategies should be considered in the training of the machine learning models. Table 3 also shows the same descriptive parameters just for patients who required ICU, in parentheses and bold.

The chart **(a)** in Fig 3 displays the distribution per age of infected people, considering the distribution of those who survive, and those who did not, furthermore, inserted at the top-left of **(a)** the chart shows a constant increase in the percentage of the mortality rate of people as a function of the age. The chart **(b)** in Fig. 3 shows the distributions of the quantitative features Days in-home (top-left-side) i.e., the distribution of days that people infected have been treated in the home, Days in-hospital care (top-right-side) i.e., the distribution of days that people were in hospital care and Days in ICU, i.e., the distribution of days that people required ICU.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2022 doi:10.20944/preprints202212.0330.v1

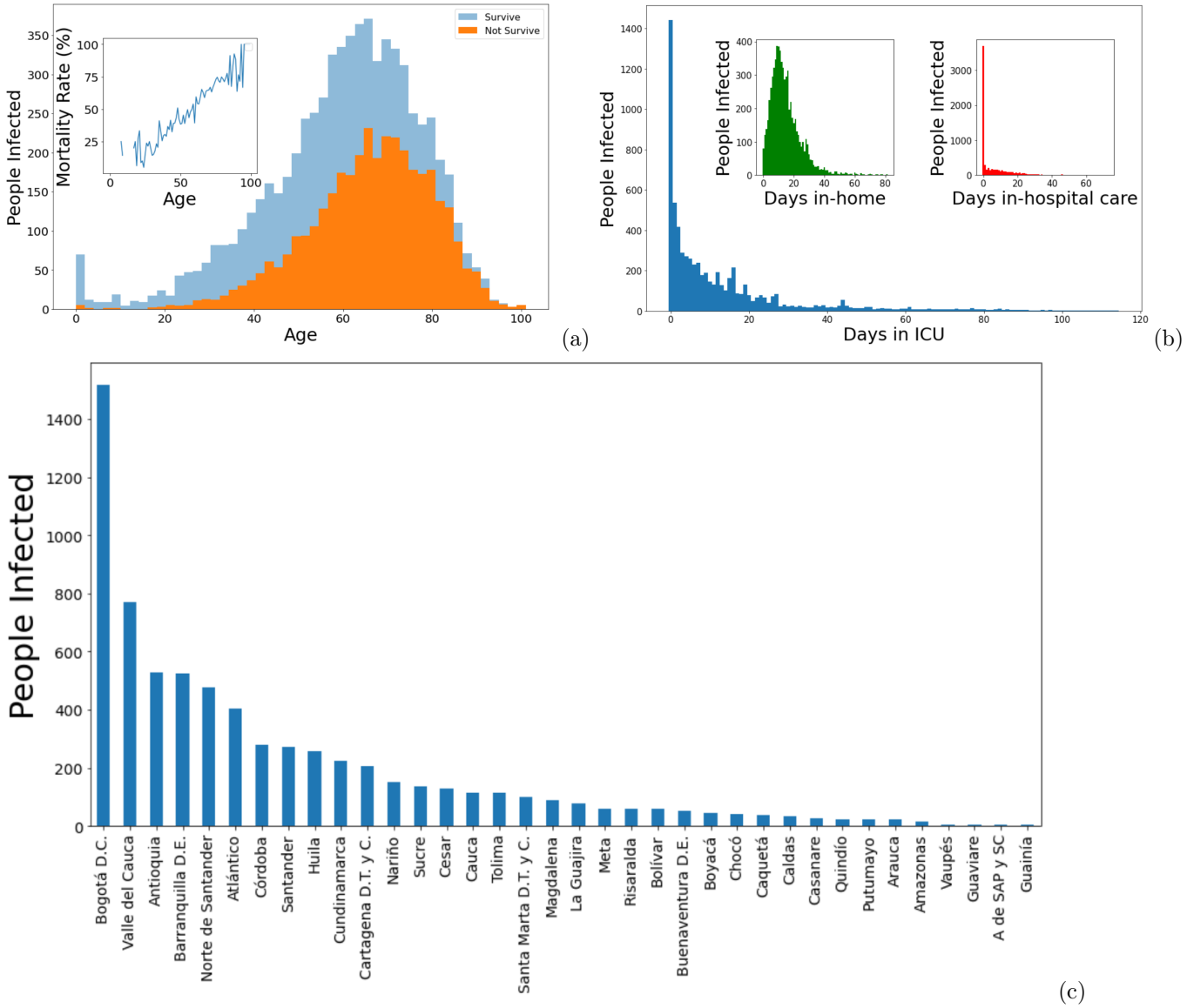

Fig 3. Distribution of people infected by quantitative variables and department (a) Upper-Left. The increasing mortality rate by age. Main chart. The distribution of people infected by age, considering survivors (blue) and not survivors (orange). (b) Upper-Left. Distribution of people infected according to the recovery days in-home. Upper-Right. Distribution of people infected according to recovery days in-hospital care. Main chart. Distribution of people infected according to recovery days in ICU. (c) Distribution of people infected by Departments, considering Departments as an administrative unit in Colombia.

3.3 Modeling

The final dataset was used to train four different machine learning models: logistic regression, random forest, extreme learning machine (ELM), and extreme gradient boosting (XGBoost). The aim was to predict whether a patient would be readmitted to ICU after being diagnosed positive for COVID-19. The data was highly imbalanced; therefore, several cutting-edge methodologies for dealing with imbalanced data were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2022 doi:10.20944/preprints202212.0330.v1

Table 4. Compilation of the best obtained results.

Score	Logistic Regression[1]	Random Forest[2]	XGBoost[3]	ELM[4]
AUC-ROC	0.88	0.94	0.94	0.85
Precision (Class 1)	0.04	0.08	0.07	0.07
Recall (Class 1)	0.81	0.84	0.85	0.79
f1-score (Class 1)	0.08	0.14	0.14	0.12
Accuracy	0.85	0.92	0.91	0.91

[1] Logistic Regression cross validation with parameter *class_ weight* in balanced.
[2] Random Forest with the parameter *class_ weight* equal to balanced.
[3] Extreme Gradient Boosting with a parameter *scale_ pos_ weight* equal to 100.
[4] Extreme Learning Machine with random oversampler as a balancing class method.

Summary of the best AUC scores for the different machine learning methods and imbalance techniques.

tested for each trained model: near-miss, random-over-sampler, Smote-Tomek, and balanced bagging classifier. In some cases, built-in options included in some classifiers for handling imbalanced data were tested. 70% of the data were used for training and 30% for testing. All calculations were performed with five-fold cross-validation and randomized grid-search using the python libraries *scikit-learn* [33] and *scikit-elm* [34] for the models, and the *imbalanced-learn* library for handling the imbalanced data [35]. To check the performance of all models, the AUC-ROC is used.

The first method, and baseline for our results, was a logistic regression model with cross-validation considering the hyperparameter *class_ weight* as *balanced*. The second technique used was the random forest model with the *class_ weight* parameter set to *balanced*. In the third technique, XGBoost models were trained, including the hyperparameter *scale_ pos_ weight*. The ELM model was trained with only the four imbalance techniques.

The main performance metric observed was AUC-ROC, but measures such as precision, recall, and even accuracy were necessary due to in some cases despite the AUC-ROC, the accuracy could display a good performance, on the other side, a measure as recall could be indicating not a good performance in the minority class.

Logistic regression was tested in different settings, five-fold cross-validation taking class weight balanced to cover the imbalance in data, and near-miss, random over sampler, and balance bagging classifier were also tested. The trained models produce similar results regarding performance. Although the technique near-miss produced the worst performance compared to the rest, its value in recall was similar. In A logistic regression model without considering the imbalance, the accuracy is 0.99, but the recall in class 1 is 0.

The random forest models were tested with randomized search and five-fold cross-validation using the hyperparameter class weight balanced, whose performance was equivalent to that of logistic regression models. Therefore, models with an equivalent recall were chosen. Thirdly, the extreme boosting models also were tested with randomized search cross-validation considering the hyperparameter *scale_ pos_ weight* due to the imbalance in data.

Results and Discussion

Afterward, the best were chosen based on the best performance measures to run the models using the strategies for imbalance classes, even those embedded into them. The logistic regression, random forest, and extreme gradient boosting models ran with 50 iterations, and the extreme learning models with 100 iterations. However, both random forest and XGBoost show better performance than logistic regression and extreme

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2022 doi:10.20944/preprints202212.0330.v1

learning machines; even so, the performing measures of these are according to state-of-the-art and the input variables proposed.

Fig 4. Performance metrics. Summary of the performance metrics obtained for the best machine learning algorithms after testing different balancing methods.

Table 4 sums up the results obtained, showing general measures as AUC-ROC and Accuracy, and measures for each class as precision, recall, and f1-score. In contrast, the general measures show a good performance, but the values by class do not display a good prediction, as [36] said, in an imbalanced dataset, precision-recall is more informative than the AUC-ROC, even so, contributes to getting a general idea about of the performance of the models. However, these are still a better indicator insomuch as

 doi:10.20944/preprints202212.0330.v1

looking at the best model to predict people admitted in ICUs.

One of the relevant problems identified in the performance indicators is the low value for precision; according to this, there are many false positives, that is, patients who were not admitted to ICUs as admitted patients, this result overestimates the number of people who will need intensive care equipment and healthcare personal which could be costly in economic terms and additionally creating panic scenarios. The above would depict a logistic situation that is difficult to handle due to any healthcare system would not be equipped to care for this proportion, despite that, the recall value provides a good performance predicting people as they need an ICU, and they need it, the true positives, this will assure that most people who need an ICU will have, but, as we saw, with an over-estimating the number of cases.

The results previously exposed set some concerns about the performance of the proposed models, however, the AUC-ROC values in comparative terms in [37] and [18] whose AUC-ROC was 0.98 in both cases and the best AUC-ROC proposed in this paper is 0.94 for Random Forest and XGBoost, where the AUC-ROC value appears as the most crucial indicator even whether the dataset is imbalanced, although as [38] pointed out recall and precision are essential in imbalanced datasets, nevertheless, it insists that AUC-ROC as the most important metric, the performance of the models is similarly despite [37] considers variables as age, Body Mass Index (BMI), sex, smoker/non-smoker status and diagnoses, temporal data included time and date of COVID-19 PCR tests, hospital, and ICU admissions (due to COVID-19), use of mechanical ventilation, medications ministered, laboratory tests performed and vital signs, which is a considerable difference regarding the information used in this work, even considering the research in [32] where 909 variables were considering, and the AUC-ROC reported was 0.95, thus most of the studies reported AUC-ROC values in the same scale, but there were a generalized use of information often not public as medical information recorded daily in hospitals, information from medical record and information about comorbidities of patients.

With the information previously exposed, the role of the AUC-ROC metric is understandable as regards the performance of the model, but as it stands out in [38] sensitivity and specificity contribute to improving the performance per class since in practice, these could imply resources, thus, whether measures as recall and precision enhance, the overall performance increase and with these the reliability of the model. Finally, it is important to notice that the use of open datasets presents a challenge concerning obtaining the best information possible such as in the preprocessing in the development of the models, for instance, the possibility to obtain the days that a patient's stay in the home, hospital care or ICU, suggests that even the model could be updated daily to obtain a new prediction which in practice would allow making decisions on time.

Conclusions

In this paper, we presented the results on the modeling of Intensive Care Units (UCIs) occupancy of COVID-19 patients using machine learning methods trained with open data. Our findings showed that openly available data without compromising personal information could be used to make predictions with high scores.

Preprocessing is an integral part of data analysis, insomuch as, being openly accessible data, the information contained is limited, which leads to designing strategies to extract implicit information. In this case, identifying failures in capturing data suggests cleaning operations and minimizing errors.

As highlighted, the datasets used in the different proposals in state-of-the-art consider datasets with a set of medical variables that are usually not in open datasets.

 doi:10.20944/preprints202212.0330.v1

Therefore, a prior would be suitable to think that the more medical variables are considered, the better the performance measures are obtained. Although it seems evident, the results obtained with some demographic variables and how long the patients were during the infection provide good results. Further studies could include data from the other COVID-19 waves, as well as data from other countries for comparison.

Considering the current situation of the pandemic, especially the vaccination, that in itself modified the virus behavior into the population, is not comparable with the results obtained in this study, insomuch as this was accomplished before the development of vaccines and this is not considering changes induced in the population dynamics.

Acknowledgments

Espinosa-Leal wishes to acknowledge CSC – IT Center for Science, Finland, for computational resources. Fajardo-Moreno wishes to acknowledge Microsoft AI For Good Research Lab, USA, for Azure ML Studio and data scientist resources for this project.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases. 2020;20(5):533–534.
2. Wu CJ, Brooks D, Chen K, Chen D, Choudhury S, Dukhan M, et al. Machine learning at facebook: Understanding inference at the edge. In: 2019 IEEE international symposium on high performance computer architecture (HPCA). IEEE; 2019. p. 331–344.
3. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annual review of clinical psychology. 2017;13:23.
4. Espinosa-Leal L, Chapman A, Westerlund M. Autonomous Industrial Management via Reinforcement Learning. Journal of Intelligent & Fuzzy Systems. 2020;39(6):8427–8439.
5. 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.
6. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence. 2021;3(3):199–217.
7. von Borzyskowski I, Mateen B, Mazumder A, Wooldridge M. Data science and AI in the age of COVID-19; 2021. Available at `https://www.turing.ac.uk/sites/default/files/2021-06/data-science-and-ai-in-the-age-of-covid_full-report_2.pdf` (Accesed: 2021/08/06).
8. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nature medicine. 2020;26(8):1183–1192.

9. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M. Mapping the landscape of artificial intelligence applications against COVID-19. Journal of Artificial Intelligence Research. 2020;69:807–845.

10. Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nature communications. 2020;11(1):1–12.

11. Bonadia N, Carnicelli A, Piano A, Buonsenso D, Gilardi E, Kadhim C, et al. Lung Ultrasound Findings Are Associated with Mortality and Need for Intensive Care Admission in COVID-19 Patients Evaluated in the Emergency Department. Ultrasound in medicine & biology. 2020;46(11):2927–2937. doi:10.1016/j.ultrasmedbio.2020.07.005.

12. Kottlors J, Zopfs D, Fervers P, Bremm J, Abdullayev N, Maintz D, et al. Body composition on low dose chest CT is a significant predictor of poor clinical outcome in COVID-19 disease - A multicenter feasibility study. European journal of Radiology. 2020;132:109274–109274. doi:10.1016/j.ejrad.2020.109274.

13. Salles Neto LLd, Martins CB, Chaves AA, Konstantyner TCRdO, Yanasse HH, Campos CBLd, et al. Forecast UTI: aplicativo para previsão de leitos de unidades de terapia intensiva no contexto da pandemia de COVID-19. Epidemiologia e Serviços de Saúde. 2020;29.

14. Allenbach Y, Saadoun D, Maalouf G, Vieira M, Hellio A, Boddaert J, et al. Development of a multivariate prediction model of intensive care unit transfer or death: A French prospective cohort study of hospitalized COVID-19 patients. PLOS ONE. 2020;15(10):e0240711. doi:10.1371/journal.pone.0240711.

15. Colombi D, Bodini FC, Petrini M, Maffi G, Morelli N, Milanese G, et al. Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia. Radiology. 2020;296(2):E86–E96. doi:10.1148/radiol.2020201433.

16. Schalekamp S, Huisman M, van Dijk RA, Boomsma MF, Freire Jorge PJ, de Boer WS, et al. Model-based Prediction of Critical Illness in Hospitalized Patients with COVID-19. Radiology. 2020; p. 202723. doi:10.1148/radiol.2020202723.

17. Chao H, Fang X, Zhang J, Homayounieh F, Arru CD, Digumarthy SR, et al. Integrative Analysis for COVID-19 Patient Outcome Prediction. ArXiv. 2020; p. arXiv:2007.10416v1.

18. Schwab P, DuMont Schütte A, Dietz B, Bauer S. Clinical Predictive Models for COVID-19: Systematic Study. J Med Internet Res. 2020;22(10):e21439. doi:10.2196/21439.

19. Wollenstein-Betech S, Cassandras CG, Paschalidis IC. Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator. medRxiv : the preprint server for health sciences. 2020; p. 2020.05.03.20089813. doi:10.1101/2020.05.03.20089813.

20. Mejía-Vilet JM, Córdova-Sánchez BM, Fernández-Camargo DA, Méndez-Pérez RA, Morales-Buenrostro LE, Hernández-Gilsoul T. A risk score to predict admission to the intensive care unit in patients with COVID-19: the ABC-GOALS score. Salud Pública de México. 2020;doi:10.21149/11684.

21. Cohen JP, Morrison P, Dao L, Roth K, Duong TQ, Ghassemi M. COVID-19 image data collection: Prospective predictions are the future. arXiv preprint arXiv:200611988. 2020;.

22. Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. PLOS ONE. 2020;15(7):e0236618. doi:10.1371/journal.pone.0236618.

23. Vaid A, Somani S, Russak AJ, De Freitas JK, Chaudhry FF, Paranjpe I, et al. Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation. J Med Internet Res. 2020;22(11):e24018. doi:10.2196/24018.

24. Li X, Ge P, Zhu J, Li H, Graham J, Singer A, et al. Deep learning prediction of likelihood of ICU admission and mortality in COVID-19 patients using clinical variables. PeerJ. 2020;8:e10337.

25. Cheng FY, Joshi H, Tandon P, Freeman R, Reich DL, Mazumdar M, et al. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. Journal of clinical medicine. 2020;9(6):1668.

26. Dan T, Li Y, Zhu Z, Chen X, Quan W, Hu Y, et al. Machine Learning to Predict ICU Admission, ICU Mortality and Survivors' Length of Stay among COVID-19 Patients: Toward Optimal Allocation of ICU Resources. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2020. p. 555–561.

27. Podder P, Mondal MRH. Machine Learning to Predict COVID-19 and ICU Requirement. In: 2020 11th International Conference on Electrical and Computer Engineering (ICECE). IEEE; 2020. p. 483–486.

28. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. NPJ digital medicine. 2021;4(1):1–7.

29. Jamshidi E, Asgary A, Tavakoli N, Zali A, Esmaily H, Jamaldini SH, et al. Using Machine Learning to Predict Mortality for COVID-19 Patients on Day Zero in the ICU. medRxiv. 2021;.

30. Lorenzen SS, Nielsen M, Jimenez-Solem E, Petersen TS, Perner A, Thorsen-Meyer HC, et al. Developing machine learning models for predicting intensive care unit resource use during the COVID-19 pandemic. medRxiv. 2021;.

31. Heo J, Han D, Kim HJ, Kim D, Lee YK, Lim D, et al. Prediction of patients requiring intensive care for COVID-19: development and validation of an integer-based score using data from Centers for Disease Control and Prevention of South Korea. Journal of intensive care. 2021;9(1):1–9.

32. Dan T, Li Y, Zhu Z, Chen X, Quan W, Hu Y, et al. Machine Learning to Predict ICU Admission, ICU Mortality and Survivors' Length of Stay among COVID-19 Patients: Toward Optimal Allocation of ICU Resources. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2020. p. 555–561.

33. 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;12:2825–2830.

 doi:10.20944/preprints202212.0330.v1

34. Akusok A, Leal LE, Björk KM, Lendasse A. Scikit-ELM: an extreme learning machine toolbox for dynamic and scalable learning. In: International Conference on Extreme Learning Machine. Springer; 2019. p. 69–78.

35. Lemaître 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(17):1–5.

36. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):1–21. doi:10.1371/journal.pone.0118432.

37. Lorenzen SS, Nielsen M, Jimenez-Solem E, Petersen TS, Perner A, Thorsen-Meyer HC, et al. Developing machine learning models for predicting intensive care unit resource use during the COVID-19 pandemic. medRxiv. 2021;.

38. Podder P, Khamparia A, Mondal M, Rahman MA, Bharati S. Forecasting the Spread of COVID-19 and ICU Requirements. International Journal of Online & Biomedical Engineering. 2021;17(5).