Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 238–247

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Explainable learning to analyze the outcome of COVID-19 patients using clinical data

Daniel Olañeta[a,b], 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, A Coruña, Spain*
[b]*VARPA Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, 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, 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, A Coruña, Spain*

Abstract

Patients at high risk of contracting COVID-19 require specialized monitoring throughout their illness to ensure optimal treatment at each stage. To support this monitoring, Computer-Aided Diagnosis (CAD) methods analyze clinical data to estimate the most likely outcome for each patient, using various clinical variables such as symptoms, medical history, and laboratory results to predict outcomes. Despite the numerous proposals for COVID-19 diagnosis using CAD methods, the lack of explainability in many machine learning models poses a challenge in incorporating these methods into clinical practice. Additionally, other crucial tasks such as estimating the risk of death or severe forms of the disease must be considered to identify cases that require greater monitoring. To overcome these challenges, we propose an explainable methodology for estimating the risk of hospitalization and death in COVID-19 patients using clinical data. Our methodology employs four machine learning algorithms, three feature selection methods, and a decision tree to provide explainability. Our approach achieves an accuracy of 86.16% ± 0.74% for the estimation of hospitalization risk with 29 features, and an accuracy of 86.40% ± 1.80% for the estimation of the risk of death with 26 features. Moreover, our methodology provides valuable insights into the relationship between clinical variables and patient outcomes, which can inform more robust and informed clinical decision-making and improve our understanding of the disease. We demonstrate the potential of our transparent and effective CAD methods to support clinical decision-making in COVID-19 patient care and further research, offering a promising solution to overcome the challenges in incorporating CAD methods into clinical practice.

Keywords: feature selection; explainable machine learning; decision trees; outcome estimation; COVID-19

* Corresponding author. Tel.: +34 981167000 Ext. 6039
E-mail address: joaquim.demoura@udc.es

1877-0509
10.1016/j.procs.2023.10.008

1. Introduction

The COVID-19 pandemic has had a significant impact on public health worldwide, with over 300 million cases and 5 million deaths reported as of February 2023, according to WHO reports. COVID-19 is an acute infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that can severely affect the respiratory tract tissues [4]. In severe cases, patients may require hospitalization, mechanical ventilation, or admission to an Intensive Care Unit (ICU) [29]. Given the acute nature of this infection, it is essential to quickly evaluate the patient's condition and manage healthcare resources as efficiently and effectively as possible, particularly for patients who require more personalized monitoring.

Several studies have shown that pre-existing health conditions, such as diabetes, hypertension, and obesity, can increase the risk of severe illness and death in COVID-19 patients [22, 8, 31]. Additionally, clinical variables obtained from laboratory tests, such as blood tests, chest X-rays [16, 17], and CT scans, can provide valuable information to predict a patient's risk of severe illness or death [9]. In this context, Computer-Aided Diagnosis (CAD) methods, especially those supported by artificial intelligence (AI) strategies, can retrospectively analyze the correlation between pre-existing conditions, clinical variables, and patient outcomes [25]. These methods can not only help better understand the impact of variables from a clinical perspective but also facilitate the development of automatic methods that can estimate a patient's probable outcome based on their previous records.

However, one of the primary challenges in developing CAD methods is that many of them are based on machine learning (ML) algorithms with a black-box structure, making it difficult to include these methodologies in ordinary clinical practice [14]. The lack of explainability is one of the primary challenges with black-box ML models. In recent years, there has been an increased interest in explainable artificial intelligence (often abbreviated as XAI) [7] whose aim is to leverage different techniques to give machine learning pipelines the capability to explain the taken decisions, being the shapley additive explanations algorithm (denoted as SHAP) [15] one of the most popular strategies. Moreover, other white-box algorithms, like the decision trees, can help overcome this challenge by providing an automatically derived set of rules that explain the pathway followed by the model to make a decision, which can help clinicians to understand and accept the model's decisions. Another significant challenge in developing CAD methods is selecting the most appropriate set of features to characterize the problem. To address this issue, significant efforts have been made in the research community to propose automatic methods that can select the optimal subset of features for a given medical imaging problem, being specially useful when working with a considerable amount of features [5]. Decision trees can also be useful in this regard, as they provide a clear understanding of the relationship between different features and their impact on patient outcomes, helping to identify the most relevant features for accurate predictions.

Several studies have investigated methods for diagnosing, assessing the severity of, and predicting outcomes for COVID-19 patients and identifying those at higher risk. Apart from that, some of these works also include XAI methods in their pipelines to give explainability to the decisions made by the classification models. As reference, Bottrighi *et al.* [1] proposed a study that uses several ML models to estimate the risk of COVID-19 patients based on records with 43 different features, including demographics, chest X-ray and computerized tomography (CT) findings, complications, and treatments. Liu *et al.* [13] used omics-data and ML to perform COVID-19 diagnosis and severity prediction. Thimoteo *et al.* [24] used explainable boosting machine and logistic regression models to diagnose COVID-19 through blood test results. The authors also used black-box models, such as support vector machine and random forest, which are complemented by the SHAP algorithm to provide explainability in this context. Rostami *et al.* [20] developed a COVID-19 diagnosis method using blood test data that adds explainability with a decision tree. Weizman *et al.* [28] developed a scoring system based on an ML model to predict the outcome of hospitalized patients, including the risk of being admitted to the ICU and the risk of death. Another interesting contribution in this field is the work of Yagin *et al.*, [30] that uses the SHAP method to identify gene biomarkers that can help to determine a COVID-19 positive or negative. Despite the abundance of studies that have been conducted on COVID-19, there is still a pressing need for more comprehensive and interpretable models that can better estimate the risk of severe illness and death in COVID-19 patients. Previous studies have focused on different aspects of COVID-19 diagnosis, severity evaluation, and outcome prediction, but none of them have provided an exhaustive study of the risk estimation that includes explainability and feature selection analysis. The use of explainable models is particularly important for the successful adoption of these methods in clinical practice, as clinicians need to understand the reasoning behind the model's predictions to trust and accept them.

In this work, we make a significant contribution to the existing literature by providing a comprehensive study of risk estimation for COVID-19 patients. To achieve this, we utilize four state-of-the-art ML algorithms and three representative feature selection approaches to choose the optimal subset of features that can maximize the performance of the ML models. One of the most significant contributions of our work is the use of decision trees to increase the explainability of the ML models' decisions. This feature is particularly crucial for clinicians to understand and accept the model's decisions and successfully adopt this methodology, given the black-box nature of many ML models.

In particular, we conduct two representative analyses in our study. The first estimates the risk of hospitalization, while the second estimates the risk of death. These analyses provide critical information for allocating healthcare resources more efficiently and identifying patients who require more personalized monitoring and intensive care. In both scenarios, our study includes an exhaustive analysis of the optimal subset of features and a thorough understanding of how the number of these features affects the performance of the models with different feature selection methods. This information is essential for improving the accuracy of the risk estimation and enhancing the generalizability of the methodology to other medical applications.

We believe that our proposed methodology can offer valuable insights to clinicians, medical researchers, and healthcare professionals in the fight against the COVID-19 pandemic. By providing an explainable approach to risk estimation, our work can help clinicians better understand the impact of variables from a clinical perspective and facilitate better management of healthcare resources for those patients who require more personalized monitoring. Our proposed approach can also be generalized to other medical applications, where the development of explainable models can aid in the interpretation and acceptance of the model's predictions.

2. Dataset

The dataset used in this study was specifically designed for our research purposes and was provided by the Complexo Hospitalario Universitario de A Coruña (Galicia, Spain). It comprises records of 3217 COVID-19 patients (that corresponds with 3217 unique patients) and contains 29 variables related to clinical features, outcomes (survival or death), and cohort information (hospitalized or non-hospitalized). The clinical variables included in the dataset are age, sex, height (in centimeters), weight (in kilograms), and body mass index (BMI), which was calculated from the height and weight variables. In addition to these variables, the dataset also includes information about pre-existing health conditions that patients may have had, such as diabetes, arterial hypertension (AHT), chronic obstructive pulmonary disease (COPD), liver disease (LD), leukemia, lymphoma, neoplasm, asthma, human immunodeficiency virus (HIV), solid organ transplant (transplant), chemotherapy in the last 3 months, corticosteroids in the last 3 months, and biological treatment, including the type of biological treatment if applicable. Furthermore, the dataset includes the results of various clinical lab tests, such as lymphocyte count, percentage of lymphocytes, D-dimer test results, lactate dehydrogenase (LDH), creatinine, glomerular filtration rate (GFR), C-reactive protein (CRP), ferritin, and interleukin-6 (IL-6) protein test results. It is worth mentioning that the selection of the most relevant variables to characterize each patient for the proposed studies was done in accordance with the Head of Infectious Diseases Department of the aforementioned institution.

3. Methodology

we are able to match the conclusions drawn from the feature ranking with the clinical explainability provided by the decision tree

In this section, we present our explainable framework that combines a feature selection process with four types of ML models and a decision tree to add the explainability. The aim of separating the classification with high-accuracy black-box models and the decision trees (a glass-box model) is to match the clinical explainability provided by the decision trees with the discussion of the feature rankings. This balances the need for accuracy with the need for interpretability. An overview can be seen in Fig. 1. This proposal is divided into four different steps. In the 1st step, a data curation and balancing process is carried out. For the 2nd step, three feature selection algorithms were chosen to obtain the optimal subsets of features that will be then fed to the ML models. In the 3rd step, the 4 chosen ML models are trained. Finally, in the 4th step, the Decision Tree is trained. This trained model is analyzed afterward, to study the inferred rules that can give explainability about the decisions taken by the algorithm. Each step of our methodology is explained in more detail in the following subsections.

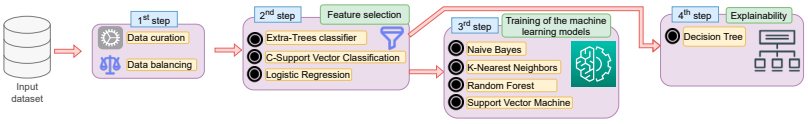

Fig. 1: Overview of the methodology that depicts the 4 followed steps. 1st step: data curation and balancing. 2nd step: feature selection with 3 different methods. 3rd step: training of 4 different ML models. 4th step: training of decision tree and explainability.

3.1. Data curation and balancing

As the dataset was obtained in a real clinical context, it is necessary to address some common issues that can arise. Firstly, it is important to distinguish between two types of variables: discrete and continuous. Discrete variables correspond to preconditions and treatments for each patient, and missing values were filled with a 0-padding in these cases. In contrast, missing values in continuous variables (e.g., height, weight) were filled with -1. Moreover, as another step of data curation, the variables are normalized to a range [0, 1] involving the maximum and the minimum of each one. Another important problem with the dataset is class imbalance. In the first analysis (Hospitalized vs Non-Hospitalized), the majority of cases correspond to the hospitalized class because patients are more likely to have data recorded when they have a severe condition. In the second analysis (Survival vs Death), the dataset is influenced by the mortality rate of COVID-19, where a majority of hospitalized patients survive and a smaller fraction die. To address class imbalance, we employed the SMOTE algorithm [26], which is a data augmentation strategy that synthesizes new samples from existing ones. SMOTE was chosen over simpler methods like duplicating samples of the minority class because it provides new and valuable information to the model, and can help improve the accuracy of the analysis. By using SMOTE, we were able to balance the dataset without losing valuable information. It is important to note that SMOTE can present important limitations. Firstly, it can increase the risk of overfitting, given that the algorithm generates new samples interpolating real data, that leads to a strong correlation between the original and the generated data. Moreover, there is also a risk that the generated data may deviate from the true distribution of the original data. Added to the previous points, other important problem is that, in case the original data presents noise or outliers, this could also be reflected in the new generated data. Considering these potential issues, it is important to ensure that the trained classification models are properly evaluated. In particular, there is no oversampling applied to the test set, to ensure that the models' performance is assessed on real-world data.

3.2. Feature selection

Once we have a properly curated and balanced dataset, we employ three commonly used state-of-the-art algorithms, based on machine learning models, to assign a weight to each feature: Extra-Trees Classifier (ETC), C-Support Vector Classification (C-SVC), and Logistic Regression (LR). Using three different scoring algorithms provides a more robust understanding of the most relevant features, as having multiple perspectives can reinforce the discussion. In particular, the 3 different scoring algorithms help to ensure that there is an agreement between the methods regarding the most relevant features, that represents an alternative from relying on a single ranking. Additionally, the feature selection process is useful for reducing the dimensionality of the data, removing redundant and irrelevant features that could introduce noise to the models. To this end, we used a Recursive Feature Elimination (RFE) strategy [12] to obtain smaller feature sets recursively by removing the least important features, finally building a ranking from the most to the least important. In this context, the classifiers are used to evaluate the importance of each individual feature. This is measured depending on the impact that the features have on the performance of the model. The three feature scoring methods used in this study are explained in detail below:

1. **Extra-Trees Classifier (ETC)** [10]: This method randomly selects several thresholds for each feature and chooses the most appropriate one as the split rule. This approach helps to avoid overfitting.
2. **C-Support Vector Classification (C-SVC)** [19]: This algorithm is based on Support Vector Machines and separates data points by searching for an optimal hyperplane that maximizes the margins between the two classes.
3. **Logistic Regression (LR)** [3]: This method estimates, for each sample, the probability of belonging to a specific class based on a set of independent variables.

3.3. Training of the ML models

In this step, the ML models are trained using the optimal subset of features obtained in the previous step. To provide a more thorough analysis, we have chosen four different common state-of-the-art algorithms: Naive Bayes (NB), k-Nearest Neighbors (kNN), Random Forest (RF), and Support Vector Machine (SVM). This thorough analysis helps to assess the generalizability and robustness of the selected features with different modeling techniques, also minimizing the risk of introducing biases that could appear from tailoring those features to a particular classifier. Thus, this ensures the independence between the 2 main steps of the methodology (feature selection and classification).

1. **Naive Bayes (NB)** [27]: This model is a probabilistic classifier based on the Bayes theorem that assumes strong independence among the features that characterize the problem.
2. **k-Nearest Neighbors (kNN)** [11]: This classifier stores the samples during the training phase, transferring almost the full weight of the computation to inference. During inference, when the model is presented with a new sample whose class is unknown, the distance between that sample and the samples stored during the training phase is calculated, selecting the k smallest distances. Finally, the input sample will be classified with the same class that has the majority of those k neighbors. To avoid the situation where k/2 samples belong to one class and the other k/2 samples belong to the other class, k must be an odd number.
3. **Random Forest (RF)** [2]: This method builds a forest composed of several decision trees during training time. Then, the selected output at the inference stage is the most-voted class among all those decision trees.
4. **Support Vector Machine (SVM)** [18]: This ML algorithm searches for the optimal hyperplane that maximizes the distance between two classes. Once a new sample is fed to this model, the position on one side or another of the hyperplane will determine its classification.

3.4. Explainability

In the fourth step, we train a Decision Tree [21] model to provide explainability to the decisions made by the algorithm, which is one of the strongest points of our work. The aim is to automatically infer a set of rules that can be extremely useful for clinicians, as it helps them understand not only the most important features but also the relationship between them and their contribution to deriving the outcome of the patient. To avoid creating excessively large decision trees, we train the decision tree with only the top 5 features selected by the corresponding feature selection method. With this, we also avoid the creation of unnecessary rules caused by the existence of outliers or mislabeled samples. Gini [23] and entropy [6] were chosen as the splitting criteria, and the optimum method was obtained using cross-validated grid-search. The visualization of the trained decision tree displays the rule that must be followed at each node. For example, if GFE ≤ 0.039 is the rule, the left branch is followed if the condition is True and the right branch if it is False. The value of the selected criteria (gini or entropy), the percentage of samples available at that point of the tree, the ratio of samples for each class, and the majority class are also displayed.

4. Results and Discussion

This section presents the results and discussion for the two provided analyses. The first analysis estimates the risk of a patient requiring hospitalization (being Non-Hospitalized the positive class and Hospitalized the negative class), while the second analysis estimates the risk of death (with Death as the positive class and Survival as the negative class). Firstly, the performance obtained by the chosen classifiers regarding the different feature selection methods and the number of features that compose the corresponding optimal subset, the feature rankings to discuss the most relevant variables according to the 3 feature selection methods, and the evolution of the performance when increasing the number of features are analyzed. Finally, the explainability of the trained decision trees is also discussed for both analyses. For experimentation purposes, the models are evaluated with the most appropriate classification metrics, including Accuracy, Recall, Precision, and F1-Score. It is important to consider that accuracy will be biased with the imbalanced problem proposed in this work, artificially inflating the actual performance. To overcome this issue, we have also included the recall and precision, that evaluate the actual performance for the positive class, without biases. It is also worth noting that the dataset was split using a 10-fold cross-validation to ensure that training and validation sets are independent. As the training process can be performed 10 times, this makes it possible to provide the mean and the standard deviation of the metrics.

Analysis I: Estimation of the risk of hospitalization. The global results in terms of accuracy for analysis I are shown in Table 1. In this case, the RF algorithm achieves the highest performance regardless of the chosen feature

Table 1: Overall comparison of the used classifiers and feature selection methods for the analysis I, where the metrics of the approach with the highest accuracy are highlighted in gray. The performance can be compared regarding the accuracy and the number of features of each approach.

Classifier	Feature selection method	# of features	Accuracy	Recall	Precision
NB	ETC	25	72.99% ± 1.13%	94.51% ± 2.13%	66.06% ± 1.16%
	C-SVC	13	72.82% ± 1.00%	95.37% ± 1.35%	65.72% ± 1.13%
	LR	7	72.48% ± 1.52%	94.88% ± 1.43%	65.52% ± 1.40%
kNN	ETC	19	80.66% ± 1.30%	91.14% ± 1.52%	75.37% ± 1.93%
	C-SVC	29	80.97% ± 1.75%	92.19% ± 1.62%	75.32% ± 2.00%
	LR	29	80.34% ± 1.80%	91.19% ± 1.74%	74.95% ± 1.85%
RF	ETC	28	86.08% ± 1.18%	89.14% ± 1.98%	84.02% ± 1.60%
	C-SVC	28	85.78% ± 0.97%	89.00% ± 1.68%	83.64% ± 1.12%
	LR	29	86.16% ± 0.74%	89.52% ± 1.33%	83.91% ± 1.43%
SVM	ETC	25	68.71% ± 1.80%	67.71% ± 2.95%	69.13% ± 2.53%
	C-SVC	20	77.19% ± 1.99%	87.35% ± 2.29%	72.60% ± 1.88%
	LR	23	77.38% ± 1.55%	87.57% ± 1.85%	72.74% ± 1.66%

Table 2: Feature ranking for analysis I, comparing the results obtained by the 3 feature selection methods.

Position	*Feature selection method*			*Position (cont.)*	*Feature selection method (cont.)*		
	ETC	**C-SVC**	**LR**		**ETC**	**C-SVC**	**LR**
1	Creatinine	Neoplasm	Neoplasm	**16**	Neoplasm	LD	Creatinine
2	Age	Diabetes	Diabetes	**17**	Height	Lymphocytes	Lymphocytes (%)
3	LDH	CCS	COPD	**18**	Asthma	CRP	Lymphocytes
4	Lymphocytes	COPD	CCS	**19**	BMI	Chemotherapy	CRP
5	AHT	IL-6	GFR	**20**	COPD	AgeRange	AgeRange
6	CRP	Asthma	Asthma	**21**	LD	Sex	LDH
7	GFR	AHT	IL-6	**22**	CCS	LDH	Height
8	Lymphocytes (%)	Lymphoma	Lymphoma	**23**	Lymphoma	Leukemia	Sex
9	Ferritine	GFR	AHT	**24**	Transplant	Biological	Biological
10	AgeRange	Transplant	Transplant	**25**	Biological Type	Height	Chemotherapy
11	Weight	HIV	LD	**26**	Chemotherapy	Biological Type	Biological Type
12	Diabetes	Ferritine	Leukemia	**27**	Leukemia	D-Dimer	D-Dimer
13	D-Dimer	Creatinine	BMI	**28**	Biological	Age	HIV
14	IL-6	Lymphocytes (%)	Weight	**29**	HIV	Weight	Age
15	Sex	BMI	Ferritine				

selection algorithm, with the lowest accuracy of 85.78% ± 0.97% and the highest accuracy of 86.16% ± 0.74%. The second highest-performing configuration is the kNN, with the lowest accuracy of 80.34% ± 1.80% and the highest accuracy of 80.97% ± 1.75%. It is worth noting the case of SVM, which obtains satisfactory performances with Logistic Regression and Linear SVC as the feature selection methods, achieving the highest accuracy of 77.38% ± 1.55% in the first case, but drops when using the Extra-Trees Classifier with an accuracy of 68.71% ± 1.80%. Finally, NB shows consistency among the 3 feature selection methods, with the lowest accuracy of 72.48% ± 1.52% and the highest accuracy of 72.99% ± 1.13%. Overall, it can be concluded that the highest-performing approach uses Random Forest as a classifier and Logistic Regression as a feature selector.

The feature ranking for analysis I is shown in Table 2. The ExtraTreesClassifier method produces a different ranking in the top positions compared to LinearSVC and LogisticRegression. However, there is an agreement between LinearSVC and LogisticRegression in these positions, with Neoplasm, Diabetes, COPD, CCS, Asthma, Lymphoma, GFR, and AHT among the top 10 features. In the case of the ExtraTreesClassifier, AHT is also given high importance (position 5), but the method assigns high scores to other features such as Creatinine, Age, LDH, and variables related to the amount of Lymphocytes. It is noteworthy that the ExtraTreesClassifier gives great importance to Age and AgeRange variables to estimate the risk of hospitalization, while they are less relevant in the other two feature selection methods.

To examine the impact of the number of features on model performance, we selected the highest performing approach, which was Random Forest with Logistic Regression as the feature selection method. Fig. 2 shows the evolution of performance as the number of features increases. Overall, the performance tends to improve as the number of features increases, stabilizing around 20 features despite a slight improvement with the entire set of 29 features. For explainability, we considered the approach with the highest overall performance, selecting the decision tree built with Logistic Regression as the feature selection method. Fig. 3 and Fig. 4 show the half right and half left parts of the decision tree, respectively. The most relevant features in the half right part of the decision tree are GFR, Diabetes and Neoplasm. In particular, GFR is the most important feature in the half right part of the decision tree, as it is used multiple times to discriminate between Hospitalized and Non-Hospitalized. In this part of the tree, Diabetes is also present in one node, and Neoplasm appears in two cases. The most significant variables in the half left part of the tree are directly related to the immune system and breathing capacity of patients. In particular, the most important variables are Diabetes, CCS, Neoplasm, and COPD. This is notable because COVID-19 typically affects immunocompromised patients more severely, and COPD significantly increases the risk of hospitalization, as the disease severely affects the lungs. However, the number of occurences of each variable only partially determines

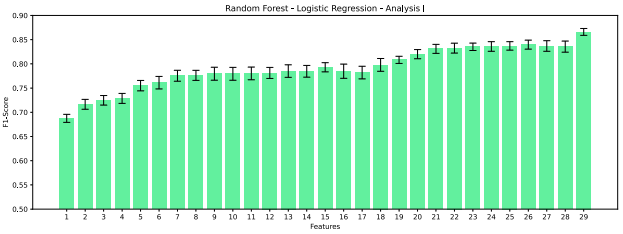

Fig. 2: Evolution of the performance given the number of features in the analysis I, considering the approach with the highest accuracy (Random Forest classifier with Logistic Regression as feature selector).

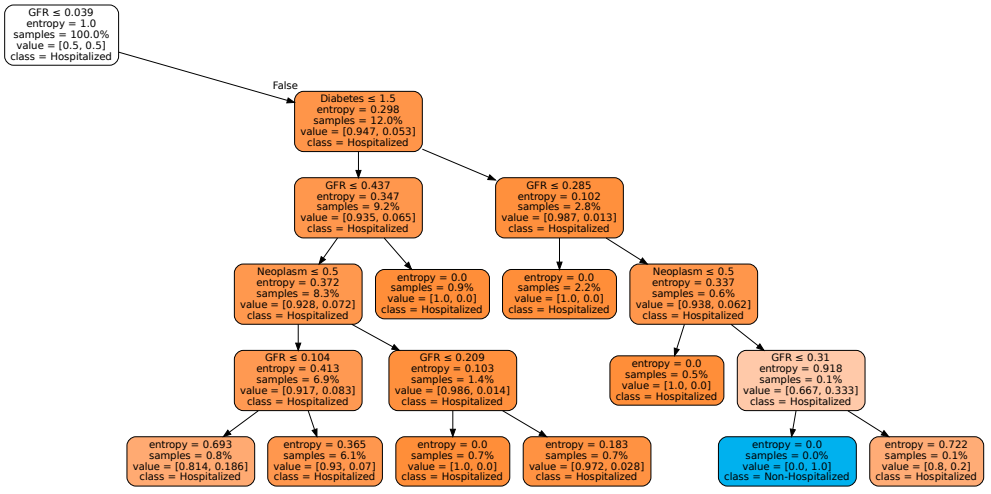

Fig. 3: Half right of the Decision Tree that has been considered to add explainability to Analysis I.

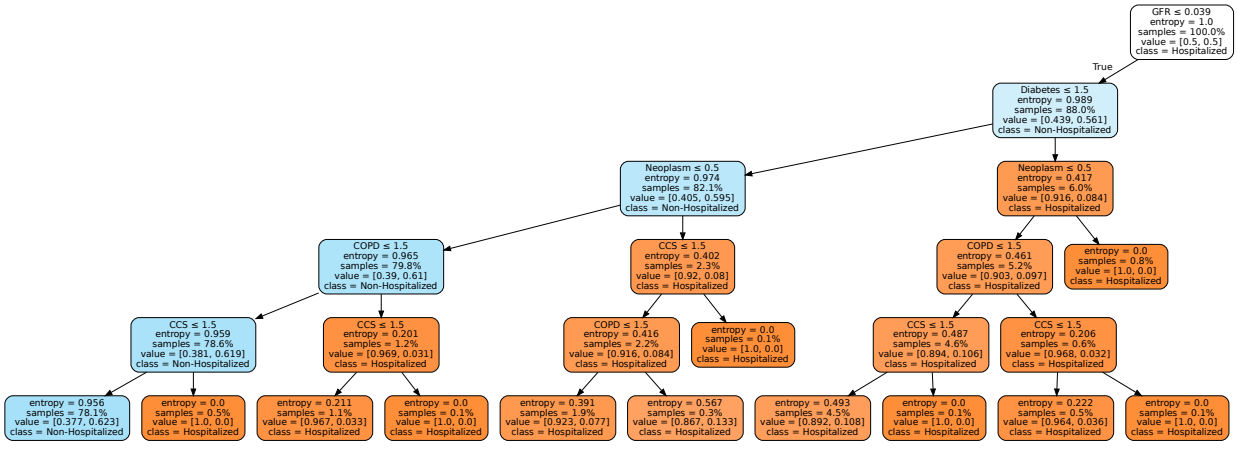

Fig. 4: Half left of the Decision Tree that has been considered to add explainability to Analysis I.

their actual importance. This can be complemented with the position of the nodes in the tree. Particularly, GFR and Diabetes and Neoplasm show to have high discrimination capabilities, being positioned in the shallowest levels.

Analysis II: estimation of the risk of death. Table 3 shows the results obtained in terms of accuracy for the analysis I, regarding the used classifier and the feature selection method, also providing the optimal number of features obtained in each case. There, it can be seen that the best performance is achieved by the Random Forest algorithm,

Table 3: Accuracy and number of features comparison given the approaches chosen for experimentation in analysis II. The metrics of the approach with the highest accuracy are highlighted in gray.

Classifier	Feature selection method	# of features	Accuracy	Recall	Precision
NB	ETC	2	72.14% ± 1.40%	81.42% ± 1.95%	68.66% ± 2.53%
	C-SVC	29	69.88% ± 2.26%	85.10% ± 2.38%	65.24% ± 2.00%
	LR	29	70.52% ± 2.31%	84.98% ± 2.55%	65.92% ± 1.97%
kNN	ETC	19	83.29% ± 1.24%	89.96% ± 1.20%	79.36% ± 2.39%
	C-SVC	29	82.58% ± 1.84%	90.36% ± 1.88%	78.20% ± 2.70%
	LR	29	83.06% ± 1.48%	90.97% ± 1.91%	78.52% ± 2.46%
RF	ETC	17	86.03% ± 1.55%	90.39% ± 2.16%	83.10% ± 1.82%
	C-SVC	28	86.15% ± 2.38%	90.36% ± 3.29%	83.28% ± 2.07%
	LR	26	86.40% ± 1.80%	90.58% ± 2.98%	83.64% ± 2.09%
SVM	ETC	19	74.06% ± 2.19%	84.80% ± 1.83%	69.79% ± 3.13%
	C-SVC	28	73.88% ± 2.79%	82.33% ± 2.87%	70.42% ± 3.38%
	LR	25	74.46% ± 2.17%	83.28% ± 1.86%	70.81% ± 3.06%

Table 4: Feature ranking for analysis II, showing the results obtained by the 3 feature selection methods.

Position	*Feature selection method*			*Position (cont.)*	*Feature selection method (cont.)*		
	ETC	**C-SVC**	**LR**		**ETC**	**C-SVC**	**LR**
1	Lymphocytes (%)	Lymphocytes	Lymphocytes	**16**	COPD	Biological	Leukemia
2	Age	Lymphocytes (%)	Lymphocytes (%)	**17**	Diabetes	LDH	Biological
3	Creatinine	Height	Height	**18**	BMI	CCS	HIV
4	Lymphocytes	BMI	BMI	**19**	Asthma	Ferritine	Diabetes
5	AgeRange	Asthma	Asthma	**20**	IL-6	IL-6	Lymphoma
6	LDH	Creatinine	Creatinine	**21**	CCS	D-Dimer	D-Dimer
7	CRP	Transplant	Transplant	**22**	LD	CRP	Ferritine
8	D-Dimer	LD	LD	**23**	Lymphoma	HIV	IL-6
9	GFR	COPD	COPD	**24**	Chemotherapy	Neoplasm	CRP
10	Ferritine	Lymphoma	Chemotherapy	**25**	Transplant	Diabetes	Age
11	Weight	AHT	GFR	**26**	Biological Type	Age	AgeRange
12	Sex	Leukemia	Weight	**27**	Leukemia	AgeRange	AHT
13	Height	GFR	CCS	**28**	Biological	Biological Type	Biological Type
14	AHT	Chemotherapy	LDH	**29**	HIV	Weight	Neoplasm
15	Neoplasm	Sex	Sex				

using the Logistic Regression as feature selection method, with an accuracy of 86.40% ± 1.80% and with an optimum number of 26 features. In particular, the Random Forest algorithm demonstrates a high performance regardless of the feature selection method, with a minimum mean accuracy of 86.03% and an optimal subset of 17 features. The next best performing method is the algorithm kNN, with the lowest mean accuracy of 82.58% ± 1.84% and the highest mean accuracy of 83.29% ± 1.24%. Finally, this ranking is completed by the SVM and the NB, with a performance that is notably lower than the other 2 methods. In particular, the lowest performance of the SVM is 73.88% ± 2.79% and the highest performance of 74.46% ± 2.17%. In the case of NB, the lowest accuracy is 69.88% ± 2.26% while the highest accuracy is 72.14% ± 1.40%.

The evaluation of how the number of features affects the performance can be seen in Fig. 5 using as reference the optimal approach for this analysis (with Random Forest as classifier and Logistic Regression as feature selection method). The evolution depicts a trend of improvement that converges from 25 features onward.

Regarding the feature rankings, they can be found in Table 4. For this analysis, there is greater agreement between ExtraTreesClassifier and the other two feature selection methods, with Lymphocytes and Lymphocytes (%) ranked among the most relevant features, but there are still significant differences. Looking at the rankings produced by LinearSVC and LogisticRegression, the variables related to the amount of Lymphocytes, BMI, Creatinine, Transplant, COPD, LD, and Creatinine are among the top 10. In the case of ExtraTreesClassifier, in addition to Lymphocytes, Age, AgeRange, Creatinine, LDH, CRP, D-Dimer, and Ferritine are also ranked in the top 10. It is worth noting that ExtraTreesClassifier gives a high importance to Age and AgeRange, whereas LinearSVC and LogisticRegression rank these variables significantly lower, as in the case of Analysis I.

For this second analysis, the set of rules of the trained decision tree can be seen in Fig. 6. In this case, the root node starts discriminating between those patients with and without missing values for Lymphocytes (%). In particular, the mentioned percentage of Lymphocytes and the absolute count appear several times in the nodes of the decision tree. Nevertheless, other variables as Asthma, BMI and Height are considered by the decision tree to build the rules. In the case of Asthma, this precondition is relevant given that it directly affects the breathing capacity of the patients, worsening the risk of contracting severe COVID-19. Regarding the variable BMI, it is well-known that overweight and obesity are risk factors for many pathological conditions. In a similar way as in the previous analysis, the location of the variables is also relevant having that the percentage of Lymphocytes is one of the most discriminative features, given that it is placed high in the decision tree, as well as the variables Asthma and BMI.

5. Conclusions

In this work, we have proposed an explainable methodology to identify patients at a higher risk of hospitalization (Hospitalized/Non-Hospitalized scenario) or death (Survival/Death scenario) in two different analyses. The explain-

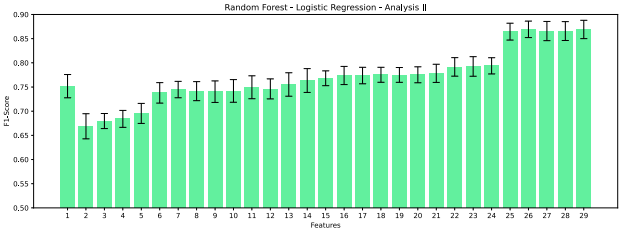

Fig. 5: Evolution of the performance regarding the number of features in the analysis II, taking the approach with the highest accuracy (Random Forest classifier with Logistic Regression as feature selector).

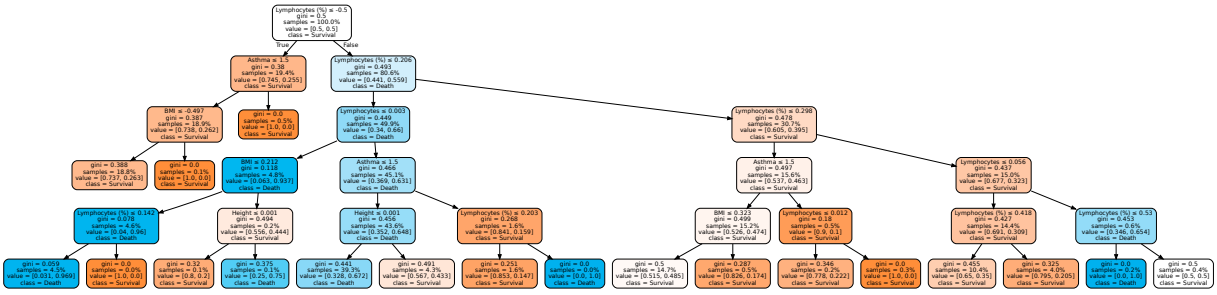

Fig. 6: Decision Tree that has been trained to add explainability to Analysis II.

ability of this methodology can help clinicians to better understand the disease and make more informed decisions based on retrospective data, which is also important for the potential inclusion of the methodology in clinical practice. For this purpose, we selected four different ML algorithms (Naive Bayes, k-Nearest Neighbors, Random Forest, and Support Vector Machine), three feature selection methods (Extra-Trees Classifier, Linear Support Vector Classification, and Logistic Regression), and an additional Decision Tree to provide explainability. We also performed an exhaustive analysis to find the optimal subset of features. The results demonstrate that the proposed methodology is suitable for the presented problem, with an accuracy of 86.16% ± 0.74% for identifying patients at high risk of hospitalization and 86.40% ± 1.80% for identifying patients at high risk of death. Regarding explainability, the analysis I shows the great importance of Glomerular Filtration Rate, Neoplasm, Diabetes, Chronic Obstructive Pulmonary Disease, and Corticosteroids (variables that are directly related to immunity and breathing capacity) in identifying patients at high risk of hospitalization. In the case of analysis II, the absolute count and percentage of Lymphocytes are notably relevant variables to determine the risk of death of a COVID-19 patient. Asthma (directly related to the breathing capacity of the patients), BMI, and Height (which determine if a patient is overweight) are also relevant. These analyses could be complemented using other sources of data in future works to improve the performance of the classification models and add more clinical features to the decisions taken. Moreover, we could also include other frameworks of explainability, like the SHAP algorithm, to complement the evaluation of the methodology.

Acknowledgements

This work was supported by Ministerio de Ciencia e Innovación, Government of Spain through the research project with [grant numbers RTI2018-095894-B-I00, PID2019-108435RB-I00, TED2021-131201B-I00, and PDC2022-133132-I00]; Consellería de Educación, Universidade, e Formación Profesional, Xunta de Galicia, Grupos de Referencia Competitiva, [grant number ED431C 2020/24], predoctoral grant [grant number ED481A 2021/196]; CITIC, Centro de Investigación de Galicia [grant number ED431G 2019/01], receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

References

[1] Bottrighi, A., et al., 2022. A machine learning approach for predicting high risk hospitalized patients with COVID-19 SARS-cov-2. BMC Medical Informatics and Decision Making 22. doi:10.1186/s12911-022-02076-1.

[2] Breiman, L., 2001. Random forests. Machine learning 45, 5–32. doi:10.1023/A:1010933404324.

[3] Cheng, Q., Varshney, P.K., Arora, M.K., 2006. Logistic regression for feature selection and soft classification of remote sensing data. IEEE Geoscience and Remote Sensing Letters 3, 491–494. doi:10.1109/LGRS.2006.877949.

[4] Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W.C., Wang, C.B., Bernardini, S., 2020. The covid-19 pandemic. Critical reviews in clinical laboratory sciences 57, 365–388. doi:10.1080/10408363.2020.1783198.

[5] de Moura, J., Novo, J., Ortega, M., 2019. Deep feature analysis in a transfer learning-based approach for the automatic identification of diabetic macular edema, in: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. doi:10.1109/IJCNN.2019.8852196.

[6] Du, M., Wang, S.M., Gong, G., 2011. Research on decision tree algorithm based on information entropy. Advanced Materials Research 267, 732–737. doi:10.4028/www.scientific.net/amr.267.732.

[7] Dwivedi, R., et al., 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys 55, 1–33. doi:10.1145/3561048.

[8] Fang, L., Karakiulakis, G., Roth, M., 2020. Are patients with hypertension and diabetes mellitus at increased risk for covid-19 infection? The lancet respiratory medicine 8, e21. doi:10.1016/S2213-2600(20)30116-8.

[9] Gao, Y., Li, T., Han, M., Li, X., Wu, D., Xu, Y., Zhu, Y., Liu, Y., Wang, X., Wang, L., 2020. Diagnostic utility of clinical laboratory data determinations for patients with the severe covid-19. Journal of medical virology 92, 791–796. doi:10.1002/jmv.25770.

[10] Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine Learning 63, 3–42. doi:10.1007/s10994-006-6226-1.

[11] Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003. KNN model-based approach in classification, in: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, pp. 986–996. doi:10.1007/978-3-540-39964-3_62.

[12] Li, F., Yang, Y., 2005. Analysis of recursive feature elimination methods, in: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM. doi:10.1145/1076034.1076164.

[13] Liu, X., Hasan, M.R., Ahmed, K.A., Hossain, M.Z., 2023. Machine learning to analyse omic-data for COVID-19 diagnosis and prognosis. BMC Bioinformatics 24. doi:10.1186/s12859-022-05127-6.

[14] London, A.J., 2019. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Center Report 49, 15–21. doi:10.1002/hast.973.

[15] Lundberg, S., Lee, S.I., 2017. A unified approach to interpreting model predictions. doi:10.48550/ARXIV.1705.07874.

[16] Morís, D.I., de Moura, J., Novo, J., Ortega, M., 2021. Data augmentation approaches using cycle-consistent adversarial networks for improving COVID-19 screening in portable chest x-ray images. Expert Systems with Applications 185, 115681. doi:10.1016/j.eswa.2021.115681.

[17] Morís, D.I., de Moura, J., Novo, J., Ortega, M., 2021. Cycle generative adversarial network approaches to produce novel portable chest x-rays images for covid-19 diagnosis, in: ICASSP 2021, pp. 1060–1064. doi:10.1109/ICASSP39728.2021.9414031.

[18] Noble, W.S., 2006. What is a support vector machine? Nature Biotechnology 24, 1565–1567. doi:10.1038/nbt1206-1565.

[19] Novakovic, J., Veljovic, A., 2011. C-support vector classification: Selection of kernel and parameters in medical diagnosis, in: 2011 IEEE 9th international symposium on intelligent systems and informatics, IEEE. pp. 465–470. doi:10.1109/SISY.2011.6034373.

[20] Rostami, M., Oussalah, M., 2022. A novel explainable COVID-19 diagnosis method by integration of feature selection with random forest. Informatics in Medicine Unlocked 30, 100941. doi:10.1016/j.imu.2022.100941.

[21] Safavian, S., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics 21, 660–674. doi:10.1109/21.97458.

[22] Singh, A.K., Gupta, R., Ghosh, A., Misra, A., 2020. Diabetes in covid-19: Prevalence, pathophysiology, prognosis and practical considerations. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14, 303–310. doi:10.1016/j.dsx.2020.04.004.

[23] Tangirala, S., 2020. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications 11. doi:10.14569/ijacsa.2020.0110277.

[24] Thimoteo, L.M., Vellasco, M.M., Amaral, J., Figueiredo, K., Yokoyama, C.L., Marques, E., 2022. Explainable artificial intelligence for COVID-19 diagnosis through blood test variables. JCAES 33, 625–644. doi:10.1007/s40313-021-00858-y.

[25] Vaishya, R., Javaid, M., Khan, I.H., Haleem, A., 2020. Artificial intelligence (ai) applications for covid-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14, 337–339. doi:10.1016/j.dsx.2020.04.012.

[26] Wang, J., Xu, M., Wang, H., Zhang, J., 2006. Classification of imbalanced data by using the smote algorithm and locally linear embedding, in: 2006 8th international Conference on Signal Processing. doi:10.1109/ICOSP.2006.345752.

[27] Webb, G.I., Keogh, E., Miikkulainen, R., Miikkulainen, R., Sebag, M., 2011. Naïve bayes, in: Encyclopedia of Machine Learning. Springer US, pp. 713–714. doi:10.1007/978-0-387-30164-8_576.

[28] Weizman, O., et al., 2022. Machine learning-based scoring system to predict in-hospital outcomes in patients hospitalized with COVID-19. Archives of Cardiovascular Diseases 115, 617–626. doi:10.1016/j.acvd.2022.08.003.

[29] Wunsch, H., 2020. Mechanical ventilation in covid-19: interpreting the current epidemiology. doi:10.1164/rccm.202004-1385ED.

[30] Yagin, F.H., Cicek, İ.B., Alkhateeb, A., Yagin, B., Colak, C., Azzeh, M., Akbulut, S., 2023. Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. Computers in Biology and Medicine 154, 106619. doi:10.1016/j.compbiomed.2023.106619.

[31] Yu, W., Rohli, K.E., Yang, S., Jia, P., 2021. Impact of obesity on covid-19 patients. Journal of Diabetes and its Complications 35, 107817. doi:10.1016/j.jdiacomp.2020.107817.