IM - ORIGINAL

Mortality predictors in patients with COVID-19 pneumonia: a machine learning approach using eXtreme Gradient Boosting model

N. Casillas[1,2] · A. M. Torres[2] · M. Moret[1] · A. Gómez[1] · J. M. Rius-Peris[2,3] · J. Mateo[2]

Abstract
Recently, global health has seen an increase in demand for assistance as a result of the COVID-19 pandemic. This has prompted many researchers to conduct different studies looking for variables that are associated with increased clinical risk, and find effective and safe treatments. Many of these studies have been limited by presenting small samples and a large data set. Using machine learning (ML) techniques we can detect parameters that help us to improve clinical diagnosis, since they are a system for the detection, prediction and treatment of complex data. ML techniques can be valuable for the study of COVID-19, especially because they can uncover complex patterns in large data sets. This retrospective study of 150 hospitalized adult COVID-19 patients, of which we established two groups, those who died were called Case group ($n = 53$) while the survivors were Control group ($n = 98$). For analysis, a supervised learning algorithm eXtreme Gradient Boosting (XGBoost) has been used due to its good response compared to other methods because it is highly efficient, flexible and portable. In this study, the response to different treatments has been evaluated and has made it possible to accurately predict which patients have higher mortality using artificial intelligence, obtaining better results compared to other ML methods.

Introduction

SARS-CoV-2 is a betacoronavirus belonging to Coronaviridae family, whose first detection was in December 2019 in Wuhan, China. Its pathogenesis is associated with extensive lung damage that generates a hypoxemic state in the individual who contracts it. It's on the host where the activation of different inflammatory pathways occurs, presenting a cytokines overproduction and causing a chaotic inflammatory response and coagulation disorders. However, not all patients affected by SARS-CoV-2 will develop this severe form of the disease, since a high percentage of patients will suffer from a mild or even asymptomatic illness [1]. On the one hand, the presence of a pro-inflammatory state generated edema and cellular infiltration at the lung level, producing a diffuse interstitial pattern and as consequence development of acute respiratory distress syndrome (ARDS) [2]. Elevation of certain circulating cytokines has been observed such as: interleukin-6 (Il-6), interlukine-2 (Il-2), Tumor necrosis factor-alpha (TNF-alpha), and colonies (G-CSF). Il-6 play an essential role in the pathogenesis of the respiratory syndrome due to SARS-CoV-2 since, in addition to being a pro-inflammatory cytokine, it is involved in modulating the initiation of coagulation, increasing the platelet and leukocyte adhesion. In addition, it has been proposed as a prognostic marker of the disease and predictor of long-term pulmonary fibrotic sequelae in patients with persistently elevated Il-6 [3–5]. That is why drugs that targeted this interleukin or its cellular receptors, such as tocilizumab or baricitinib, were quickly used. In a retrospective multicentre study whose main objective was the need for intubation and/or mortality, the use of tocilizumab monotherapy was compared with corticosteroids at intermediate doses, pulses of corticosteroids

✉ N. Casillas
  nazaretcasillas@hotmail.es

  M. Moret
  marta.moretrueda@gmail.com

  A. Gómez
  adrgom02@gmail.com

1. Departament of Internal Medicine, Hospital Virgen de la Luz, Cuenca, Spain

2. Neurobiological Research Group, Institute of Technology, Castilla-La Mancha University, Cuenca, Spain

3. Departament of Pediatrics, Hospital Virgen de la Luz, Cuenca, Spain

or the combination of pulses of corticosteroids with toci-lizumab, without finding significant differences between the various arms [6–8]. Although it is true, that the initiation of high doses of corticosteroids was conditioned to the development of the hyperinflammatory state, which could translate into a late initiation of immunosuppressive treatment [9]. Regarding, studies published with baricitinib in combination with remdesivir or intravenous corticosteroids, such as Dexamethasone or Methylprednisolone, have shown benefits in terms of shorter recovery time in those with high-flow O2 supplements and non-invasive mechanical ventilation (NIMV), without evaluating its effect in monotherapy. Initially, steroids have been evaluated in multiple studies due to their anti-inflammatory activity whose objective was the cessation of the pro-inflammatory cascade whose direct consequence was ARDS, concluding its no effect on the reduction of mortality although it did decrease hypoxia and the risk of respiratory failure. [10–13]. There are mixed results regarding the use of corticosteroids. However, the latest publications conclude that the use of corticosteroids has shown a reduction in mortality in moderate-severe patients who require supplemental O2 or invasive mechanical ventilation (IMV). While in mild patients a harmful effect cannot be ruled out [14].

On the other hand, SARS-CoV-2's high mortality is closely related to the occurrence of coagulation disorders. Disseminated intravascular coagulation (DIC) is the most frequently alteration detected. At the beginning of SARS-CoV-2 infection, a hypercoagulatory state occurs with an increase in macro- and micro-vascular thrombotic events, preferably venous, although they have also been observed at the arterial level. However, the consumption of platelets as well as coagulation factors generates a chaotic state that favors the appearance of bleeding events that have been carried out with an increase in mortality in critically ill patients. Several studies have confirmed that the intensification of anticoagulant treatment has not been related to a lower incidence of thrombosis, while it has generated an increase in bleeding events [15]. While other authors argue that prophylactic anticoagulation doses are insufficient in high-risk patients [16]. Certainly, it is necessary to perform a correct initial evaluation of the coagulation status of patients as well as to carry out periodic analytical follow-ups during hospitalization with: prothrombin time (PT), thromboplastin time (APTT), platelets, antithrombin (AT), D-dimer (DD), and degradation products of the fibrin (FDP) [17–21]. DD has been one of the most studied parameters in the COVID-19 population. It's a fibrin degradation product that results from both the conversion of fibrinogen to fibrin by thrombin and by the cross-linking of fibrin by activated factor XIII and by the degradation of fibrin by plasmin. It is therefore part of coagulation and the activation of fibrinolysis, and for this reason it can be elevated in thrombotic and haemorrhagic events [22]. Various studies have reported an elevated level of DD in a large proportion of patients and a progressive elevation has even been observed in those patients who develop ARDS and therefore a more serious disease. During the early days of the pandemic, DD levels marked the start of treatment with low-weight heparins at anticoagulant doses to prevent thrombus formation in these patients. This led to the appearance of haemorrhagic events, which could be related to the situation of clinical severity such as a higher dosage of anticoagulant treatment. Therefore, DD can be a predictor of severity, even mortality.

Our work is based on the classification of patients with COVID-19 and the objective has been to evaluate the response to the different treatments and to predict which patients were at higher risk of developing severe respiratory distress syndrome. For this we have implemented an extreme gradient boosting algorithm (XGB), which in addition to being a supervised algorithm, is a variant of gradient boosting and which aims to increase both speed as running in ML model. It is a decision tree based ML algorithm that uses a gradient reinforcement framework. The XGB implementation provides several advanced features for model fitting, algorithm improvement, and computing environments [23, 24]. Therefore, this algorithm was developed to create a new tool that allows doctors to make decisions based on real clinical data. ML techniques can be valuable for the study of COVID-19, especially because they can uncover complex patterns in large data sets [25–27]. The results obtained by the proposed method confirmed that this system classifies COVID-19 patients with greater precision than other ML methods.

Material

Study design and population

This retrospective observational study includes patients older than 15 years, thus excluding the paediatric population who were admitted to the Virgen de la Luz Hospital in Cuenca. All patients included in this study were those with a confirmed diagnosis for COVID-19 through polymerase chain reaction (RT-PCR) by nasopharyngeal exudate and/or Rapid test of IgM/IgG antibodies admitted in a period comprised between March 15 and April 30 [28, 29].

- Symptom and vital signs that were taken into account: Cough, expectoration, arthralgias o myalgia, asthenia and anorexia, dyspnoea, fever >37.5, saturation (<93%), pulse and respiratory rate.
- Analytical parameters and their levels upon admission to the emergency room: Platelets (k/mcl) <150, Leukocytes (k/mcl) low or normal levels, Lymphocytes (k/mcl) <1,

D-dimer (ng/ml) >1000, Ferritin (ng/ml) >1000, Creatine-kinase (IU/l) >150, Lactate dehydrogenase (IU/l) >450 and Reactive C protein (mg/l) >100.

– Chest X-ray with the presence of unilateral or bilateral interstitial infiltrates. For this, the radiographic evaluation scale ERVI was used.

In addition to the epidemiological context in which we found ourselves, was taken into account. Secondly, the demographic characteristics as well as the presence of comorbidities of the sample were collected in Table 1, as well as the clinical, analytical and radiological presentation [1, 30–32]. An observational study was carried out where it was grouped into two groups based on the objective variable that was mortality. A group was obtained "Case patients" where were those patients who finally died during admission and the control group who survived.

Biomarkers studied

Upon admission to the Emergency Department, patients with symptoms for COVID-19 underwent laboratory tests and radiography, which determined levels of platelets, leukocytes, lymphocytes, prothrombin and thromboplastin times, fibrinogen, DD, c-reactive protein (CRP), procalcitonin (PCT), lactate dehydrogenase (LDH) and pressure of oxygen in arterial blood (PAO2) was performed prior to any treatment [20, 33]. 24 h after admission, the analysis was completed, extracted in the Emergency Department, with profiles of liver, kidney, and ferritin function, etc. The previously mentioned parameters were collected in Table 1. In addition to these parameters, other parameters such as albumin, total proteins, creatinine, ions (sodium, potassium, total or ionic calcium) were included [34, 35]. During hospital admission, analytical controls were carried out depending on the clinical evolution of the patient and the physician in charge of the case. Our database included both the initial values obtained in the emergency room or 24 h after hospitalization, and the worst value presented during admission.

Methods

In our study, the supervised learning algorithm eXtreme Gradient Boosting (XGBoost) was used due to its good response compared to other methods and because it is highly efficient, flexible and portable. XGBoost provides parallel tree reinforcement with fast and accurate resolution mode.

Five widely known machine learning (ML) methods will be used to train patient classification models into two groups. The methods included DT [36–38], GNB [37, 39], KNN [36, 37, 40, 41], SVM [36, 37, 42, 43] and the proposed XGB method [23, 24] and were implemented using MatLab statistical software and machine learning (Matlab 2020a).

Initially we split our data into 70% for training and the rest for testing. A total of 5 k-fold was cross-validated. In addition, machine learning techniques often have one or more hyperparameters that allow a different algorithm adjustment during the training process. The different values of these hyperparameters (number of splits, learners, neighbours, distance metric, distant weight, kernel, box constraint level, multiclass method, etc.) for each method lead to algorithms with different prediction performances to obtain the best possible performance. To optimize these hyperparameters for each ML technique used in this study, each model was trained with a Bayesian optimization approach. Bayesian optimization aims to estimate which hyperparameter settings would maximize algorithm performance from previous attempts, based on the assumption that there is a relationship between the various hyperparameters, and the performance achieved by the algorithm. The hyperparameters adjusted for XGB were maximum depth, gamma, learning rate, and n estimators.

Patient data was not shared between the training and testing subsets to prevent the algorithm from being tested with the same patient data used for training. Figure 1 shows the process followed to carry out the complete study. As can be seen, the subjects to be studied were selected first. Once the database was created, the training and validation of the implemented Machine Learning methods was carried out.

The area under the AUC and the balanced precision were used as performance measures to be maximized. 100 repetitions were used to calculate the mean and standard deviation values of the different calculated results, given the stochastic nature of the boot and machine learning in all simulations. The experiments were repeated uniformly at random to reduce the effects of noise in the data, calculate accurate AUC values, and obtain statistically significant results.

Results

This section describes the results obtained with the patient records used for training and validation in COVID-19 pneumonia classification. The performance of the proposed system has been compared with different classification ML methods accepted in the scientific community.

The results obtained with both test and training records in the classification of mortality in COVID-19 patients are discussed below.

It should be noted that the performance of the proposed system has been compared with different ML classification methods widely used in the literature.

Table 1 Demographic characteristics, comorbidities, clinical, and laboratory findings on admission

	All adults (*n* 151)	Case patients (*n* 53)	Controls (*n* 98)	*P* value
Age, median, year	69 (28–97)	71 (49–92)	67 (28–97)	0.536
Sex				<0.0001
-Male	93 (62%)	43 (46%)	50 (54%)	
-Female	58 (38%)	10 (17%)	48 (83%)	
Comorbidities				
-High Blood Pressure	82 (54%)	32 (60%)	50 (51%)	0.193
-Dyslipemia	58 (38%)	24 (45%)	34 (35%)	0.146
-Diabetes	42 (28%)	25 (48%)	17 (18%)	<0.0001
-Coagulopathy disease	35 (23%)	17 (32%)	18 (18%)	0.046
-Hyperuricemia	27 (18%)	11(21%)	16 (16%)	0.321
-COPD*	15 (10%)	7 (13%)	8 (8%)	0.238
-OSA*	15 (10%)	7 (13%)	8 (8%)	0.238
-Dementia	12 (8%)	6 (11%)	6 (6%)	0.206
-Asthma	4 (2%)	2 (4%)	2 (2%)	0.439
-Autoimmune disease	7 (4%)	1 (2%)	6 (6%)	0.226
-Active Cancer	3 (2%)	2 (4%)	1 (1%)	0.286
BMI				0.011
-Normal	17 (11%)	4 (8%)	13 (13%)	
-Overweight I	11 (7%)	4 (8%)	7 (8%)	
-Overweight II	33 (22%)	6 (11%)	27 (27%)	
-Obesity I	19 (12%)	3 (6%)	16 (16%)	
-Obesity II	15 (10%)	7 (13%)	8 (8%)	
-Obesity III	3 (2%)	1 (2%)	2 (2%)	
Clinical debut				
-Dyspnoea	114 (75%)	45 (85%)	69 (70%)	0.035
-Cough	111 (73%)	41 (77%)	70 (71%)	0.278
-Fever	95 (63%)	35 (66%)	60 (61%)	0.827
-Diarrhoea	31 (21%)	9 (17%)	22 (22%)	0.283
-Nausea and vomiting	16 (10%)	3 (6%)	13 (13%)	0.126
-Neurological symptoms (anosmia/ageusia)	13 (9%)	4 (7%)	9 (9%)	0.495
Initial laboratory tests, median				
-Hemoglobin (g/dl)	13.56	13.67	13.5	0.361
-Platelets (k/mcl)	192	189	194	0.233
-Lymphocytes (k/mcl)	0.58	0.39	0.68	0.360
-Fibrinogen (mg/dl)	408	402	410	0.750
-D-dimer (ng/ml)	1983	4743	676	<0.0001
-Ferritin (ng/ml)	1319	1748	1112	0.610
-Creatinine (mg/dl)	1.27	1,42	1.18	0.346
-Albumin (g/dl)	4.44	5.22	3.43	0.102
-Creatine-kinase (IU/l)	237	331	188	0.028
-Lactate dehydrogenase (IU/l)	725	778	697	0.959
-Reactive C protein (mg/l)				
	137	171	118	0.194
COVID-19 Treatment				
-Antibiotics	139 (92%)	50 (94%)	89 (91%)	0.692
-Azithromycin	128 (85%)	42 (79%)	86 (88%)	0.095
-Hydroxychloroquine	118 (78%)	37 (70%)	81 (83%)	0.042
-Antiviral				
Lopinavir-ritonavir	43 (28%)	16 (30%)	27 (27%)	0.212
Darunavir-cobicistat	30 (20%)	14 (26%)	16 (16%)	0.266

Table 1 (continued)

	All adults (n 151)	Case patients (n 53)	Controls (n 98)	P value
Remdesivir	0 (0%)	0 (0%)	0 (0%)	-
-Steroids	93 (61%)	37 (70%)	56 (57%)	0.087
Pulse	80 (53%)	33 (62%)	47 (48%)	0.065
-Immunosuppressant				
Tocilizumab	13 (9%)	8 (15%)	5 (5%)	0.105
Baricitinib	2 (1%)	1 (2%)	1 (1%)	0.302
Cyclosporine	6 (4%)	3 (5%)	3 (3%)	0.423
-N-acetylcysteine	66 (44%)	27 (51%)	39 (40%)	0.126
-Anticoagulant				
Prophylactic dose	66 (44%)	12 (23%)	54 (55%)	<0.0001
Intermediate dose	59 (39%)	29 (55%)	30 (31%)	0.003
Therapeutic dose	16 (10%)	9 (17%)	7 (7%)	0.061
Coagulation disorders during admission	13 (7%)	8 (15%)	5 (5%)	
-Bleeding	6 (54%)	4 (57%)	2(40%)	0.114
Major bleeding	2 (33%)	2 (50%)	0 (0%)	
-Thrombosis	7 (63%)	4 (57%)	3(60%)	0.358
Pulmonary embolism (PE)	4 (57%)	3 (75%)	1 (33%)	
Deep venous thrombosis (DVT)	2 (28%)	0 (0%)	2 (67%)	
Both (PE + DVT)	1 (14%)	1 (25%)	0 (0%)	

Therapy administered and complications during hospitalization *COPD* Chronic Obstructive Lung Disease, *OSA* Obstructive sleep apnoea
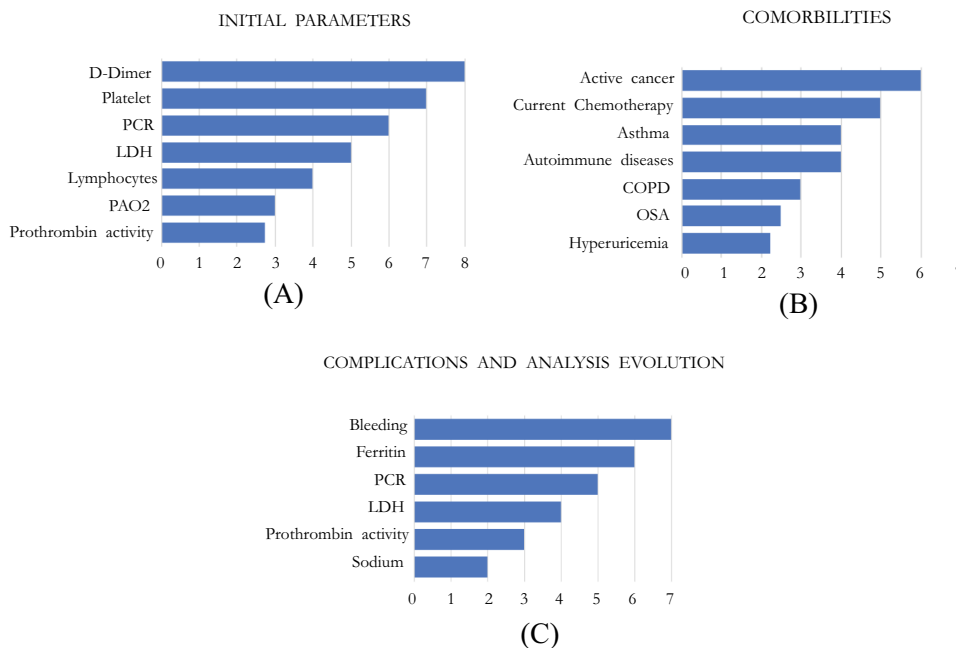

Fig. 1 Training and validation scheme for machine learning methods
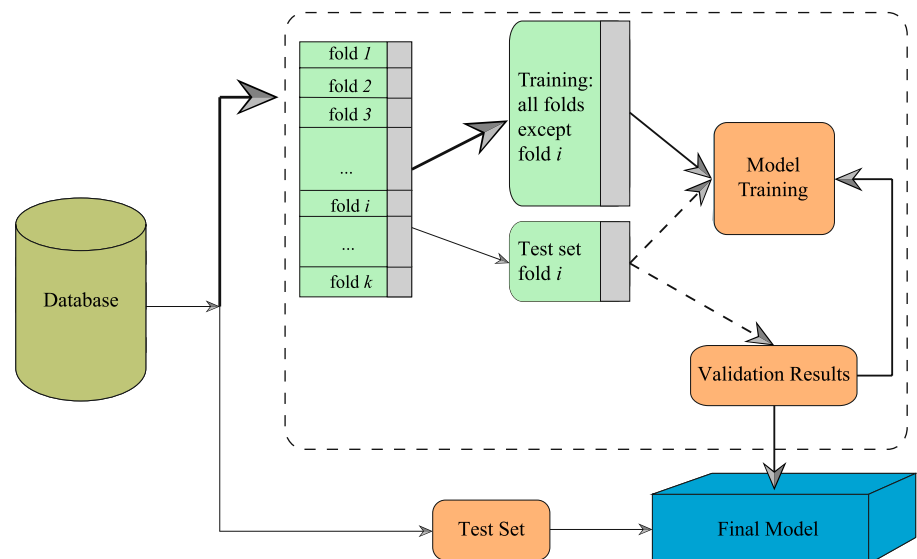
Comparison between Case patient's vs Control group

A total of 151 confirmed or highly suspected COVID-19 patients were collected during the previously referred study period. The median age of the sample was 69 (28–97) years, 53 died during hospital admission classifying them within the group Case patient's vs 98 who survived (35% vs 65%), being predominantly men (62% vs 38%)

Both groups Case patient's vs control were compared to evaluate the differences that influenced the clinical evolution. Among the characteristics of the sample collected in Table 1, it should be noted that comorbidities were homogeneously distributed between both groups, except for Diabetes Mellitus (48% vs 18%, p = <0.0001) and previous coagulopathies (32% vs 18%, $p = 0.046$) that were found more frequently in the Case patients' group. Regarding the symptoms referred to their admission, dyspnoea was described more

INITIAL PARAMETERS

COMORBILITIES

COMPLICATIONS AND ANALYSIS EVOLUTION

(A)

(B)

(C)

frequently in the Case patient's (85% vs 70%, p = 0.035), however no significant differences were observed in terms of fever or extrapulmonary symptoms such as gastrointestinal or neurological.

Initial blood tests showed significantly elevated levels of DD (4743 vs 676, $p \leq 0.0001$) and creatine-kinase (CK, 331 vs 188, $p = 0.028$). However, no notable differences were found in terms of platelet, lymphocyte, ferritin or LDH levels. On the other hand, the treatments received during the hospital stay did not show clear differences between the two groups. It should be noted the differences observed in the dosage of anticoagulant treatment, with the use of prophylactic doses being observed more frequently (23% vs 55%, $p \leq 0.0001$) in the control group, compared to a higher frequency of intermediate doses (55% vs 31%, $p \leq 0.003$) used in the Case patients- group.

ML analysis

The most important analytical predictors of mortality were the initial DD, platelets, lymphocytes levels, LDH, CRP, PAO2 and prothrombin activity collected upon admission to hospital. Furthermore, there were other powerful predictors of mortality among the comorbidities of the patients, such as asthma, CODP and OSA, history of hyperuricemia. The presence of these comorbidities conditioned the clinical course and outcome of the disease. Other strong predictors were active cancer with chemotherapy treatment, autoimmune diseases, and lung diseases which have been collected in the Fig. 2. In addition, critical events that occurred during admission as well as analytical values during hospital stay

Table 2 The table displays the mean values and standard deviation of balanced accuracy, recall, precision, AUC, F_1 score, MCC, DYI and Kappa score of the ML system and the proposed model evaluated in this work

Method	Balanced accuracy (%)	Recall	Precision	AUC
GNB	77,32	77,41	76,77	0,77
SVM	78,89	80,01	79,92	0,80
DT	82,66	82,76	82,08	0,82
KNN	85,84	85,94	85,23	0,85
XGB	93,22	93,33	92,55	0,93
Method	F1 score	MCC	DYI	Kappa
GNB	77,09	68,61	77,32	68,84
SVM	79,78	71,00	80,02	71,24
DT	82,42	73,35	82,67	73,60
KNN	85,59	76,18	85,85	76,43
XGB	92,94	82,72	93,22	82,99

were included in the analysis. It was observed that critical events such as bleeding as well as the alterations maintained in the levels of ferritin, CRP, LDH, prothrombin activity were powerful predictors of mortality in patients with COVID-19 pneumonia. Another important variable in the prediction of mortality in our sample was the sodium level during hospital period. The predictor variables were evaluated by different Machine learning classification methods for the prediction of mortality, among which we find XGB, KNN, DT, SVM and GNB collected in Table 2.

On the one hand, the receiver operating characteristic (ROC) has been applied to compare the classification
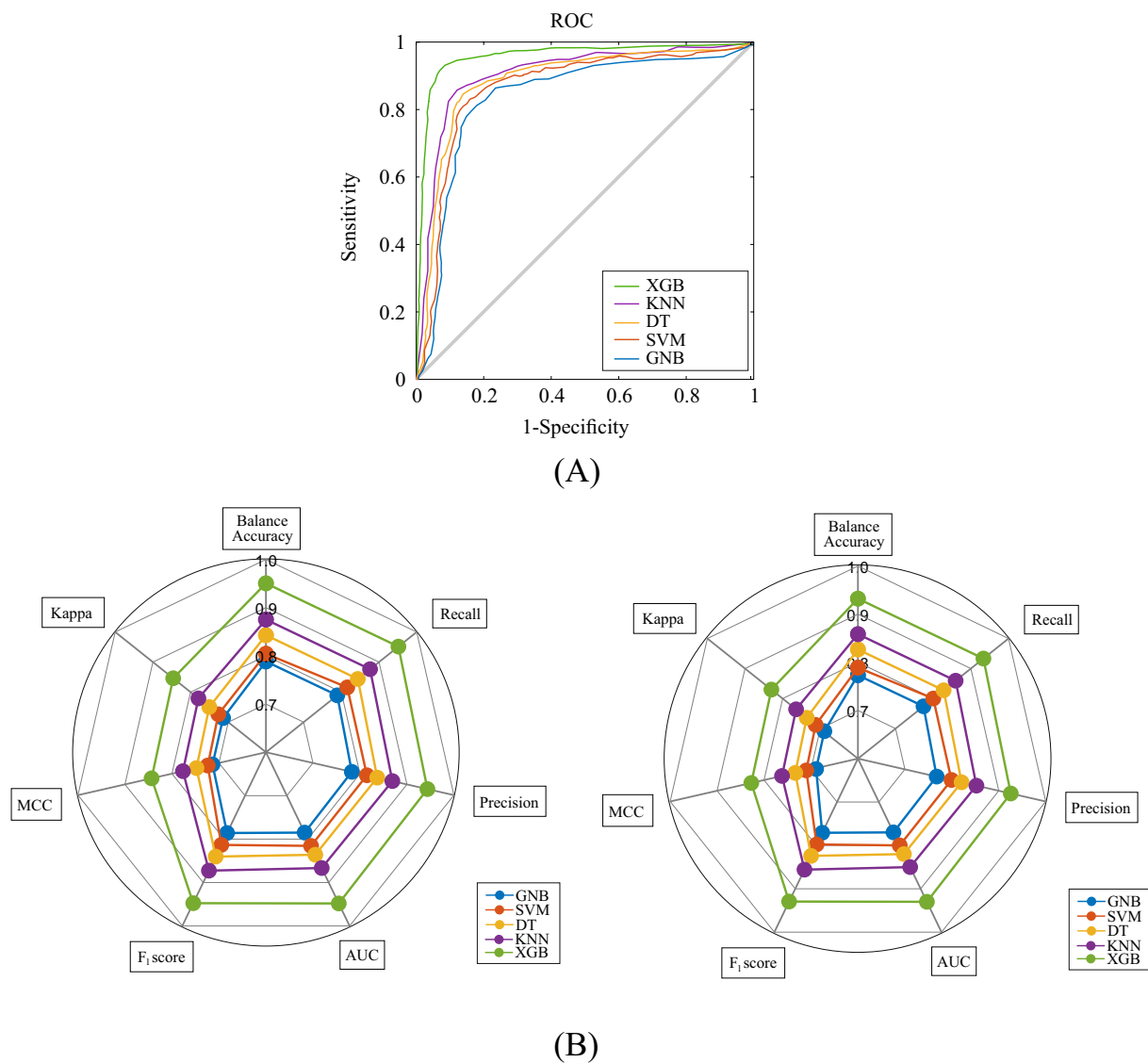
Fig. 3 Section **A**: ROC curves for the five assessed machine learning predictors. Section **B**: The figure shows the radar plot of the training phase (left) and test (right) for the prediction of mortality in COVID-19 patients

capability of the proposed system with other ML methods as can be seen in Section A of Fig. 3. By means of sensitivity and specificity, we can represent the results obtained among deceased patients versus survivors. We observed that there were clear differences in power between the different methods for predicting mortality. Also, Table 2 shows diverse parameters of different classification methods such as SVM, DT, GNB, KNN and the proposed system for classification of COVID-19 patients as well as healthy patients. The SVM and GNB based systems obtain lower classification values than other methods with accuracy values close to 79%. In turn, DT and KNN improve this value reaching 86%. About the proposed XGB method, the values obtained were much higher than those of the other methods, reaching an accuracy of close to 94%. It is true that the KNN and DT methods are

close in their Precision and Recall values to the proposed method, but they do not achieve the same percentage. SVM and GNB achieve values close to 82% in the F1 score value but none of these achieve the results obtained by the proposed method. For the evaluation of these methods, values of balanced accuracy, recovery, precision and F1 score are used, which allows us to select the best method that allows obtaining a precision close to 95%. XGB proved to have a strong performance for prediction compared to other methods as SVM or GNB for which the AUC was 0.8 vs 0.77 respectively. The 7.37% improvement obtained with XGB in the DYI classifies it as the best of the studied methods.

On the other hand, to check the performance of the proposed XGB system in the classification of the two classes analysed, other parameters widely used also in the literature

such as AUC, MCC, DYI and Kappa index were used. For this check, one of the most reliable statistical indices available, the Matthews Correlation Coefficient (MCC), was used. This coefficient produces a high score only if the prediction performed well in all four categories of the matrix. Results in all four categories of the confusion matrix (true positives, false negatives, true negatives and false positives), proportionally to the size of the positive items and the size of the negative items in the dataset. Again, the proposed method achieves a higher MCC value (closer to 1) compared to the other methods analysed. As can been observed in Table 2, DT and KNN are the best performers compared to the others but do not achieve the XGB value. Another parameter used was the Kappa index where again XGB outperforms KNN and DT which are the best performing methods. XGB method obtains a better score in most of the metrics, which indicates that it is a balanced model. All models score high on the training data set, however they go down with the test data, showed on Section B of Fig. 3. In the case of XGB, it behaves similarly in both data sets, confirming it as the best model among those compared.

Discussion

This study aimed to describe the usefulness of new machine learning algorithms in the search for data patterns and to identify the relationships between all fields. Machine Learning allows us to predict the mortality of the population of patients with Pneumonia due to COVID-19 in a way that allows clinicians to anticipate those patients at risk of developing an acute distress syndrome associated with the disease, which could translate into a reduction so much of mortality. An initial evaluation of comorbidities, clinical and laboratory presentation was carried out using biochemical and haematological parameters at hospital admission that could later be used as predictors. The results found show a clear association between initial alterations, mainly in analytical parameters such as levels of DD, platelets, CRP, LDH, lymphocytes, PAO2 and prothrombin activity with mortality, translating into a state of hyperinflammation initial in the Case patients- group, being independent of the treatments received during their hospitalization. Cell activation after tissue damage generates an inflammatory response with expression of different tissue factors that will produce alterations in coagulation times and in the activation of prothrombin, generating hypercoagulatory states and increased risk of thrombosis. Doctors use the DD level to detect thrombosis, however it can be elevated in other situations such as bleeding, cancer or even pregnancy among others. Its elevation in COVID-19 patients has been widely studied, and there are multiple causes that can generate it, including inflammation, endothelial dysfunction, hypoxia,

also age and comorbidities. This elevation has been related to an increase in mortality, and therefore its initial detection as well as its follow-up is necessary. During the hospital stay, a worsening of the levels of Ferritin, CRP, LDH and prothrombin activity was observed, as well as the appearance of haemorrhagic events were associated with higher mortality [33, 44]. In addition, the decrease in lymphocyte count has been associated with an increased risk of serious disease. However, it was not a predictor of mortality during follow-up in our sample. Other authors have observed that a lower initial lymphocyte count as well as a sustained decrease is an early predictor of severe disease [45].

On the other hand, the burden of comorbidities presented by the patients also had a greater influence on mortality. Patients with a high burden of comorbidities are subjected to chronic stress that generates a decrease in immunity and therefore a greater susceptibility to infections. Also, the presence of cardiovascular risk factors such as hypertension or diabetes conditions prior vascular damage, which will entail a greater risk of acute cardiovascular events during the infectious process. In our study, the comorbidities with the greatest association were mainly presented by patients with active cancer and undergoing chemotherapy, despite having a small sample concordance with the results obtained in other studies carried out. Patients with active solid and hematological cancers, although mainly the latter, have a higher risk of developing a more severe disease from COVID-19 and a higher mortality from it [46]. Although the differences between treatment with conventional chemotherapy vs biological therapies with monoclonal antibodies were not taken into account in our study. On the other hand, previous pulmonary pathology was other predictor variables for mortality. Although the frequency of these comorbidities in the sample studied was low, we cannot rule out that the frequency of these diseases is conditioned by greater social isolation as it is considered high risk during the first wave. Most of them were patients <65 years of age, with other associated comorbidities, with mortality in the COPD group of 46%. Therefore, although the incidence was low, these patients have a higher risk of developing severe disease and higher mortality [47]. This is mainly due to lower resistance to the virus. Several studies have observed that lack of expression of angiotensin-converting enzyme 2 (ACE2) confers protection against SARS-CoV-2 infection. ACE2 is used by the virus to enter the cell and spread, and it is overexpressed in smokers and COPD patients and may be an aggravating factor to develop a more serious disease. In addition, these patients have pre-existing lung damage that, associated with a chronic inflammatory state, will condition a barrier and deficient immunity against the virus.

Regarding the history of hyperuricemia, it was associated with higher mortality. Elevated uric acid has already been described in various lung diseases and in COVID-19

Table 3 The table shows the results of the comparison of the proposed method and the various published methods for the prediction of mortality in COVID-19 patients

Author	Classifier	ACC	AUC	F_1 value (%)
Li et al. [50]	LASSO	–	>80	–
Hu et al. [51]	AUROCs	–	88	–
Mele et al. [52]	PM2.5	–	–	77.8
Di Castelnuovo et al. [53]	RF	83.4	–	90.4
Proposed	XGB	93.22	93	92.94

patients. However, the role of uric acid is controversial since it can act as an antioxidant as well as an immunity stimulant. On the one hand, the release of uric acid deposits secondary to tissue damage generates their elevation in the bloodstream, thus exerting an immune stimulus and triggering a generalized inflammation as a consequence of the increased production of cytokines such as Il-1, Il-6 and TNF-alpha. However, the lowered levels of uric acid could be related to a loss of its antioxidant and immunomodulatory power in situations of significant oxidative stress. Various authors have investigated in this regard, Bo Chen et al. [48] presented a cohort of cases where an association was observed between U-shaped uric acid levels, that is, high uric acid levels were associated with higher mortality and need for IMV, while low uric acid levels increased the risk of death, admission to ICU and IMV. Similar results regarding hypouricemia were observed by In's Dufour et al. [49] on a cohort of ICU patients who presented lower uric acid levels upon admission to the unit. Further studies are needed to clarify the behaviour of uric acid and its pathophysiology during acute SARS-CoV-2 infection.

Table 3 presents a summary of different techniques that have been used to predict mortality in COVID-19 patients. As can be seen, the proposed model achieves more accurate values. The results show that the proposal can effectively improve the performance of other classification methods. Due to this, the proposed system may be a useful tool to classify patients with SARS-CoV-2 Pneumonia according to risk.

Machine Learning allows obtaining greater diagnostic precision with a smaller sample size but a large number of data established, thus improving the results obtained by traditional statistical methods. These results allow the clinician Obtain a more sensitive tool for the identification of higher risk factors that allows more energetic initial action. Artificial intelligence is in continuous progress, in recent months several studies have been published on its validity in the evaluation of CT images in patients with COVID-19 pneumonia, reaching a sensitivity and specificity > 90%, and greatly exceeding those results obtained by trained medical personnel.

Conclusion

In conclusion, using artificial intelligence techniques, we have been able to obtain highly reliable predictive variables, being able to detect those patients with the highest risk and establish early measures that allow modifying the course of the disease. Within the numerous factors influence the mortality of patients who develop COVID-19 pneumonia, we highlight D-dimer, platelets, CRP, LDH, lymphocytes, PAO2 and prothrombin activity, as well as the changes kept the hospitalization. The development of bleeding because of this pro-inflammatory state has also been associated with high mortality. Therefore, knowledge of the clinical characteristics of patients affected by SARS-CoV-2 and their evolution is of crucial importance for the development of Clinical Practice Guidelines that facilitate decision-making for physicians.

References

1. Arishnan A, Hamilton James P, Alqahtani Saleh A, Woreta Tinsay A (2021) A narrative review of coronavirus disease 2019 (COVID-19): clinical, epidemiological characteristics, and systemic manifestations. Internal and emergency medicine, pp 1–16
2. García-Salido A (2020) Revisión narrativa sobre la respuesta inmunitaria frente a coronavirus: descripción general, aplicabilidad para sars-cov-2 e implicaciones terapéuticas. In Anales de Pediatria, vol. 93, pp 60–e1. Elsevier
3. d'Alessandro M, Bergantini L, Cameli P, Curatola G, Remediani L, Bennett D, Bianchi F, Perillo F, Volterrani L, Mazzei M et al (2021) Serial kl-6 measurements in COVID-19 patients. Internal and Emergency Medicine, pp 1–5
4. Rello J, Belliato M, Dimopoulos M-A, Giamarellos-Bourboulis EJ, Jaksic V, Martin-Loeches I, Mporas I, Pelosi P, Poulakou G, Pournaras S, et al (2020) Update in COVID-19 in the intensive care unit from the 2020 hellenic athens international symposium. Anaesthesia Critical Care & Pain Medicine
5. Saghazadeh A, Rezaei N (2020) Towards treatment planning of COVID-19: rationale and hypothesis for the use of multiple immunosuppressive agents: anti-antibodies, immunoglobulins, and corticosteroids. Int Immunopharmacol 84:106560

6. Lan S-H, Lai C-C, Huang H-T, Chang S-P, Li-Chin L, Hsueh P-R (2020) Tocilizumab for severe COVID-19: a systematic review and meta-analysis. Int J Antimicrob Agents 56(3):106103

7. Jesús R-B, Jerónimo P, Jordi C, Pablo R, Inmaculada J, María Y, José RA, Juan B, EstherAznar M, Pedro GD et al (2021) Treatment with tocilizumab or corticosteroids for COVID-19 patients with hyperinflammatory state: a multicentre cohort study (SAM-COVID-19). Clin Microbiol Infect 27(2):244–252

8. Zhang X, Zhang Y, Qiao W, Zhang J, Qi Z (2020) Baricitinib, a drug with potential effect to prevent sars-cov-2 from entering target cells and control cytokine storm induced by COVID-19. Int Immunopharmacol 106749

9. Sarkar S, Khanna P, Soni KD (2021) Are the steroids a blanket solution for COVID-19? A systematic review and meta-analysis. J Med Virol 93(3):1538–1547

10. Villar J, Ferrando C, Martínez D, Ambrós A, Muñoz T, Soler JA, Aguilar G, Alba F, González-Higueras E, Conesa LA et al (2020) Dexamethasone treatment for the acute respiratory distress syndrome: a multicentre, randomised controlled trial. Lancet Respir Med 8(3):267–276

11. Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, Hohmann E, Chu HY, Luetkemeyer A, Kline S et al (2020) Remdesivir for the treatment of COVID-19. N Engl J Med 383(19):1813–1826

12. Gutierrez-Lorenzo M, Cuadros-Martínez CM (2020) Baricitinib in the treatment of sars-cov-2 infection. Revista Espanola de Quimioterapia: Publicacion Oficial de la Sociedad Espanola de Quimioterapia 33(4):294–295

13. Rodriguez-Garcia JL, Sanchez-Nievas G, Arevalo-Serrano J, Garcia-Gomez C, Jimenez-Vizuete JM, Martinez-Alfaro E (2021) Baricitinib improves respiratory function in patients treated with corticosteroids for sars-cov-2 pneumonia: an observational cohort study. Rheumatology 60(1):399–407

14. Sterne JAC, Murthy S, Diaz JV, Slutsky AS, Villar J, Angus DC, Annane D, Azevedo LCP, Berwanger O, Cavalcanti AB et al (2020) Association between administration of systemic corticosteroids and mortality among critically ill patients with COVID-19: a meta-analysis. JAMA 324(13):1330–1341

15. Lemos ACB, do Espírito Santo DA, Salvetti MC, Gilio RN, Agra LB, Pazin-Filho A, Miranda CH (2020) Therapeutic versus prophylactic anticoagulation for severe COVID-19: a randomized phase ii clinical trial (hesacovid). Thrombosis Res 196:359–366

16. Hadid T, Kafri Z, Al-Katib A (2021) Coagulation and anticoagulation in COVID-19. Blood Rev 47:100761

17. Gorbalenya AE, Baker SC, Baric R, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM, Neuman BW, et al (2020) Severe acute respiratory syndrome-related coronavirus: the species and its viruses–a statement of the coronavirus study group

18. Sarzi-Puttini P, Giorgi V, Sirotti S, Marotto D, Ardizzone S, Rizzardini G, Antinori S, Galli M (2020) COVID-19, cytokines and immunosuppression: what can we learn from severe acute respiratory syndrome? Clin Exp Rheumatol 38(2):337–342

19. Luo H, You C, Si-wei L, Yue-qiang F (2021) Characteristics of coagulation alteration in patients with COVID-19. Ann Hematol 100(1):45–52

20. Zeng F, Huang Y, Guo Y, Yin M, Chen X, Xiao L, Deng G (2020) Association of inflammatory markers with the severity of COVID-19: a meta-analysis. Int J Infect Dis 96:467–474

21. Shi C, Wang C, Wang H, Yang C, Cai FEI, Zeng F, Cheng F, Liu Y, Deng TZB et al (2020) The potential of low molecular weight heparin to mitigate cytokine storm in severe COVID-19 patients: a retrospective cohort study. Clin Transl Sci 13(6):1087–1095

22. Ferrer R, Navas A, Adda M, Artigas A (2008) Papel de la coagulación en la fisiopatología de la lesión pulmonar aguda: Paralelismo con la sepsis. Med Intensiva 32(6):304–311

23. Chang W, Liu Y, Wu X, Xiao Y, Zhou S, Cao W (2019) A new hybrid xgbsvm model: application for hypertensive heart disease. IEEE Access 7:175248–175258

24. Chen W, Fu K, Zuo J, Zheng X, Huang T, Ren W (2017) Radar emitter classification for large data set based on weighted-xgboost. IET Radar Sonar Navig 11(8):1203–1207

25. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS One 14(11):e0224365

26. Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X, Liang H (2017) Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Sci Rep 7(1):1–9

27. Mateo J, Rius-Peris JM, Maraña-Pérez AI, Valiente-Armero A, Torres AM (2021)Extreme gradient boosting machine learning method for predicting medical treatment in patients with acute bronchiolitis. Biocybernet Biomed Eng

28. Yamayoshi S, Sakai-Tagawa Y, Koga M, Akasaka O, Nakachi I, Koh H, Maeda K, Adachi E, Saito M, Nagai H et al (2020) Comparison of rapid antigen tests for COVID-19. Viruses 12(12):1420

29. Mohamadian M, Chiti H, Shoghli A, Biglari S, Parsamanesh N, Esmaeilzadeh A (2021) COVID-19: virology, biology and novel laboratory diagnosis. J Gene Med 23(2):e3303

30. Wong HYF, Hiu YSL, Ambrose H-TF, SiuTing L, Thomas W-YC, Christine SYL, Macy M-SL, Jonan CYL, Keith W-HC, om Wai-Hin C et al (2020) Frequency and distribution of chest radiographic findings in patients positive for COVID-19. Radiology 296(2):E72–E78

31. Sánchez-Oro R, Nuez JT, Martínez-Sanz G (2020) La radiología en el diagnóstico de la neumonía por sars-cov-2 (covid-19). Med Clin 155(1):36

32. Alsharif W, Qurashi A (2021) Effectiveness of COVID-19 diagnosis and management tools: a review. Radiography 27(2):682–687

33. Zheng T, Liu X, Wei Y, Li X, Zheng B, Gong Q, Dong L, Zhong J (2021) Laboratory predictors of COVID-19 mortality: a retrospective analysis from Tongji hospital in Wuhan. Mediators Inflamm 2021

34. De Carvalho H, Richard MC, Chouihed T, Goffinet N, Bastard QL, Freund Y, Kratz A, Dubroux M, Masson D, Figueres L et al (2021) Electrolyte imbalance in COVID-19 patients admitted to the emergency department: a case–control study. Internal Emerg Med 1–6

35. Gheorghe G, Ilie M, Bungau S, Stoian Anca MP, Bacalbasa N, Diaconu CC (2021) Is there a relationship between COVID-19 and hyponatremia? Medicina 57(1):55

36. Han J, Pei J, Kamber M (2016) Data mining: concepts and techniques. Third edition

37. Azevedo A (2019) Data mining and knowledge discovery in databases. In: Advanced methodologies and technologies in network architecture, mobile computing, and data analytics

38. Rivera-Lopez R, Canul-Reich J (2018) Construction of near-optimal axis-parallel decision trees using a differential-evolution-based approach. IEEE Access 6:5548–5563

39. Biplab KD, Himadri SD (2020) Gfnb: Gini index-based fuzzy Naive bayes and blast cell segmentation for leukemia detection using multi-cell blood smear images. Med Biol Eng Comput 58(11):2789–2803

40. Zhang S, Li X, Zong M, Zhu X, Wang R (2017) Efficient knn classification with different numbers of nearest neighbors. IEEE Trans Neural Netw Learn Syst 29(5):1774–1785

41. Xing W, Bei Y (2019) Medical health big data classification based on knn classification algorithm. IEEE Access 8:28808–28819

42. Shuang Yu, Li X, Zhang X, Wang H (2019) The ocs-svm: an objective-cost-sensitive svm with sample-based misclassification cost invariance. IEEE Access 7:118931–118942

43. Kafai M, Eshghi K (2017) Croification: accurate kernel classification with the efficiency of sparse linear svm. IEEE Trans Pattern Anal Mach Intell 41(1):34–48

44. Al-Samkari H, Karp Leaf RS, Dzik WH, Carlson JCT, Fogerty AE, Waheed A, Goodarzi K, Bendapudi PK, Bornikova L, Gupta S et al (2020) COVID-19 and coagulation: bleeding and thrombotic manifestations of sars-cov-2 infection. Blood 136(4):489–500

45. Wang F, Nie J, Wang H, Zhao Q, Xiong Y, Deng L, Song S, Ma Z, Mo P, Zhang Y (2020) Characteristics of peripheral lymphocyte subset alteration in COVID-19 pneumonia. J Infect Dis 221(11):1762–1769

46. García-Suárez J, De La Cruz J, Cedillo Á, Llamas P, Duarte R, Jiménez-Yuste V, Hernández-Rivas JÁ, Gil-Manso R, Kwon M, Sánchez-Godoy P et al (2020) Impact of hematologic malignancy and type of cancer therapy on COVID-19 severity and mortality: lessons from a large population-based registry study. J Hematol Oncol 13(1):1–12

47. Leung JM, Niikura M, Yang CWT, Sin DD (2020) COVID-19 and copd. Eur Respir J, 56(2)

48. Chen B, Lu C, Hong-Qiu G, Li Y, Zhang G, Lio J, Luo X, Zhang L, Hu Y, Lan X et al (2021) Serum uric acid concentrations and risk of adverse outcomes in patients with COVID-19. Front Endocrinol, 12

49. Dufour I, Werion A, Belkhir L, Wisniewska A, Perrot M, De Greef J, Schmit G, Yombi JC, Wittebole X, Laterre P-F et al (2021) Serum uric acid, disease severity and outcomes in COVID-19. Crit Care 25(1):1–12

50. Li M, Zhang Z, Cao W, Liu Y, Du B, Liu Q, Uddin MN, Jiang S, Chen C, Chen C et al (2021) Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. Sci Total Environ 764:142810

51. Chuanyu H, Liu Z, Jiang Y, Shi O, Zhang X, Kelin X, Suo C, Wang Q, Song Y, Kangkang Yu et al (2020) Early prediction of mortality risk among patients with severe COVID-19, using machine learning. Int J Epidemiol 49(6):1918–1929

52. Mele M, Magazzino C (2021) Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence. Environ Sci Pollut Res 28(3):2669–2677

53. Di Castelnuovo A, Bonaccio M, Costanzo S, Gialluisi A, Antinori A, Berselli N, Blandi L, Bruno R, Cauda R, Guaraldi G et al (2020) Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian corist study. Nutr Metab Cardiovasc Dis 30(11):1899–1913