Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Comprehensive analysis of clinical data for COVID-19 outcome estimation with machine learning models

Daniel I. Morís [a,b], Joaquim de Moura [a,b,*], Pedro J. Marcos [c], Enrique Míguez Rey [d], Jorge Novo [a,b], Marcos Ortega [a,b]

[a] Centro de Investigación CITIC, Universidade da Coruña, Campus de Elviña, s/n, 15071 A Coruña, Spain
[b] Grupo VARPA, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, Xubias de Arriba, 84, 15006 A Coruña, Spain
[c] Dirección Asistencial y Servicio de Neumología, Complejo Hospitalario Universitario de A Coruña (CHUAC), Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, Sergas, 15006 A Coruña, Spain
[d] Grupo de Investigación en Virología Clínica, Sección de Enfermedades Infecciosas, Servicio de Medicina Interna, Instituto de Investigación Biomédica de A Coruña (INIBIC), Área Sanitaria A Coruña y CEE (ASCC), SERGAS, 15006 A Coruña, Spain

ARTICLE INFO

Keywords:
COVID-19
Machine learning
Clinical data
Feature selection
Classification

ABSTRACT

COVID-19 is a global threat for the healthcare systems due to the rapid spread of the pathogen that causes it. In such situation, the clinicians must take important decisions, in an environment where medical resources can be insufficient. In this task, the computer-aided diagnosis systems can be very useful not only in the task of supporting the clinical decisions but also to perform relevant analyses, allowing them to understand better the disease and the factors that can identify the high risk patients. For those purposes, in this work, we use several machine learning algorithms to estimate the outcome of COVID-19 patients given their clinical information. Particularly, we perform 2 different studies: the first one estimates whether the patient is at low or at high risk of death whereas the second estimates if the patient needs hospitalization or not. The results of the analyses of this work show the most relevant features for each studied scenario, as well as the classification performance of the considered machine learning models. In particular, the XGBoost algorithm is able to estimate the need for hospitalization of a patient with an AUC-ROC of 0.8415 ± 0.0217 while it can also estimate the risk of death with an AUC-ROC of 0.7992 ± 0.0104. Results have demonstrated the great potential of the proposal to determine those patients that need a greater amount of medical resources for being at a higher risk. This provides the healthcare services with a tool to better manage their resources.

1. Introduction

The COVID-19 is an infectious disease declared as global pandemic by the World Health Organization (WHO) in 11th March 2020 [1]. Due to its rapid spread, this disease has emerged as a challenge for the healthcare systems worldwide. This challenge is even greater during the most critical moments of the pandemic, as the great hospital and ICU admission rates can lead to a lack of medical resources surpassing the capacity of the health systems [2]. When this occurs, the clinicians must take important decisions, giving more attention to those patients that seem to need more personalized clinical care and monitoring. The COVID-19 is characterized for being usually more threatening to those patients that have previous pathological conditions, as well as for those patients of a higher age [3]. In the same way, it tends to be more dangerous for those patients that have an immunosuppression condition, as they present a reduced capacity to attack the pathogen. Therefore, this information is crucial for the clinicians to pay special attention to those who have a greater risk of experiencing a severe form of the disease and/or that present a greater risk of death.

When a patient with COVID-19 needs medical attendance, the healthcare workers must have the clinical information of each patient to determine those cases that need more attention. The clinical data should include previous relevant conditions of the person (such as previous diseases) as well as other parameters that evaluate the state of the patient at a particular moment (such as measurements of a blood test) [4]. In the same way, healthcare workers usually register the outcome of the patient and some descriptive information of their stay at the medical center indicating the death or survival of the patient, the time spent at the hospital if necessary, the need for mechanical ventilation or the stay at ICU for the most severe cases. This allows the

* Corresponding author at: Centro de Investigación CITIC, Universidade da Coruña, Campus de Elviña, s/n, 15071 A Coruña, Spain.
E-mail addresses: daniel.iglesias.moris@udc.es (D.I. Morís), joaquim.demoura@udc.es (J. de Moura), pedro.jorge.marcos.rodriguez@sergas.es (P.J. Marcos), enrique.miguez.rey@sergas.es (E.M. Rey), jnovo@udc.es (J. Novo), mortega@udc.es (M. Ortega).

https://doi.org/10.1016/j.bspc.2023.104818
Received 26 June 2022; Received in revised form 22 November 2022; Accepted 5 March 2023
Available online 9 March 2023

healthcare systems to gather significant amounts of these clinical data, that can be used for retrospective studies. Thus, these datasets can be used to develop Computer-Aided Diagnosis (CAD) systems [5], in order to help the healthcare workers to take decisions on their daily practice, an aspect that is especially helpful during the most critical moments of the pandemic [6–9]. Furthermore, these retrospective studies are useful for clinicians to have a better understanding of the disease and the way the different variables worsen or improve the expected evolution of each patient.

During the last years, the machine learning strategies have emerged as powerful techniques to solve problems of classification and regression with large amounts of data, thanks to their great capability to find patterns and infer the most probable output [10]. However, most of the machine learning models act as black boxes, meaning that they are unable to explain the followed rules to obtain the output given a certain input information. This hinders the process of carrying the algorithmic solution to the daily clinical practice and makes it impossible to understand the weight that each feature impact has in the output. In the same way, there is a possibility that some features are redundant and even useless, aspects that could lead the model to lose effectiveness. These situations are more complex when the number of features is relatively high, making it impossible to perform a manual analysis. In this scope, alongside the machine learning approaches, many algorithms of feature selection have been proposed during the last years [11]. These algorithms were created with the aim of analyzing the importance of each feature and to reduce the dimensionality of the original dataset, removing the useless and the less important features. This aspect also allows reducing the computational requirements in terms of time and resources, making the problem more affordable.

The estimation of the outcome of the patients is a critical task that has been assessed in many scenarios. Among all of them, it is remarkable that some works have proposed the estimation of the outcome for patients that suffer from heart conditions, such as heart failure and stroke, as reference, the work from Tripoliti et al. [12] and the work from Asadi et al. [13]. On the other hand, the work from Alizadeh et al. [14] proposes the application of machine learning models to predict the outcome of neurosurgery procedures, using representative variables of the patient, such as age, gender or comorbidities among others. Nascimben et al. [15] use gene expression data from patients with bladder cancer to estimate the tumor stage and the risk of death, applying machine learning models and bio-statistical techniques. Finally, in the work of Yuan et al. [16], the authors use machine learning to estimate the risk of survival/death in patients with lung cancer using a longitudinal cohort of patients given their corresponding electronic health records.

In the case of the COVID-19 disease, its study is very relevant under the current situation and, therefore, we can find several contributions that work with clinical variables to estimate the evolution of the patients. As reference, the work from Aktar et al. [17] proposes the application of different machine learning algorithms to estimate if hospitalized patients require intensive care or if they can be shifted to a normal ward using relevant clinical data provided by blood tests. The work from Ali et al. [18] develops a methodology that applies several machine learning algorithms in 2 different COVID-19 datasets. The first one is composed of attributes that can be obtained without performing any kind of laboratory analysis such as the age group of the patient, the gender, the race, the ethnicity, potential symptoms, stay in hospital or stay in ICU, among others. The second one is a clinical dataset that is composed of relevant markers obtained after a blood test such as red blood cells, hemoglobin, leukocytes, or potassium, among many others. The aim of training the models using the first dataset is to estimate the most probable outcome of a patient: survival or death. In the case of the second dataset, the purpose is to classify patients as COVID-19 positive or negative using the provided clinical data. Regarding the contribution of Domínguez-Olmedo et al. [19], the authors propose the use of machine learning to estimate the risk of death of patients given their clinical data provided by the laboratory tests. Magunia et al. [20] perform a study that uses several machine learning algorithms to estimate the outcome of hospitalized patients that need to be admitted to the ICU. This analysis also allows the authors to understand which are the most relevant features to estimate that outcome. Hernández-Pereira et al. [21] propose the application of several machine learning algorithms and feature selection methods to estimate the risk that a patient has of requiring hospitalization or ICU admission.

In addition to all the previous reference works, we can also mention the methodology proposed by Xu et al. [22], that predicts in-hospital mortality and other relevant outcomes like ICU transfer using different machine learning algorithms (Least absolute shrinkage and selection operator (LASSO), Elastic net logistic regression (LR) and eXtreme Gradient Boosting (XGBoost)). On the other hand, the contribution of Polilli et al. [23] uses logistic regression and Cox modeling to predict the risk of hospitalization, death, the need for oxygen support and the need for intensive therapy (the last 2 in patients younger than 70 years old). In the case of Pishgar et al. [24], the authors include process mining in their methodology to exploit available time information. Van der Velde et al. [25] present a model called APOP (Acute Presenting Older Patient) that computes the in-hospital mortality risk of older people. The data include comorbidities and other important characteristics of the patients that could put them on a higher risk of death. It is also remarkable the work from Bendavid et al. [26], that proposes the use of the XGBoost algorithm to predict respiratory failure and invasive mechanical ventilation in patients with COVID-19 based on clinical features.

Nevertheless, diagnosing and estimating the outcome of a patient can also be done using other types of data (such as imaging data). Given that the main affectation of the COVID-19 is located in the lungs, chest imaging modalities are very useful in this scope. As reference, Chamberlin et al. [27] propose a deep learning model to automatically diagnose and prognosticate the evolution of COVID-19 using chest X-ray images. The work from Chaudhary et al. [28] proposes the use of the Fourier-Bessel series expansion-based decomposition method to extract features from chest X-ray and CT images that feed a classifier trained to perform an automatic diagnosis of COVID-19. Moreover, the work from Bermejo-Peláez et al. [29] segments the extension of different lesion subtypes in CT images. Then, this information is used to predict the outcome of COVID-19 patients. De Moura et al. [30,31] study the separability among 3 different classes (control cases, patients with pulmonary pathologies others than COVID-19 and COVID-19) using datasets of images captured by fixed and portable X-ray devices. To this end, authors consider 6 different network architectures, with 2 types of DenseNet, 2 types of ResNet and 2 types of VGG. Finally, it is also remarkable that some works combine the features extracted from images with clinical features. In particular, as reference, the methodology proposed by Sinha et al. [32] merges features extracted from CT images with clinical parameters to prognosticate the need for invasive mechanical ventilation using machine learning algorithms.

These works have some interesting results for the health community as they provide statistical analyses of the available datasets, exhaustive experiments with several machine learning algorithms and metrics as well as analyses of the importance of each feature. However, these works lack of an exhaustive analysis of how the amount of used features impact the performance. Given the significant gap in the literature, in this work, we propose the use of several machine learning models (Support Vector Machine, Decision Tree, XGBoost, k Nearest Neighbor and Multilayer Perceptron) to perform several retrospective studies on clinical data of COVID-19 patients. To do so, we use a dataset provided by the Complexo Hospitalario Universitario de A Coruña (CHUAC) that was specifically designed for the purposes of this study. We perform an exhaustive analysis of the available variables, using 3 relevant feature selection methods that provide the importance of each feature on the inference process: Fisher Scoring, Mutual Information

Table 1
Description of the basic information variables that exist in the dataset for each patient.

Variable	Description
Exitus	Indicates if the patient died or survived
Cohort	Non-Hospitalized or Hospitalized
Age Range	Older ages are related with higher risk
Age	The same as the previous one
Sex	Pathology affects differently males and females
Height	Height of the patient
Weight	Overweight and obesity mean with higher risk
BMI	Higher BMI is associated with higher weight

and a Variance-based Ranking. In the same way, we also performed several experiments to understand how the number of used features impact in the performance of the classification models. One of the most outstanding points of our contribution is that we have selected the 2 most critical scenarios in the case of a severe COVID-19 infection, with regard to hospitalization and the health evolution of the patients. Overall, the aim of this work is to study the impact of the different variables included in the dataset to make estimations in the 2 following scenarios:

- Given a patient that comes into emergencies, estimate the need for hospitalization.
- Given a patient that is admitted to the hospital, estimate the risk of death.

In summary, the analyses provided in this contribution are useful to determine the most relevant features for each studied scenario and the performance that the different machine learning models can achieve while addressing the considered problems. Under a clinical point of view, the results help to understand which are the most important features to estimate the outcome of a patient and therefore allowing to manage the resources more efficiently, prioritizing those cases of higher risk. In particular, the analyses show that age variables are notably relevant for both the first scenario and the second scenario, as well as the Arterial Hypertension and the Diabetes, 2 prevalent conditions in the studied datasets. Furthermore, indicators of renal activity are also important to determine the outcome of the patients. The main difference between scenarios is that, for the second one, the previous conditions related with cancer are much more relevant than in the first scenario. Regarding the novelty of the work, these are the most remarkable key points of our contribution:

- 2 critical scenarios are studied to estimate the outcome of COVID-19 patients, supported by an exhaustive experimentation that is performed using 5 different state-of-the-art machine learning models.
- 3 different feature selection methods are used to rank from the most important features to the least important. This ranking was then used to thoroughly study how the amount of features impacts the performance of the models.
- The most appropriate model is objectively selected for each analysis scenario. It is remarkable that a high performance is obtained despite the difficulty of the problem that is being proposed.
- The analyses and the developed systems presented in this contribution are useful for the clinicians to take decisions in the 2 critical scenarios that are described. These decisions support a better management of the health resources.

The manuscript is structured as follows: in Section 2, we describe the used dataset, while in Section 3 we describe the steps of the proposed methodology. Lately, in Section 4, we present the results obtained and their discussion after the experimental validation was performed. Finally, Section 5 summarizes the main conclusions that can be extracted after the development of the work.

Table 2
Description of the previous conditions that are present in the dataset for each patient.

Variable	Description
AHT	AHT is associated with a higher risk
Diabetes	Diabetes is associated with a higher risk
COPD	Patients with COPD suffer from breathing problems
Asthma	Asthma directly affects breathing capabilities
LD	It can affect certain organic mechanisms to fight the COVID-19
Leukemia	Patients with leukemia are often immunosupressed
Lymphoma	Patients with lymphoma are often immunosupressed
Neoplasm	Patients with neoplasm are often immunosupressed
HIV	Patients with HIV are often immunosupressed
Transplant	Patients with transplant are often immunosupressed
Chemotherapy	Patients under chemotherapy are often immunosupressed
Biological	Patients under biological treatments are often immunosupressed
CCS	Patients treated with CCS are often immunosupressed

Table 3
Description of the clinical variables that are available in the dataset for each patient.

Variable	Description
LYMP	A low count of lymphocytes can be indicative of an illness
LYMP (pct.)	The same as the previous one
D-Dimer Test	This variable is related with blood coagulation
LDH	This variable can be indicative of damage in certain body tissues
Creatinine	This variable can be indicative of kidney disorders
GFR	The same as the previous one
CRP	This variable presents abnormalities when the patient has an inflammation
Ferritin	This variable is related with the amount of blood iron
IL-6	The results of an IL-6 test are linked with immune response

2. Materials

In this section, we firstly describe the dataset that was used for the purposes of this work in Section 2.1. After that, we explain in detail the software and hardware resources necessary for the development of this work in Section 2.2.

2.1. COVID-19 CHUAC dataset

This dataset was specifically designed for the purposes of this work, and it was provided by the Complexo Hospitalario Universitario de A Coruña (CHUAC). It is important to remark that the selection of the most relevant and useful variables to characterize each patient to make the studies herein proposed was done in accordance with the Head of Infectious Diseases Department of the mentioned institution. In particular, the dataset is composed of 2067 unique patients where, for each patient, the dataset provides the variables Exitus, Cohort, Age Range, Age, Sex, Height, Weight and Body Mass Index (abbreviated as BMI) whose detailed description can be seen in Table 1.

Apart from that, the dataset also includes attributes that indicate if the patient has relevant previous conditions (*i.e.*, diseases or treatments that a patient can have). These conditions are Arterial Hypertension (abbreviated as AHT) [33], Diabetes Mellitus (abbreviated as Diabetes) [34], Chronic Obstructive Pulmonary Disease (abbreviated as COPD) [35], Asthma [36], Liver Disease (abbreviated as LD) [37], 3 cancer conditions as is the case of Leukemia, Lymphoma and Neoplasm [38], Human Immunodeficiency Disease (abbreviated as HIV) [39], Solid Organ Transplant (abbreviated as Transplant) [40], Chemotherapy within the last 3 months (abbreviated as Chemotherapy) [41], Biological Treatment within the last 3 months (abbreviated as Biological) [42] and Corticosteroids (abbreviated as CCS) [43]. Each condition is more deeply explained in Table 2.

Finally, there are some other relevant clinical measurements, as is the case of the Count of Lymphocytes (abbreviated as LYMP) and Percentage of Lymphocytes (abbreviated as LYMP (pct.)) [44], D-Dimer

Table 4
List of the libraries used in this work, with their version and their description.

Name	Version	Description
matplotlib	3.5.1	Library used for graphical visualizations of the data
numpy	1.21.4	Library that allows to create and work with arrays in Python
pandas	1.3.5	Python library for data analysis
scikit-learn	1.0.1	Library that allows to create machine learning models
xgboost	1.5.2	Python implementation of the XGBoost algorithm

Table 5
Description in detail of the hardware resources that were used for the purposes of this work.

Name	Description
OS	Ubuntu 20.04.3 LTS (Focal Fossa)
Kernel	Linux 5.13.0-41-generic
Architecture	x86-64
CPU	11th Gen Intel(R) Core(TM) i7-11700K @ 3.60 GHz
Motherboard	Gigabyte Z590 AORUS ELITE
RAM	2 x 32GiB DIMM DDR4 Synchronous 3200 MT/s CRUCIAL BL32G32C16U4BL.M16FB
SSD	Western Digital WDS100T1X0E-00AFY (1 TB)
HDD	Seagate IronWolf ST4000VN008-2DR16 (4 TB)

Test [45], Lactate Dehydrogenase (abbreviated as LDH) [46], Creatinine and Glomerular Filtration Rate (abbreviated as GFR) [47], C-Reactive Protein test (abbreviated as CRP) [48], Ferritin [49] and IL-6 protein test (abbreviated as IL-6) [50]. Each of these variables is detailed in Table 3.

2.2. Software and hardware resources

The implementation that was made for the purposes of this work has been done in Python 3 (version 3.8.10) using several libraries that are detailed in Table 4. Regarding the hardware resources used for the development of the work, the particular specifications can be seen in Table 5.

3. Methodology

In this work, we propose 2 different analysis scenarios that are more deeply explained in Section 3.1: the estimation of the risk of hospitalization and the estimation of the risk of death for a given patient. For that aim, we have developed an exhaustive analysis procedure that is common for both scenarios that, however, is able to perform each independent estimation working with different versions of the dataset and extracting the correspondent conclusions for each case. In particular, this methodology, whose pipeline can be graphically seen in Fig. 1, is divided in several steps, where each step is detailed in a different subsection. Specifically, the methodology starts with a data processing (detailed in Section 3.2) to ensure that the quality of the input dataset is satisfactory. This is followed by a feature selection process (detailed in Section 3.3) that chooses the best set of features to characterize the problem, reducing its dimensionality and allowing to understand which are the most important features. Finally, the dataset is used to train the classification models (a process detailed in Section 3.4), providing the validation results of the work. To finish this section, we also briefly describe the metrics that were used to validate our methodological proposal.

3.1. Description of the analyses

In this work, we conduct 2 analyses with different versions of the original COVID-19 CHUAC dataset. The description of each experiment is detailed below.

Analysis I. Estimation of Non-Hospitalized/Hospitalized patients. In this first scenario, we build a version of the dataset using *Cohort* as the output of the classification model. For this particular problem, we consider Non-Hospitalized as the negative class and Hospitalized as the positive class. In this version of the dataset, we included the patients of the Non-Hospitalized cohort and the Hospitalized cohort, having a total of 2067 unique patients. This analysis exposes which are the most important features to estimate the risk of hospitalization for a given patient, an aspect that can help the expert clinicians to make a more efficient management of the available health resources.

Analysis II. Estimation of Survival/Death. In this second scenario, we build a different version of the dataset, where the variable that indicates the death or the survival of the patient is chosen as the output of the classification model. For this version of the dataset, we only included those patients of the Hospitalized cohort, thus having a total of 1783 unique patients. In this particular problem, Survival is considered as the negative class and Death as the positive class. This analysis can help both to understand which are the most relevant features to determine the death risk of a patient and to focus more on those cases that need a greater monitoring and health care.

3.2. Data processing

Given that the dataset was built in a real clinical scenario context, it is necessary to deal with common issues, such as the missing values. Considering that the dataset is composed of both discrete and numerical variables, the way of dealing with missing values must be different for each case. On one hand, with regard to discrete variables, we filled the missing values with a zero-padding. On the other hand, it is slightly different in the case of the numeric variables. In the latter context, a 0 could refer to a meaningless situation (as, for example, it is impossible for a patient to have 0 leukocytes). Thus, we decided to fill these cells applying a padding with −1.

Moreover, in this step, we also needed to analyze the format of the variables, to know in which manner they should be converted. Therefore, for that aim, it was necessary to ensure that all the numerical variables had the same format to be converted to either decimal or integer. In the same way, it was also necessary to make sure that all the discrete variables had the proper format to be converted to a binary format. This refers to the precondition variables, as the dataset specifies if the patient has a given precondition or not.

3.3. Feature selection

With the aim of reducing the dimensionality of the original dataset and analyzing the impact of each future on the performance of the models, we use 3 different methods of feature selection. The use of several methods aims at offering a greater perspective of which are the most relevant features to estimate the outcome on each scenario. It is important to remark that the feature selection algorithms are actually based on statistical measures to score the correlation or dependence between input variables that can be filtered to choose the most relevant features. Generally speaking, each method will give a score to each feature that will be used to build a ranking where the most important features will be placed at the top positions while the less important features will be placed at the bottom positions. In particular, we use the method of Fisher Scoring [51] and Mutual Information [52], that were previously used in other similar clinical problems with satisfactory results. Moreover, we also use a ranking based on the variance of each feature (namely, Variance-based Ranking) to include a more exhaustive analysis of the feature selection process.

Overall, the 3 chosen feature selection methods can be divided in 2 main categories. Fisher Scoring and Mutual Information can be described as 2 methods that consider the correlation between each individual feature and the target output, while the Variance-based Ranking method only focuses on variability regarding each individual feature without taking into account its correlation against the target output. Thus, it is important to remark that Variance-based Ranking tends to

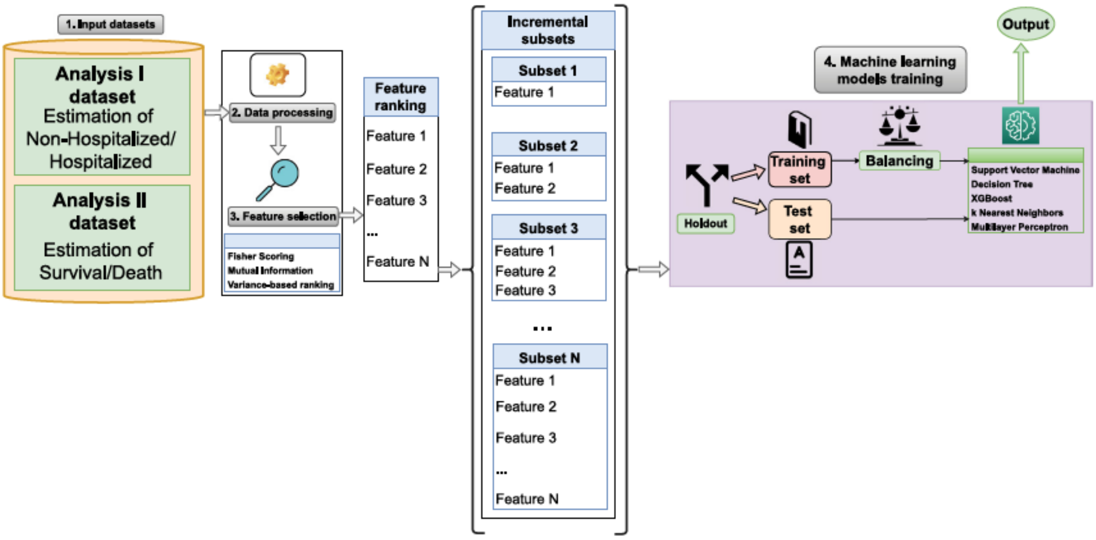


Fig. 1. Overall description of the pipeline of the proposed methodology.

give higher scores to numerical variables over discrete variables, as the numerical variables usually have a greater variability. Each of the selected methods is described more deeply below:

Fisher Scoring: this statistical method can be easily applied to classification problems. Given the values of a feature f and a set of m classes denoted as $C = \{c_1, c_2, c_3, \dots, c_m\}$, the Fisher Score is computed using the mathematical expression stated in Eq. (1):

$$Score_f = \sum_{c \in C} n_c \frac{(\mu_{cf} - \mu_f)^2}{{\sigma_{cf}}^2} \tag{1}$$

where n_c refers to the number of samples per class, μ_{cf} refers to the mean value of the feature f considering only the samples that belong to the class c, μ_f refers to the overall mean value of the feature and σ_{cf} denotes the standard deviation of the feature given only the samples that belong to the class c.

Mutual Information: in this context, Mutual Information explains the uncertainty that exists in a classification problem given the information provided by a particular feature, as it is based in the concept of entropy. Therefore, the mutual information value will be higher with a stronger dependency between the feature and the target output (*i.e.*, reduction of uncertainty). In order to define the Mutual Information, firstly, we need to state the mathematical expression of entropy. Given a feature X and denoting its entropy as $H(X)$, this statistical measurement is calculated as can be seen in Eq. (2).

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{2}$$

On the other hand, we need to define the conditional entropy. Denoting Y as the target output, the conditional entropy is expressed as can be seen in Eq. (3):

$$H(X|Y) = -\sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(y_j)} \tag{3}$$

Finally, once defined these 2 concepts, the mutual information I between a feature X and a target output Y can be defined as is stated in Eq. (4):

$$I(X,Y) = H(X) - H(X|Y) \tag{4}$$

Variance-based Ranking: this approach assumes that the features with a greater variance provide a greater amount of information. Therefore, the score value given by this method is the variance $Var(X)$ of the feature X itself as stated by the mathematical expression that can be seen in Eq. (5):

$$Var(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu) \tag{5}$$

Once the ranking of features is built with the obtained scores and given a specific selection method, the features are fed to the machine learning models as the last step of the methodology.

3.4. Machine learning models training

For validation purposes, after the data is processed and the most important features are identified, the dataset is divided in a random Holdout fashion, having a 70% of the samples for training and the remaining 30% of the samples for testing. Moreover, to overcome the considerable imbalance of the 2 analysis scenarios that are proposed in this work, a random balancing is applied on the training set. This process is described in Fig. 2 and is performed as follows: given a majority class with N samples and a minority class with M samples, the difference is computed. Then, a random subset of $N - M$ samples is selected from the minority class. Finally, the obtained random subset is appended to the minority class. As result, the new generated dataset will contain the same amount of samples for both classes. In addition to this, another important detail is that, to have a better understanding of the behavior of the model, the training process is repeated 5 times, allowing to calculate the mean and the standard deviation values. Once the dataset is split, we use 5 different machine learning models that were also used in a previous state-of-the-art work that performs a similar task [19]: Support Vector Machine (SVM) [53], Decision Tree (DT) [54], XGBoost algorithm [55], k Nearest Neighbors (kNN) [56] and Multilayer Perceptron (MLP) [57]. Given the feature rankings that are obtained in the previous step of the methodology, we propose a forward training approach. This means that the model is trained with the most important feature of the ranking, then with the 2 most important features, then with the 3 most important features... until reaching the whole number of features. This procedure is necessary to understand the impact that the number of features makes on the performance of the classification models.

Regarding the setup of the classifiers, a range of parameters was empirically selected to find a high-performing combination for each of them. In the case of the SVM, several kernel functions were considered, in particular, linear, polynomial and radial basis function, being the latest the one that obtained the highest performance. In the case of

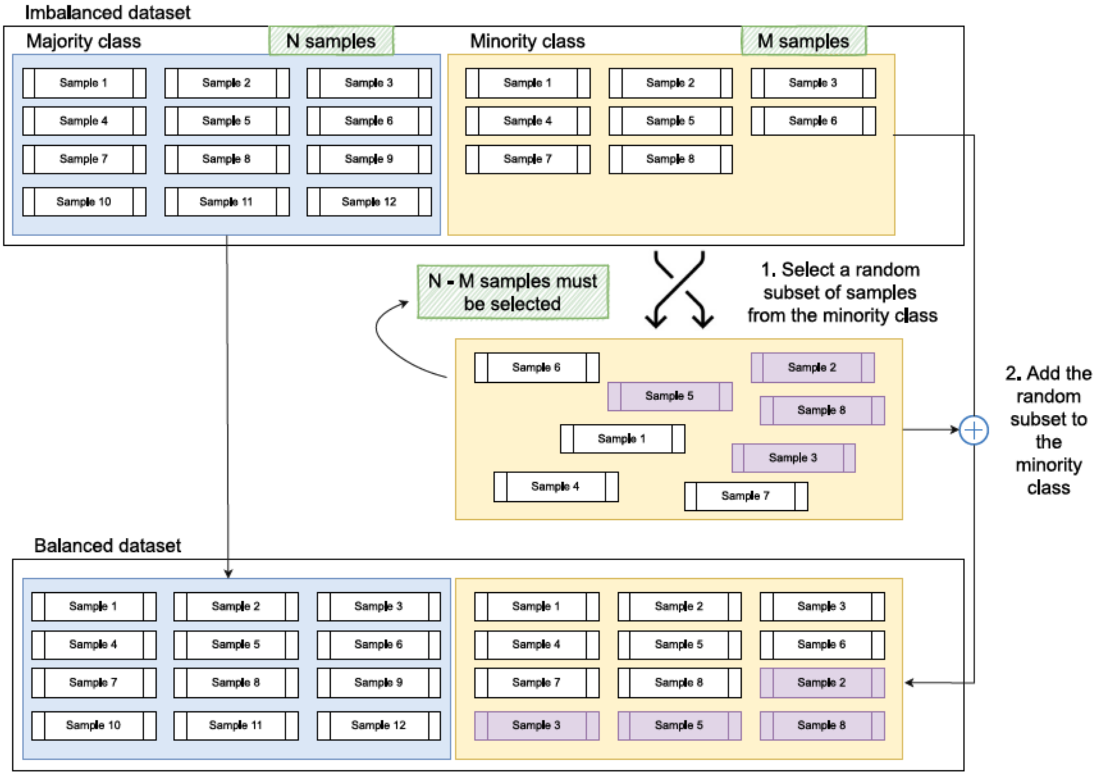


Fig. 2. Description of the method that was used to address the problem of imbalancing that exists in the original dataset.

the Decision Tree, different values of maximum depth were considered, being 3 the one that achieved the highest performance. Moreover, XGBoost was proven with different boosting strategies, retrieving the highest performance using the Dropouts meet Multiple Additive Regression Trees (DART) method [58] that was set up with a learning rate of 0.1. In the matter of kNN, different number of neighbors were taken into account, being 5 the specific chosen number. Finally, with regard to MLP, the maximum number of training iterations was set to 300, given that this was the highest-performing configuration. It is important to remark that the rest of parameters were set with the default values that are specified by the used software library.

3.5. *Evaluation metrics*

To evaluate the performance of the trained models to solve the 2 proposed tasks, we use some of the most common classification metrics that are used in the state-of-the-art. Denoting TP as true positives, TN as true negatives, FP as false positives and FN as false negatives, the sensitivity and specificity are defined as can be seen in Eqs. (6) and (7), respectively.

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

In the same way, we also provide the value of AUC-ROC. To define this metric, we first need to determine the expression of True Positive Rate (TPR) and False Positive Rate (FPR). The TPR is equivalent to the already defined expression of Sensitivity, while the FPR can be defined as 1-Specificity. Given those definitions of TPR and FPR, and denoting $d(FPR)$ as the derivative of the False Positive Rate, the expression of the AUC-ROC can be seen in Eq. (8).

$$AUC - ROC = \int_0^1 TPR \cdot d(FPR) \tag{8}$$

Table 6
Analysis of the distribution of 3 relevant features in the dataset built for the experiment I (estimation of Non-Hospitalization/Hospitalization), given by the absolute numbers and the percentages.

Feature		
Age	<65 [65, 80] >80	875 (42.33%) 727 (35.17%) 465 (22.50%)
Sex	Male Female	1173 (56.75%) 894 (43.25%)
Outcome	Non-Hospitalized Hospitalized	284 (13.74%) 1783 (86.26%)

4. Results and discussion

For the validation purposes of this work, 2 different experiments are performed, one for each of the analyses described in Section 3.1. In this way, the experiment I is conducted to estimate the risk of hospitalization for a given patient (Estimation of Non-Hospitalized/Hospitalized) while the experiment II is performed to estimate the risk of death (Estimation of Survival/Death). Hence, in this section, we discuss the results of the experiment I (Section 4.1) and the experiment II (Section 4.2). Finally, in Section 4.3, we compare the results of our proposal with those obtained in the main works of the state-of-the-art. It is important to note that each scenario has particular circumstances that lead to independent conclusions, an aspect that requires to perform an individual data and feature selection analysis for each particular case.

4.1. *Experiment I. Estimation of non-hospitalized/hospitalized patients*

The overall analysis of the distribution of 3 relevant features in the dataset can be seen in Table 6 as is the case of the Age, Sex, and Outcome, while Table 7 shows the statistical distribution of the clinical variables given the median, the first quartile and the third quartile.

Table 7
Statistical analysis of the numerical variables for the dataset built in the experiment I (estimation of Non-Hospitalized/Hospitalized).

Feature	Q1	Median	Q3
Height (cm)	156.00	163.00	170.00
Weight (kg)	68.65	80.00	90.00
BMI (kg/m^2)	26.67	29.74	32.91
LYMP (10^9/L)	0.70	1.00	1.46
LYMP (%)	11.30	17.30	25.60
D-Dimer (ng/mL)	466.25	750.00	1300.75
LDH (U/L)	202.00	264.00	362.00
Creatinine (mg/dL)	0.76	0.94	1.19
GFR (mL/min)	52.13	78.50	110.76
CRP (mg/L)	1.69	5.48	11.73
Ferritin (ng/mL)	147.50	413.50	812.75
IL-6 (ng/L)	7.60	18.85	47.05

Table 8
Analysis of the distribution of the values for each discrete available variable in the dataset built for the experiment I (estimation of Non-Hospitalization/Hospitalization) providing both the absolute numbers and the percentages.

Variable	Count (pct.)
Asthma	130 (6.29%)
Biological	18 (0.87%)
CCS	68 (3.29%)
Diabetes	355 (17.17%)
COPD	138 (6.68%)
LD	43 (2.08%)
HIV	4 (0.19%)
AHT	772 (37.35%)
Leukemia	12 (0.58%)
Lymphoma	18 (0.87%)
Neoplasm	210 (10.16%)
Chemotherapy	23 (1.11%)
Transplant	14 (0.68%)

Firstly, the distribution of the age of the patients is very similar among ranges, having the greatest amount between the ages of 0 and 65 (875 samples) that decreases as the age increases (727 samples between the ages of 65 and 80 and 465 samples for the ages greater than 80). On the other hand, the number of male patients is higher than the number of female patients, having 1173 males against 894 females.

With regard to the outcome, it can be concluded that the classes are imbalanced. In this particular case, where we have the class Non-Hospitalized alongside the class Hospitalized, the number of Hospitalized cases (1783 patients, that corresponds to the 86.26% of the dataset) is considerably greater than Non-Hospitalized (with 284 patients, that corresponds to the remaining 13.74% of the dataset), as expected. This is due to the fact that the health workers tend to register more data when the patient needs hospitalization, because they usually need more clinical testing during their hospital stay. This is the opposite in the case of the Non-Hospitalized cohort, as many of these patients are quickly released without any kind of additional testing.

On the other hand, when looking at the previous conditions of the patients, the imbalance is also noticeable, due to the fact that only few patients present a particular condition. In fact, as can be seen in Table 8, the most balanced variables are the AHT (where the 37.35% of the patients have this condition), the Diabetes (where the 17.17% of the patients have the condition) and the Neoplasm (with a 10.16% of patients that have this condition) while for the rest of the variables, the ratio of patients that have the condition is always lower than 10%. The most significant case of imbalance is the HIV, with only a 0.19% of patients that have the condition (just 4 samples). Nevertheless, despite dealing with such imbalance problems, the methodology shows to have a great robustness for this issue. This analysis shows that the selected samples are representative of the selected population, as hypertension [59] and diabetes [60] are 2 common previous conditions in western societies, explaining the high incidence of both disorders. Moreover, the fact that a great percentage of the patients are of an older age, makes them prone to suffer from more previous conditions, an aspect that is also reflected in the dataset.

Additionally, the results using the whole set of features can be seen in Table 9, with a comparison among the different used machine learning algorithms: SVM, DT, XGBoost, kNN and MLP. These results show that XGBoost obtains the best performance in terms of AUC-ROC, with a value of 0.8415 ± 0.00217. Regarding the individual metrics, the Sensitivity shows a high performance for this case, classifying correctly the 85.75% of the positive cases while shows a worse performance for the Specificity, as the XGBoost model only classifies correctly the 60.44% of the negative cases. However, it can be seen as a useful model, able to estimate the patients that need for hospitalization satisfactorily, as sensitivity is the most relevant metric to take into account for this particular case. In general, it can be seen that the results are satisfactory given the complexity of the estimation problem that is being solved. To analyze this last model more deeply, Fig. 3 shows a representative confusion matrix obtained when using the most appropriate algorithm for this experiment as reference (XGBoost) and the ROC curves of all the classifiers after training with the whole set of features. In particular, regarding the cases that were correctly classified, this confusion matrix shows 304 true positives and 40 true negatives. With regard to cases that were misclassified, it shows 56 false negatives (Hospitalized cases that were classified as Non-Hospitalized) and 14 false positives (Non-Hospitalized cases that were classified as Hospitalized). In terms of ROC curves, the performances among classifier models are quite similar, where XGBoost obtains the highest value. However, it is remarkable that kNN achieves a notable lower effectiveness, as reflected in the previous Table.

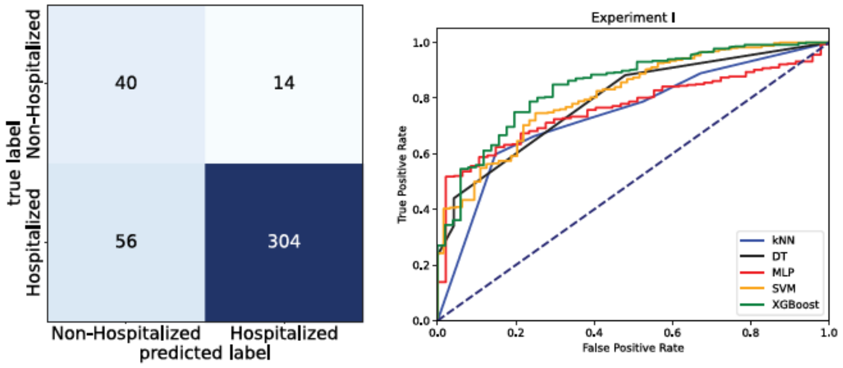
With regard to the feature selection process, the ranking of the most important features for each method can be seen in Fig. 4. From these rankings, some interesting aspects can be concluded. Firstly, in the case of the Fisher Scoring, it can be seen that the 2 most important features are Age and Age Range. This demonstrates that the main clinical criteria to decide if a patient needs to be admitted to the hospital or not is the age. A similar conclusion extracts the Mutual Information method, as the Age is considered as the most important feature and the Age Range is included within the top 4 of the ranking. Apart from the age variables, AHT also shows to be significant, as Fisher Scoring places it within the top 3, the Mutual Information within the top 6 and the Variance-based Ranking method as the third most important discrete variable. Moreover, GFR, a descriptor of the patient renal activity, is given a high score by the 3 methods, as is placed within the top 7 by the Fisher Scoring, within the top 2 by the Mutual Information method and within the top 5 by the Variance-based Ranking method.

The feature selection methods give also a great importance to BMI and Weight. In particular, Fisher Scoring places Weight and BMI at the fourth and sixth position, respectively, while Mutual Information as the third and the seventh most important features, in the same order. Moreover, in the case of the Variance-based method, both variables are placed within the top 10 of the ranking. Furthermore, the CRP is also given a great importance, being always within the top 11 of the ranking for the 3 feature selection methods. LDH also seems to be another relevant feature (top 9 for the Fisher Scoring method, top 10 for Mutual Information and top 3 for Variance-based Ranking). Finally, Diabetes, an important previous condition, is also remarkable given that it is placed within the top 10 by the Fisher Scoring method, within the top 11 by the Mutual Information method and as the fourth most important discrete variable by the Variance-based Ranking method.

In the particular case of Mutual Information, the features CCS, Chemotherapy, HIV, Leukemia, Lymphoma, LYMP and Sex are given a negligible score. Therefore, it is shown that the values of these variables are unable to precisely determine the outcome of the patient. Regarding the variable Sex, despite being placed by the Variance-based Ranking method as the second most important discrete variable, the other 2 methods give a low position to this feature, meaning that the

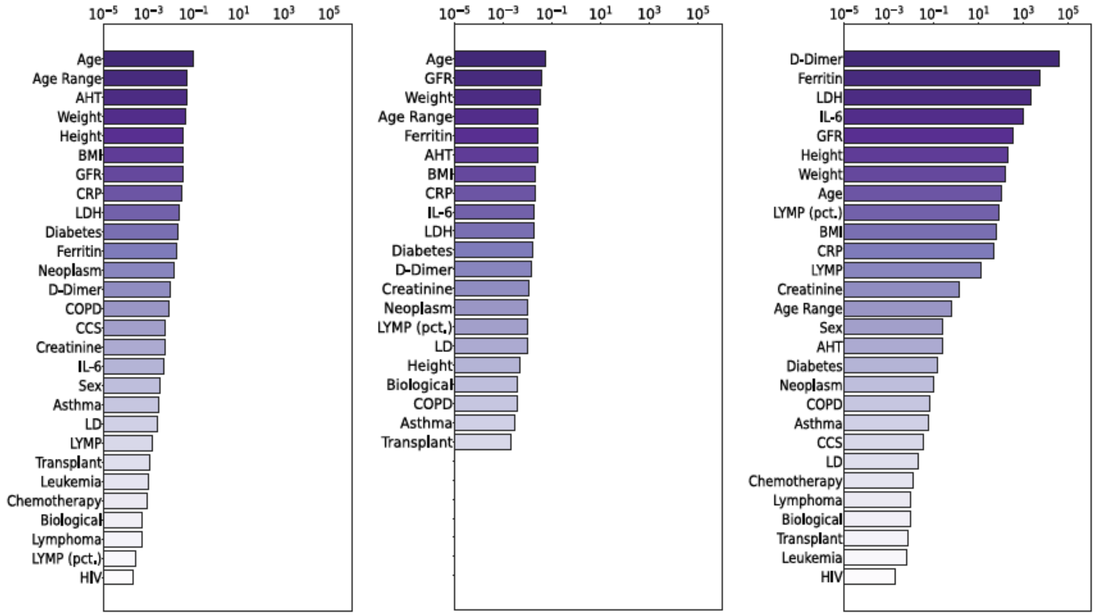
Table 9
Results obtained after training the models to estimate the output Non-Hospitalized/Hospitalized (experiment I) using all the features.

	SVM	DT	XGBoost	kNN	MLP
AUC-ROC	0.7975 ± 0.0262	0.7912 ± 0.0115	0.8415 ± 0.0217	0.6976 ± 0.0303	0.8170 ± 0.0157
Sensitivity	44.88% ± 4.33%	79.55% ± 20.38%	85.75% ± 5.02%	72.45% ± 1.50%	76.37% ± 3.20%
Specificity	90.78% ± 4.82%	58.47% ± 18.78%	60.44% ± 5.61%	61.58% ± 3.70%	71.53% ± 3.51%



(a) (b)

Fig. 3. Representative graphical results obtained in the experiment I. (a) Representative confusion matrix obtained with the most appropriate classifier for this experiment (XGBoost) training with all the features. (b) ROC curves of the classifiers used in this experiment, training with the whole set of features.



(a) (b) (c)

Fig. 4. Ranking of features according to the score given by each feature selection method for the experiment I (estimation of Non-Hospitalization/Hospitalization). The *x* axes are displayed in logarithmic scale to improve the visualization of the differences among the scores. Those features with a negligible score were removed from the chart. (a) Fisher Scoring. (b) Mutual Information. (c) Variance-based Ranking.

correlation between the values of Sex and the need for hospitalization has a great uncertainty. From this study, we can extract that the age of patient is an important variable to determine if a person needs hospitalization or not. Similarly, it can be seen that those variables related with weight also have a strong correlation with the necessity of hospitalization. In addition, those features that have a relationship with prevalent previous conditions (in this case, AHT and Diabetes) are significant, too. Finally, we can find some other important features that determine important organic processes as well, such as the renal and liver activity (LDH and GFR) and inflammation (as is the case of the CRP variable). This is a valuable conclusion, given that COVID-19 is a multi-organic disease that can affect the previously mentioned organs, because it provides a global picture of the pathology severity. In the same way, any variable related with inflammation can also help to understand the extent of the pathological condition.

In terms of the performance evolution given the used number of features from the dataset, Fig. 5 shows that there is a trend of improvement as the number of features increases when using the model with the best global performance, which in this case is provided by the XGBoost algorithm. This improvement is more noticeable when adding the first top features to the dataset, while it starts to stabilize from a certain amount of features onwards. In fact, for this scenario, it can be seen that the performance starts to flatten at around 10 features using all the feature selection methods. Another interesting aspect is that

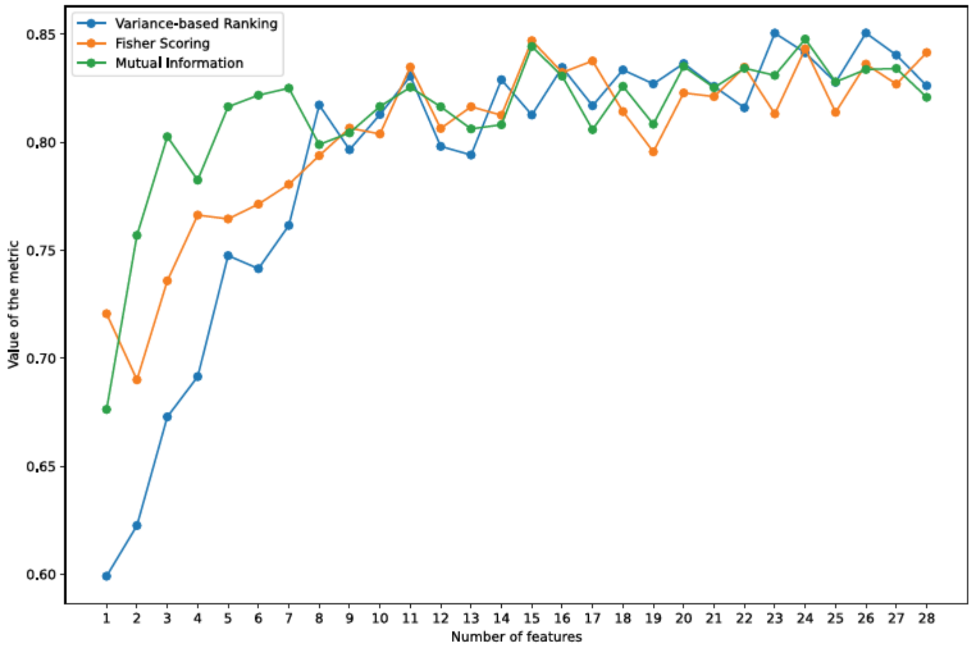


Fig. 5. Evolution of the AUC-ROC values for the problem of classifying Non-Hospitalized/Hospitalized (experiment I) given the number of features using the XGBoost algorithm.

Table 10
Analysis of the distribution of 3 relevant features in the dataset built for the experiment II (estimation of Survival/Death), given by the absolute numbers and the percentages.

Feature		
Age	<65	672 (37.69%)
	[65, 80]	670 (37.58%)
	>80	441 (24.73%)
Sex	Male	1031 (57.82%)
	Female	752 (42.18%)
Outcome	Survival	1357 (76.11%)
	Death	426 (23.89%)

Table 11
Analysis of the numerical variables of the dataset built for the experiment II (estimation of Survival/Death).

Feature	Q1	Median	Q3
Height (cm)	156.00	163.00	170.00
Weight (kg)	68.55	80.00	90.00
BMI (kg/m^2)	26.62	29.73	32.87
LYMP (10^9/L)	0.70	1.00	1.45
LYMP (%)	10.80	17.00	25.30
D-Dimer (ng/mL)	461.50	758.00	1324.25
LDH (U/L)	209.00	272.00	372.00
Creatinine (mg/dL)	0.77	0.94	1.20
GFR (mL/min)	51.91	78.50	109.94
CRP (mg/L)	1.91	5.95	12.20
Ferritin (ng/mL)	157.00	419.00	821.00
IL-6 (mg/L)	7.60	19.10	48.80

Fisher Scoring and Mutual Information have a similar behavior, while the Variance-based Ranking method starts with a lower performance, but matches the other methods when the number of features is 8. This occurs when the feature Age is initially included in the subset, being significant of the importance that the attribute implies in this particular problem. It is interesting to remark that these results are one of the main strong points of our work, as this study can be useful to understand which are the most important features and to comprehend the behavior of the model as the amount of features is increased.

4.2. Experiment II. Estimation of the Survival/Death

An overall analysis of the age ranges, sex and outcome of the patients in the dataset built for this experiment can be seen in Table 10 and the distribution of the clinical variables in terms of median, first quartile and third quartile can be seen in Table 11. Firstly, we can observe that the distribution of the age ranges is very similar as in the experiment I (*i.e.*, the amount of patients is higher for the youngest ranges and decreases as the age increases). With regard to the sex of the patients, the 57.82% are males, while the remaining 42.18% are females. Finally, the dataset shows a 76.11% of survival ratio and a 23.89% of death ratio. This is significant that the dataset for this experiment also experiences a significant imbalance. However, the proposed methodology presents a satisfactory robustness to deal with this issue.

On the other hand, the analysis of the values distribution of the discrete variables is shown in Table 12. As expected, the data is very similar to the previous case, with only slight differences. Once again, the most balanced variables are AHT (having a 41.45% of patients that present this condition), Diabetes (having a 19.24% of patients with this condition), and Neoplasm (having an 11.55% of the patients with the condition) while the rest of the variables have less than 10% of the patients with each condition. Once again, the most imbalanced variable is HIV, as only the 0.17% of the patients present this condition (just 3 samples in the dataset). In this experiment II, given that the dataset is very similar to the dataset of the experiment I, the selected samples are also representative of the population with a greater incidence of hypertension and diabetes, an aspect that is partially explained by the fact that many patients are of an older age.

Moreover, Table 13 shows the performance that was obtained using the whole set of features for this experiment II. Overall, the XGBoost is the method that achieves the best global performance with an AUC of 0.7992 ± 0.0104. These AUC values represent a high performance that, however, are influenced by the fact that the studied variables are relevant and useful but only able to partially characterize the complex problem of estimation that is being addressed. In this case, with regard to the individual metrics, the Specificity shows a satisfactory value as the model is able to classify correctly the 80.08% of the negative cases, while is considerably lower in the case of the Sensitivity, as only the 56.28% of the positive cases are classified correctly. This is an undesirable characteristic for this particular problem, as it is more

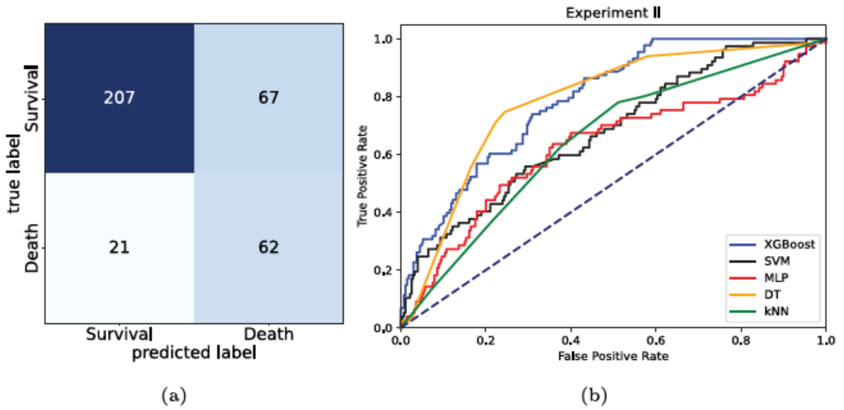


Fig. 6. Remarkable graphical results of the experiment II. (a) Confusion matrix representative of the most appropriate classifier for this experiment (decision tree) trained with the whole set of features. (b) ROC curves of the classifiers used in this experiment after training with all the features.

Table 12
Analysis of the balance of each discrete variable in the dataset built for the experiment I (estimation of Non-Hospitalized/Hospitalized), depending on that if the patients have a certain condition or not.

Variable	Count (pct.)
Asthma	121 (6.79%)
Biological	17 (0.95%)
CCS	68 (3.81%)
Diabetes	343 (19.24%)
COPD	135 (7.57%)
LD	42 (2.36%)
HIV	3 (0.17%)
AHT	739 (41.45%)
Leukemia	12 (67.30%)
Lymphoma	17 (95.34%)
Neoplasm	206 (11.55%)
Chemotherapy	22 (1.23%)
Transplant	14 (0.79%)

critical to determine precisely which are the patients with a higher risk of death. Therefore, despite providing a lower value of AUC-ROC, the decision tree is more appropriate for this problem, as it obtains the highest value of sensitivity, classifying correctly the 75.87% of the positive cases. A more visual analysis of the performance can be seen in Fig. 6, with a representative confusion matrix of the most suitable algorithm in this experiment (Decision Tree) as well as the ROC curves of all the classifiers using all the available features. The confusion matrix shows an effectiveness of 62 true positives, 207 true negatives, 21 false negatives (cases of class Death that were classified as Survival) and 67 false positives (cases classified as Death that actually belonged to class Survival). Regarding the ROC curves, SVM, kNN and MLP have a notably lower in performance compared with the remaining 2 algorithms, in the same line as stated in the Table.

With regard to the feature selection process, the ranking scores of the top features for each method can be seen in Fig. 7. Firstly, it can be seen that the Variance-based Ranking method is once again considerably different, as numerical variables are at the top of the ranking, while discrete variables are at the bottom. As the datasets that were built for the experiment I and the experiment II are very similar between them and given that this feature selection method only focuses on features without taking into account the output classes, the ranking and the scores are also very similar. With regard to the obtained rankings of the features, both the Fisher Scoring and the Mutual Information methods select the Age Range and the Age as the 2 most important features, in the same way as it happened in experiment I. Similarly, Fisher Scoring also places AHT as the third most important variable, while Mutual Information places it within the top 6 and Variance-based Ranking as the third most important discrete variable.

Apart from that, other relevant feature is the Neoplasm, as is placed within the top 3 in the case of the Fisher Scoring method, within the top 11 for the Mutual Information method and as the fifth most important discrete variable under the point of view of the Variance-based Ranking method. Once again, the Diabetes also proves to be an important feature (placed as eighth in the case of the Fisher Scoring method, tenth in the case of the Mutual Information method and as the fourth most important feature in the case of the Variance-based Ranking). In addition, Chemotherapy can also be mentioned as a remarkable variable, placed within the top 7 in the case of the Fisher Scoring, within the top 9 in the case of the Mutual Information method and within the top 10 of the discrete variables in the case of the Variance-based Ranking method. In the case of the count of lymphocytes, it is remarkable that the percentage is much more important than the absolute count. Given this percentage, Fisher Scoring gives it the ninth position, Mutual Information gives it the fifth position, while the Variance-based Ranking gives it the ninth position. Other significant variables that can be mentioned are CCS (fifth position given the Fisher Scoring algorithm, eighth position given the Mutual Information method and the eighth position of the most relevant discrete variables in the case of the Variance-based Ranking method) and Creatinine that, despite being placed as the least important numerical variable due to its lower variance, is given the top 4 by the Fisher Scoring algorithm and the top 3 by Mutual Information. Finally, it can be seen that the Mutual Information method gives a negligible score to Ferritin, Height, Weight, BMI, CRP, Sex, Asthma, LD, Biological, Transplant, and HIV.

For this second experiment, there are several differences with regard to the study of the features. Firstly, it can also be seen that age and prevalent previous conditions (AHT and Diabetes) are relevant to determine the death risk of a patient, similarly as in the experiment I. However, those variables related with Weight have less importance (in fact, Fisher Scoring places them almost at the bottom of the ranking and Mutual Information gives them negligible scores) despite the fact that Variance-based method gives them high positions due to their great variability. Similarly, the features LDH, GFR and CRP are given less importance as well that, however, is considerably less noticeable than in the previous case. On the other hand, Creatinine, a variable that is significant of the renal filtration capabilities, gains a great importance despite the drop of GFR. Once again, this demonstrates the significance of the renal activity and the outcome of COVID-19 patients. Furthermore, Neoplasm and Chemotherapy (both related with cancer scenarios) gain a great importance similarly as CCS (given that corticosteroids reduce inflammation and affect immune system) and LYMP (pct.), the last one due to the fact that amount of lymphocytes is indicative of the immune system function.

In the case of training with the whole set of available features, for a more straightforward interpretation of the results, we decided to focus on those obtained by the model with the best overall performance, (in

Table 13
Results obtained after training the models to estimate the output Survival/Death (experiment II) using all the available features.

	SVM	DT	XGBoost	kNN	MLP
AUC-ROC	0.6715 ± 0.0394	0.7839 ± 0.0298	0.7992 ± 0.0104	0.6261 ± 0.0229	0.6882 ± 0.0427
Sensitivity	22.96% ± 4.29%	75.87% ± 8.26%	56.28% ± 3.78%	52.94% ± 3.67%	55.44% ± 9.56%
Specificity	89.09% ± 1.59%	69.54% ± 7.44%	80.08% ± 1.91%	64.34% ± 2.50%	77.44% ± 3.30%

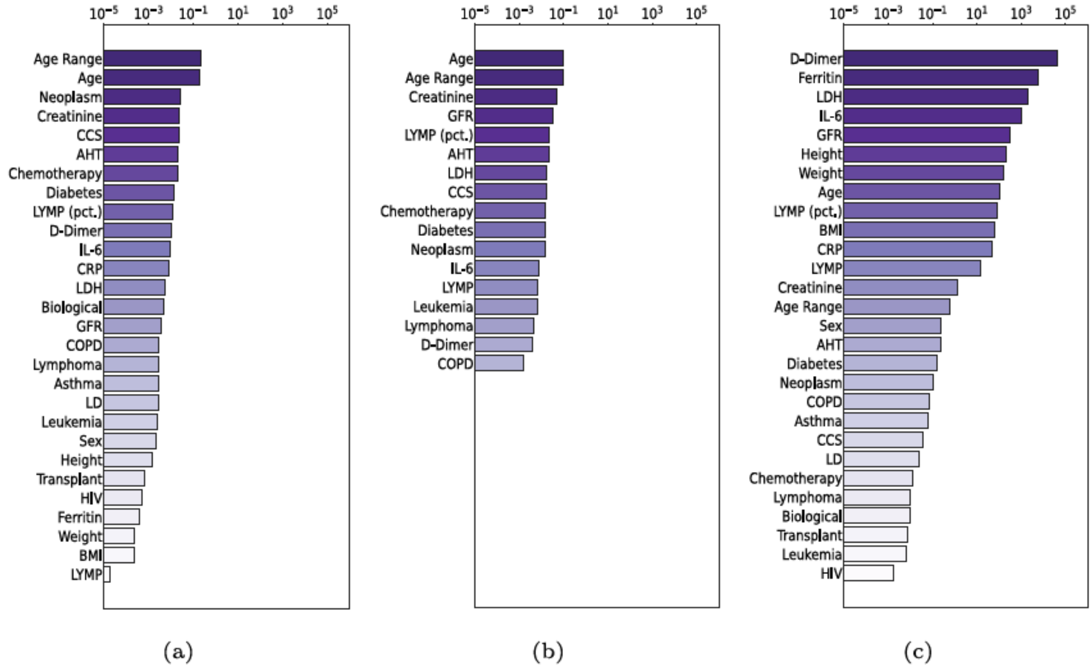


Fig. 7. Ranking of features according to the score given by each feature selection method for the experiment II (estimation of Survival/Death). The *x* axes are shown in logarithmic scale to improve the visualization of the differences among the scores. Those features with a negligible score were removed from the chart. (a) Fisher Scoring. (b) Mutual Information. (c) Variance-based Ranking.

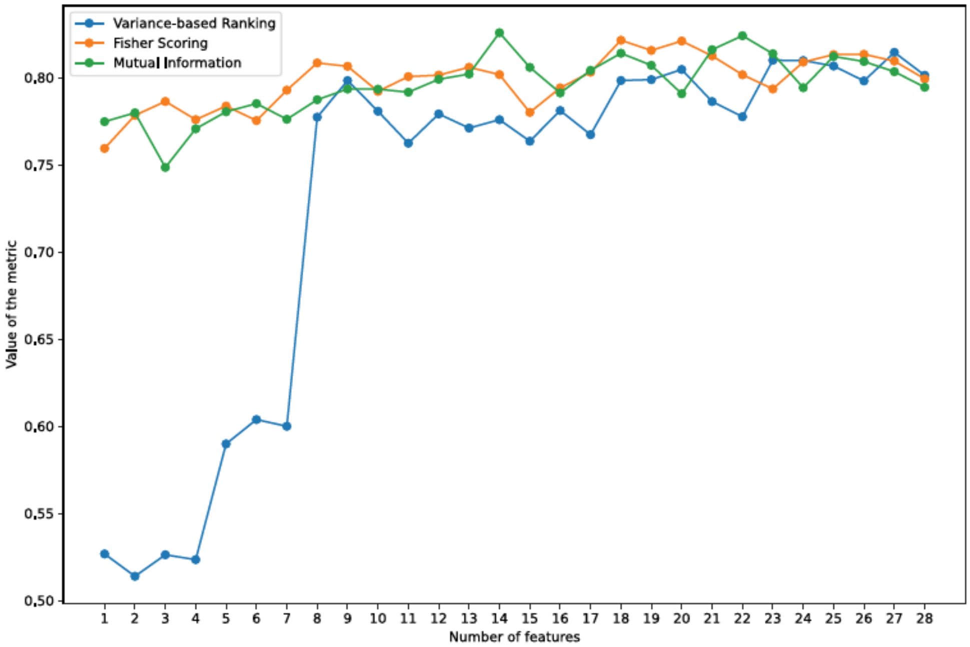


Fig. 8. Evolution of the AUC-ROC values for the problem of classifying Survival/Death (experiment II) given the number of features using the XGBoost algorithm.

Table 14
Comparison of our proposal and the main works of the state-of-the-art. It must be taken into account that the used datasets and the conditions of the experimentation can be considerably different.

		Xu et al. [22]	Polilli et al. [23]	Pishgar et al. [24]	Ali et al. [18]	Ours
Risk of hospitalization	AUC-ROC	–	0.9100	–	–	0.8415
	Sensitivity	–	80.00%	–	–	85.75%
	Specificity	–	87.00%	–	–	60.44%
Risk of death	AUC-ROC	0.8500	0.9100	0.9000	0.8100	0.7839
	Sensitivity	22.00%	89.00%	72.70%	83.00%	75.87%
	Specificity	97.00%	79.00%	80.00%	–	69.54%

this case, the XGBoost). Particularly, for the 2 first methods, the Age Range and the Age variables show to be the most important features. In the case of the Variance-based Ranking method, due to its nature, the Age Range is lower at the ranking, but it is still at the top of the discrete variables.

In terms of the performance evolution as the number of selected features grows, we decided to select the most representative classification model for such purpose, the XGBoost, which is the one that obtains the best overall performance. In particular, Fig. 8 shows a tendency of improvement when the amount of features is higher. This improvement is slight in the case of Fisher Scoring and Mutual Information, but it is considerably noticeable in the case of the Variance-ranking based method. In the latter case, the performance is significantly lower when the amount of selected features is small (in this case, less than 8 features). However, there is a performance peak while using 8 features, an aspect that represents the importance of the Age variable, as this is the first subset where the feature is added. From that number of features onward, the performance improvement starts to flatten. Once again, these results remark the strengths of our work, given that this exhaustive experimentation explains the contribution of each feature on the performance (which is a clinically relevant aspect) and the behavior of the models with each subset of features.

4.3. Comparison with other contributions

When comparing our work with the rest of the state-of-the-art approaches, it is important to note that there is a lack of publicly available data as it is difficult to find a consensus among the different criteria that it is used in the healthcare services worldwide to diagnose COVID-19 patients, hindering the process to obtain a general purpose public dataset suitable for performing a broad range of different analyses. Therefore, the available methods need to design ad-hoc datasets for specific analyses, that are restricted to the available data. Globally, we can point out that our contribution presents an exhaustive analysis of the feature selection process, that evaluates how the performance improves as the amount of features increases. Furthermore, apart from the clinical relevance of the conclusions that are extracted from the experiments, we also discuss the possible applicability of the methodology to the daily clinical practice, thanks to the trained classification models and the most relevant and useful variables that were identified during the feature selection process. We can also point out that the statistical analysis of the variables of the dataset shows that we are working with a representative dataset of the pathological conditions that are more prevalent in modern Western societies as is the case of hypertension, diabetes and other scenarios such as the overweight or obesity. This makes the analysis more relevant as it evaluates the correlation between these prevalent conditions and the increased risk of the patients that suffer from them in a COVID-19 scenario.

However, despite the difficulty of comparing our proposal with the methods found in the state-of-the-art, in this section we perform a discussion of these characteristics, pretending to be as fair as possible. In particular, Table 14 shows the comparison among our proposal and the main methods of the state-of-the-art. For our methodology, we selected the most appropriate model for each experiment (*i.e.*, XGBoost for Experiment I and Decision Tree for Experiment II). With regard to risk of hospitalization, it can be seen that the work from Polilli et al. [23] outperforms our proposal in terms of AUC and specificity, but it is worse in terms of sensitivity. As it was also explained in the discussion of the Experiment II, sensitivity is an important and critical metric for the problem that is being solved. In this sense, our work shows to be more appropriate for that task. In the case of the risk of death, the AUC-ROC of our method is the lowest among all the approaches. However, the sensitivity obtained with our proposal is considerably higher than the one obtained in the work of Xu et al. [22] and better than the sensitivity achieved in the work of Pishgar et al. [24]. In general, this demonstrates that the different approaches are appropriate to study the risk of death of a COVID-19 patient. Moreover, in the case of Ali et al., the performance is close to the one obtained in our case in terms of AUC-ROC and sensitivity. Once again, it is important to remark that the datasets used in these works are considerably different, and that, in this line, the discussion tried to be as fair as possible.

Regarding the set of features that is used in the works from the state-of-the-art, a great heterogeneity is one of the most remarkable points that can be mentioned. Overall, age and gender are usually present in these datasets for their great importance not only in the case of COVID-19 but in any clinical scenario. Some of the works presented in this comparison consider only laboratory findings. Nevertheless, other works contemplate the use of both laboratory findings and previous clinical conditions of the patients, similarly as in our proposal. Overall, it can be seen that some features are often shared among contributions, for instance, the previously-mentioned age and gender, variables directly or indirectly related with red blood cells and white blood cells, hypertension, diabetes, obesity, cardiovascular and respiratory issues (such as asthma or COPD, among others), and variables directly related with immunity.

5. Conclusions

Given that COVID-19 is a challenging disease for the healthcare services worldwide, it is important to develop useful CAD systems and provide useful information to help the clinicians to take decisions in so critical environments. In this context, the clinical data of the patients that are stored in form of electric records, can be exploited in developing these automatic methods to perform relevant tasks. For this work, we have selected 2 different study scenarios, the most critical and relevant for both the healthcare services and patients with regard to a COVID-19 severe infection: estimate if a COVID-19 patient needs hospitalization or not and if a COVID-19 patient is at low or high risk of death. This was performed using 4 different machine learning models: Support Vector Machine, Decision Tree, XGBoost, k Nearest Neighbors and Multilayer Perceptron. On each task, we performed an exhaustive statistical analysis of the most relevant aspects that can be derived from the dataset, the study of the imbalance of each class and the features that determine the previous conditions that the patients present. In the same way, this contribution includes a detailed analysis of the feature selection process using 3 different methods and the study of the impact that has the number of features used to train on the performance of the classification models. Moreover, in this manuscript, we thoroughly analyze the performance of the machine learning models used for the tasks, training with the whole set of available features.

As possible lines of future works, it could be relevant to perform other different analyses of the provided dataset after discretizing each clinical variable with its corresponding reference ranges. The analyses herein presented could also be complemented using information provided by relevant medical imaging captures, such as X-ray or CT devices with the aim to improve the performance of the classification models, integrating information and analyses from a wider range of resources. These imaging modalities are relevant as they could provide relevant biomarkers indicative of how the disease will behave in a given patient.

CRediT authorship contribution statement

Daniel I. Morís: Methodology, Software, Validation, Writing – original draft, Visualization. **Joaquim de Moura:** Methodology, Software, Validation, Writing – review & editing, Visualization, Supervision. **Pedro J. Marcos:** Conceptualization, Data curation, Methodology, Supervision. **Enrique Míguez Rey:** Data curation, Investigation, Methodology, Supervision. **Jorge Novo:** Conceptualization, Validation, Writing – review & editing, Supervision. **Marcos Ortega:** Conceptualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Funding

This research was funded by ISCIII, Government of Spain, DTS18/00136 research project; Ministerio de Ciencia e Innovación *y* Universidades, Government of Spain, RTI2018-095894-B-I00 research project; Ministerio de Ciencia e Innovación, Government of Spain through the research project with reference PID2019-108435RB-I00; CCEU, Xunta de Galicia through the predoctoral grant contract ref. ED481A 2021/196; and Grupos de Referencia Competitiva, grant ref. ED431C 2020/24; Axencia Galega de Innovación (GAIN), Xunta de Galicia, grant ref. IN845D 2020/38; CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from CCEU, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%). Funding for open access charge: Universidade da Coruña/CISUG.

References

[1] K. Yuki, M. Fujiogi, S. Koutsogiannaki, COVID-19 pathophysiology: A review, Clin. Immunol. 215 (2020) 108427, http://dx.doi.org/10.1016/j.clim.2020.108427.

[2] W.T. Siow, M.F. Liew, B.R. Shrestha, F. Muchtar, K.C. See, Managing COVID-19 in resource-limited settings: critical care considerations, 2020, http://dx.doi.org/10.1186/s13054-020-02890-x.

[3] Y.-d. Gao, M. Ding, X. Dong, J.-j. Zhang, A. Kursat Azkur, D. Azkur, H. Gan, Y.-l. Sun, W. Fu, W. Li, et al., Risk factors for severe and critically ill COVID-19 patients: a review, Allergy 76 (2) (2021) 428–455, http://dx.doi.org/10.1111/all.14657.

[4] H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, Predicting COVID-19 mortality with electronic medical records, NPJ Digit. Med. 4 (1) (2021) 1–10, http://dx.doi.org/10.1038/s41746-021-00383-x.

[5] J. Yanase, E. Triantaphyllou, A systematic survey of computer-aided diagnosis in medicine: Past and present developments, Expert Syst. Appl. 138 (2019) 112821, http://dx.doi.org/10.1016/j.eswa.2019.112821.

[6] M.E. Karar, E.E.-D. Hemdan, M.A. Shouman, Cascaded deep learning classifiers for computer-aided diagnosis of COVID-19 and pneumonia diseases in X-ray scans, Complex Intell. Syst. 7 (1) (2021) 235–247, http://dx.doi.org/10.1007/s40747-020-00199-4.

[7] P.L. Vidal, J. de Moura, J. Novo, M. Ortega, Multi-stage transfer learning for lung segmentation using portable X-ray devices for patients with COVID-19, Expert Syst. Appl. 173 (2021) 114677, http://dx.doi.org/10.1016/j.eswa.2021.114677.

[8] T. Li, W. Wei, L. Cheng, S. Zhao, C. Xu, X. Zhang, Y. Zeng, J. Gu, Computer-aided diagnosis of COVID-19 CT scans based on spatiotemporal information fusion, J. Healthc. Eng. 2021 (2021) http://dx.doi.org/10.1155/2021/6649591.

[9] D.I. Morís, J.J. de Moura Ramos, J.N. Buján, M.O. Hortas, Data augmentation approaches using cycle-consistent adversarial networks for improving COVID-19 screening in portable chest X-ray images, Expert Syst. Appl. 185 (2021) 115681, http://dx.doi.org/10.1016/j.eswa.2021.115681.

[10] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260, http://dx.doi.org/10.1126/science.aaa8415.

[11] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (6) (2017) 1–45, http://dx.doi.org/10.1145/3136625.

[12] E.E. Tripoliti, T.G. Papadopoulos, G.S. Karanasiou, K.K. Naka, D.I. Fotiadis, Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques, Comput. Struct. Biotechnol. J. 15 (2017) 26–47.

[13] H. Asadi, R. Dowling, B. Yan, P. Mitchell, Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy, PLoS One 9 (2) (2014) e88225, http://dx.doi.org/10.1371/journal.pone.0088225.

[14] B. Alizadeh, A. Alibabaei, S. Ahmadi, S.F. Maroufi, S. Ghafouri-Fard, S. Nateghinia, Designing predictive models for appraisal of outcome of neurosurgery patients using machine learning-based techniques, Interdiscip. Neurosurg. 31 (2023) 101658, http://dx.doi.org/10.1016/j.inat.2022.101658.

[15] M. Nascimben, L. Rimondini, D. Corà, M. Venturin, Polygenic risk modeling of tumor stage and survival in bladder cancer, BioData Min. 15 (1) (2022) http://dx.doi.org/10.1186/s13040-022-00306-w.

[16] Q. Yuan, T. Cai, C. Hong, M. Du, B.E. Johnson, M. Lanuti, T. Cai, D.C. Christiani, Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer, JAMA Netw. Open 4 (7) (2021) e2114723, http://dx.doi.org/10.1001/jamanetworkopen.2021.14723.

[17] S. Aktar, M.M. Ahamad, M. Rashed-Al-Mahfuz, A. Azad, S. Uddin, A. Kamal, S.A. Alyami, P.-I. Lin, S.M.S. Islam, J.M. Quinn, V. Eapen, M.A. Moni, Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: Statistical analysis and model development, JMIR Med. Inf. 9 (4) (2021) e25884, http://dx.doi.org/10.2196/25884.

[18] S. Ali, Y. Zhou, M. Patterson, Efficient analysis of COVID-19 clinical data using machine learning models, 2021, http://dx.doi.org/10.48550/arXiv.2110.09606.

[19] J.L. Domínguez-Olmedo, Á. Gragera-Martínez, J. Mata, V.P. Álvarez, Machine learning applied to clinical laboratory data in Spain for COVID-19 outcome prediction: Model development and validation, J. Med. Internet Res. 23 (4) (2021) e26211, http://dx.doi.org/10.2196/26211.

[20] H. Magunia, S. Lederer, R. Verbuecheln, B.J. Gilot, M. Koeppen, H.A. Haeberle, V. Mirakaj, P. Hofmann, G. Marx, J. Bickenbach, et al., Machine learning identifies ICU outcome predictors in a multicenter COVID-19 cohort, Crit. Care 25 (1) (2021) 1–14, http://dx.doi.org/10.1186/s13054-021-03720-4.

[21] E. Hernández-Pereira, O. Fontenla-Romero, V. Bolón-Canedo, B. Cancela-Barizo, B. Guijarro-Berdiñas, A. Alonso-Betanzos, Machine learning techniques to predict different levels of hospital care of CoVid-19, Appl. Intell. 52 (6) (2021) 6413–6431, http://dx.doi.org/10.1007/s10489-021-02743-2.

[22] Y. Xu, A. Trivedi, N. Becker, M. Blazes, J.L. Ferres, A. Lee, W.C. Liles, P.K. Bhatraju, Machine learning-based derivation and external validation of a tool to predict death and development of organ failure in hospitalized patients with COVID-19, Sci. Rep. 12 (1) (2022) http://dx.doi.org/10.1038/s41598-022-20724-4.

[23] E. Polilli, A. Frattari, J.E. Esposito, M. D'Amato, G. Rapacchiale, A. D'Intino, A. Albani, G.D. Iorio, F. Carinci, G. Parruti, Reliability of predictive models to support early decision making in the emergency department for patients with confirmed diagnosis of COVID-19: the Pescara Covid Hospital score, BMC Health Serv. Res. 22 (1) (2022) http://dx.doi.org/10.1186/s12913-022-08421-4.

[24] M. Pishgar, S. Harford, J. Theis, W. Galanter, J.M. Rodríguez-Fernández, L.H. Chaisson, Y. Zhang, A. Trotter, K.M. Kochendorfer, A. Boppana, H. Darabi, A process mining- deep learning approach to predict survival in a cohort of hospitalized COVID-19 patients, BMC Med. Inf. Decis. Making 22 (1) (2022) http://dx.doi.org/10.1186/s12911-022-01934-2.

[25] M.G.A.M. van der Velde, M.J. van der Aa, M.H.C. van Daal, M.N.T. Kremers, C.J.P.W. Keijsers, S.M.J. van Kuijk, H.R. Haak, Performance of the APOP-screener for predicting in-hospital mortality in older COVID-19 patients: a retrospective study, BMC Geriatr. 22 (1) (2022) http://dx.doi.org/10.1186/s12877-022-03274-2.

[26] I. Bendavid, L. Statlender, L. Shvartser, S. Teppler, R. Azullay, R. Sapir, P. Singer, A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from COVID-19, Sci. Rep. 12 (1) (2022) http://dx.doi.org/10.1038/s41598-022-14758-x.

[27] J.H. Chamberlin, G. Aquino, S. Nance, A. Wortham, N. Leaphart, N. Paladugu, S. Brady, H. Baird, M. Fiegel, L. Fitzpatrick, M. Kocher, F. Ghesu, A. Mansoor, P. Hoelzer, M. Zimmermann, W.E. James, D.J. Dennis, B.A. Houston, I.M. Kabakus, D. Baruah, U.J. Schoepf, J.R. Burt, Automated diagnosis and prognosis of COVID-19 pneumonia from initial ER chest X-rays using deep learning, BMC Infect. Dis. 22 (1) (2022) http://dx.doi.org/10.1186/s12879-022-07617-7.

[28] P.K. Chaudhary, R.B. Pachori, FBSED based automatic diagnosis of COVID-19 using X-ray and CT images, Comput. Biol. Med. 134 (2021) 104454, http://dx.doi.org/10.1016/j.compbiomed.2021.104454.

[29] D. Bermejo-Peláez, R.S.J. Estépar, M. Fernández-Velilla, C.P. Miras, G.G. Madueño, M. Benegas, C.G. Rivera, S. Cuerpo, M. Luengo-Oroz, J. Sellarés, M. Sánchez, G. Bastarrika, G.P. Barba, L.M. Seijo, M.J. Ledesma-Carbayo, Deep learning-based lesion subtyping and prediction of clinical outcomes in COVID-19 pneumonia using chest CT, Sci. Rep. 12 (1) (2022) http://dx.doi.org/10.1038/s41598-022-13298-8.

[30] J. de Moura, J. Novo, M. Ortega, Fully automatic deep convolutional approaches for the analysis of COVID-19 using chest X-ray images, Appl. Soft Comput. 115 (2022) 108190, http://dx.doi.org/10.1016/j.asoc.2021.108190.

[31] J. de Moura, L.R. García, P.F.L. Vidal, M. Cruz, L.A. López, E.C. Lopez, J. Novo, M. Ortega, Deep convolutional approaches for the analysis of COVID-19 using chest X-Ray images from portable devices, IEEE Access 8 (2020) 195594–195607, http://dx.doi.org/10.1109/ACCESS.2020.3033762.

[32] A. Sinha, S.P. Joshi, P.S. Das, S. Jana, R. Sarkar, An ML prediction model based on clinical parameters and automated CT scan features for COVID-19 patients, Sci. Rep. 12 (1) (2022) http://dx.doi.org/10.1038/s41598-022-15327-y.

[33] M. Zuin, G. Rigatelli, G. Zuliani, A. Rigatelli, A. Mazza, L. Roncon, Arterial hypertension and risk of death in patients with COVID-19 infection: systematic review and meta-analysis, J. Infect. 81 (1) (2020) e84–e86, http://dx.doi.org/10.1016/j.jinf.2020.03.059.

[34] S. Peric, T.M. Stulnig, Diabetes and COVID-19: Disease-management-people, Wiener Klinische Wochenschrift 132 (13) (2020) 356–361, http://dx.doi.org/10.1007/s00508-020-01672-3.

[35] F.V. Gerayeli, S. Milne, C. Cheung, X. Li, C.W.T. Yang, A. Tam, L.H. Choi, A. Bae, D.D. Sin, COPD and the risk of poor outcomes in COVID-19: A systematic review and meta-analysis, EClinicalMedicine 33 (2021) http://dx.doi.org/10.1016/j.eclinm.2021.100789.

[36] N.F. Mendes, C.P. Jara, E. Mansour, E.P. Araújo, L.A. Velloso, Asthma and COVID-19: a systematic review, Allergy, Asthma Clin. Immunol. 17 (1) (2021) 1–12, http://dx.doi.org/10.1186/s13223-020-00509-y.

[37] J. Wu, S. Song, H.-C. Cao, L.-J. Li, Liver diseases in COVID-19: Etiology, treatment and prognosis, World J. Gastroenterol. 26 (19) (2020) 2286, http://dx.doi.org/10.3748/wjg.v26.i19.2286.

[38] O.M. Al-Quteimat, A.M. Amer, The impact of the COVID-19 pandemic on cancer patients, Am. J. Clin. Oncol. (2020) http://dx.doi.org/10.1097/COC.0000000000000712.

[39] H. Mirzaei, W. McFarland, M. Karamouzian, H. Sharifi, COVID-19 among people living with HIV: a systematic review, AIDS Behav. 25 (1) (2021) 85–92, http://dx.doi.org/10.1007/s10461-020-02983-2.

[40] N. Mamode, Z. Ahmed, G. Jones, N. Banga, R. Motallebzadeh, H. Tolley, S. Marks, J. Stojanovic, M. A. Khurram, R. Thuraisingham, et al., Mortality rates in transplant recipients and transplantation candidates in a high-prevalence COVID-19 environment, Transplantation 105 (1) (2020) 212–215, http://dx.doi.org/10.1097/tp.0000000000003533.

[41] J. Jee, M.B. Foote, M. Lumish, A.J. Stonestrom, B. Wills, V. Narendra, V. Avutu, Y.R. Murciano-Goroff, J.E. Chan, A. Derkach, et al., Chemotherapy and COVID-19 outcomes in patients with cancer, J. Clin. Oncol. 38 (30) (2020) 3538–3546, http://dx.doi.org/10.1200/JCO.20.01307.

[42] A. Magdy Beshbishy, H.F. Hetta, D.E. Hussein, A.A. Saati, C. C. Uba, N. Rivero-Perez, A. Zaragoza-Bastida, M.A. Shah, T. Behl, G.E.-S. Batiha, Factors associated with increased morbidity and mortality of obese and overweight COVID-19 patients, Biology 9 (9) (2020) http://dx.doi.org/10.3390/biology9090280.

[43] C. Tang, Y. Wang, H. Lv, Z. Guan, J. Gu, Caution against corticosteroid-based COVID-19 treatment, Lancet 395 (10239) (2020) 1759–1760, http://dx.doi.org/10.1016/S0140-6736(20)30749-2.

[44] S. Tavakolpour, T. Rakhshandehroo, E.X. Wei, M. Rashidian, Lymphopenia during the COVID-19 infection: What it shows and what can be learned, Immunol. Lett. 225 (2020) 31, http://dx.doi.org/10.1016/j.imlet.2020.06.013.

[45] M. Rostami, H. Mansouritorghabeh, D-dimer level in COVID-19 infection: a systematic review, Expert Rev. Hematol. 13 (11) (2020) 1265–1275, http://dx.doi.org/10.1080/17474086.2020.1831383.

[46] L. Szarpak, K. Ruetzler, K. Safiejko, M. Hampel, M. Pruc, L. Kanczuga-Koda, K.J. Filipiak, M.J. Jaguszewski, Lactate dehydrogenase level as a COVID-19 severity marker, Am. J. Emerg. Med. (2020) http://dx.doi.org/10.1016/j.ajem.2020.05.073.

[47] M. Sepandi, M. Taghdir, Y. Alimohamadi, S. Afrashteh, H. Hosamirudsari, Factors associated with mortality in COVID-19 patients: A systematic review and meta-analysis, Iran. J. Publ. Health (2020) http://dx.doi.org/10.18502/ijph.v49i7.3574.

[48] M. Ahnach, S. Zbiri, S. Nejjari, F. Ousti, C. Elkettani, C-reactive protein as an early predictor of COVID-19 severity, J. Med. Biochem. 39 (4) (2020) 500, http://dx.doi.org/10.5937/jomb0-27554.

[49] M. Vargas-Vargas, C. Cortés-Rojo, Ferritin levels and COVID-19, Rev. Panam. de Salud PÚBlica 44 (2020) e72, http://dx.doi.org/10.26633/RPSP.2020.72.

[50] P. Du, J. Geng, F. Wang, X. Chen, Z. Huang, Y. Wang, Role of IL-6 inhibitor in treatment of COVID-19-related cytokine release syndrome, Int. J. Med. Sci. 18 (6) (2021) 1356, http://dx.doi.org/10.7150/ijms.53564.

[51] L. Sun, X.-Y. Zhang, Y. Qian, J.-C. Xu, S.-G. Zhang, Y. Tian, Joint neighborhood entropy-based gene selection method with fisher score for tumor classification, Appl. Intell. 49 (2019) http://dx.doi.org/10.1007/s10489-018-1320-1.

[52] A. Alzubaidi, G. Cosma, D. Brown, A.G. Pockley, Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information, in: 2016 International Conference on Interactive Technologies and Games (ITAG), IEEE, 2016, pp. 70–76, http://dx.doi.org/10.1109/iTAG.2016.18.

[53] W.S. Noble, What is a support vector machine? Nature Biotechnol. 24 (12) (2006) 1565–1567, http://dx.doi.org/10.1038/nbt1206-1565.

[54] Y.-Y. Song, L. Ying, Decision tree methods: applications for classification and prediction, Shanghai Arch. Psychiatry 27 (2) (2015) 130, http://dx.doi.org/10.11919/j.issn.1002-0829.215044.

[55] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.

[56] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN model-based approach in classification, in: OTM Confederated International Conferences" on the Move To Meaningful Internet Systems", Springer, 2003, pp. 986–996, http://dx.doi.org/10.1007/978-3-540-39964-3_62.

[57] F. Murtagh, Multilayer perceptrons for classification and regression, Neurocomputing 2 (5–6) (1991) 183–197, http://dx.doi.org/10.1016/0925-2312(91)90023-5.

[58] R.K. Vinayak, R. Gilad-Bachrach, Dart: Dropouts meet multiple additive regression trees, in: Artificial Intelligence and Statistics, PMLR, 2015, pp. 489–497, http://dx.doi.org/10.48550/arXiv.1505.01866.

[59] H. Reuter, J. Jordan, Status of hypertension in Europe, Curr. Opin. Cardiol. 34 (4) (2019) 342–349, http://dx.doi.org/10.1097/HCO.0000000000000642.

[60] T. Tamayo, J. Rosenbauer, S. Wild, A. Spijkerman, C. Baan, N. Forouhi, C. Herder, W. Rathmann, Diabetes in Europe: An update, Diabetes Res. Clin. Pract. 103 (2) (2014) 206–217, http://dx.doi.org/10.1016/j.diabres.2013.11.007.