Machine Learning Based Predictors for COVID-19 Disease Severity

Dhruv Patel
University of Southern California
Vikram Kher
University of Southern California
Bhushan Desai
University of Southern California
Xiaomeng Lei
University of Southern California
Steven Cen
University of Southern California
Neha Nanda
University of Southern California
Ali Gholamrezanezhad
University of Southern California
Vinay Duddalwar
University of Southern California
Bino Varghese
University of Southern California
Assad A Oberai (aoberai@usc.edu)
University of Southern California

Research Article

Keywords: COVID-19 Severity, Machine Learning, Predictive Modeling

Posted Date: November 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-108301/v1

Machine Learning Based Predictors for COVID-19 Disease Severity

Dhruv Patel[1], Vikram Kher[1], Bhushan Desai[2], Xiaomeng Lei[2], Steven Cen[2], Neha Nanda[2], Ali Gholamrezanezhad[2], Vinay Duddalwar[2], Bino Varghese[2], and Assad A Oberai[1,*]

[1]Viterbi School of Engineering, University of Southern California , Los Angeles, CA, USA
[2]Keck School of Medicine, University of Southern California, CA, USA
[*]aoberai@usc.edu

ABSTRACT

Predictors of the need for intensive care and mechanical ventilation can help healthcare systems in planning for surge capacity for COVID-19. We used socio-demographic data, clinical data, and blood panel profile data at the time of initial presentation to develop machine learning algorithms for predicting the need for intensive care and mechanical ventilation. Among the algorithms considered, the Random Forest classifier performed the best with AUC = 0.80 for predicting ICU need and AUC = 0.82 for predicting the need for mechanical ventilation. We also determined the most influential features in making this prediction, and concluded that all three categories of data are important. Finally, we determined the relative importance of blood panel profile data and noted that the AUC dropped by 0.12 units when this data was not included, thus indicating that it provided valuable data in predicting disease severity.

Introduction

The current coronavirus disease 2019 (COVID-19) pandemic has strained healthcare delivery models across the world. In the US there are over 8 million cases and 5.4% have required hospitalization. Of the hospitalized patients, to date, 5.2% have required care in the intensive care unit (ICU)[1]. Based on current projections, by January 1st 2021 the number of ICU beds needed for COVID patients will exceed the available ICU beds by 10.6%[2,3]. With this challenge in supply of ICU beds, states and counties have created detailed surge plans to ensure timely care of critically ill patients suffering with COVID-19. In order to sustain healthcare delivery through this pandemic, it is imperative to adopt a proactive approach towards utilization of healthcare resources like ICU beds and ventilators. Given the urgency for resource allocation and optimization, we sought to identify patient-level clinical characteristics at the time of admission to predict the need for ICU care and mechanical ventilation in COVID-19 patients.

Methods

Data for this study was extracted from an Institutional Review Board (IRB) approved COVID-19 REDCap[4] repository. Informed consent for the repository was waived by the USC IRB consistent with §45 CFR 46.116(f). The study was conducted in accordance with USC policies, IRB policies, and federal regulations. Subjects' privacy and confidentiality were protected according to applicable HIPAA, and USC IRB policies and procedures. The repository contained demographic, clinical and laboratory data for all COVID-19 positive patients seen at the Keck Medical Center of USC, Verdugo Hills Hospital, and Los Angeles County + USC Medical Center. Repository data elements include data from three categories: (a) socio-demographic data including age, sex, travel, contact history, and co-morbidities; (b) presenting clinical data gleaned

from symptoms and the results of an initial physical examination including fever, dyspnea, respiratory rate, and blood oxygen saturation (SpO2); (c) blood panel profile including RT-PCR, InterLeukin-6, D-Dimer, complete blood count, lipase, and C-reactive protein (CRP). They also include the outcome data, namely, the need for ICU admission and mechanical ventilation. A description of all the input features, their type, and their median, minimum and maximum values is presented in Tables 1, 2 and 3.

The study cohort comprised of 212 patients (123 males, 89 females) with an average age of 53 years (13-92 years), of which 74 required intensive care at some point during their stay, and 47 required mechanical ventilation. We note that only data obtained at the time of initial presentation, with 24 hours of initial presentation, was included as input to the predictive models, and the need for ICU admission and mechanical ventilation at any time during hospitalization were selected as outcomes.

Features with more than 30% missing data were excluded from the analysis. In the retained features, missing data was imputed using an iterative method. In this method the feature to be imputed is treated as a function of a subset of other highly-correlated features and missing values are obtained using regression[5]. This subset of features is then iterated over to arrive at the final estimate.

The retained features were used to compute the correlation of the outcome with input features. Thereafter, data was split into training (60%), validation (20%), and testing sets (20%). The training and validation data were used to train and tune the hyperparameters of supervised learning algorithms (random forests, multilayer perceptron, support vector machines and logistic regression). Among all these algorithms the Random Forest[6] (RF) classifier was found to be the most accurate and was considered for further analysis.

The tuned RF model was applied to testing data to compute the probability of ICU admission and mechanical ventilation. This was repeated with five different folds, yielding predicted probabilities for 212 subjects generated by five distinct RF models. These were used to generate an ROC curve and compute the area under the curve (AUC). The relative importance of the input features was evaluated by computing their Gini importance.

The analysis describe above was first performed with input data from all categories, that is, socio-demographic data, presenting clinical data, and blood panel profile data. Thereafter, the blood panel profile data was excluded and the analysis was performed once again. This second analysis was done to assess the relative importance of the blood panel data in predicting the outcomes.

Results

In Figure 1, we have plotted the AUC values for predicting the need for ICU and mechanical ventilation for all the algorithms considered in this study. From this figure we observe that the algorithms based on decision trees, that is, random forests,Random Forest, Extra Tree Classifier, and Gradient Boosting tend to perform better. This is likely because the simpler algorithms like logistic regression and support vector machines do not have sufficient capacity to capture the complexity in the prediction, while other algorithms like MultiLayerPerceptrons (MLP) do not have sufficient data for training. This leads to issues with robustness and over-fitting. Further, among the algorithms based on decision trees, the Random Forest (RF) classifier is the most accurate and was considered for further analysis.

For the RF predictor, we reported an AUC of 0.80, 95% CI (0.73-0.86) in predicting the need for ICU and an AUC of 0.83, 95% CI (0.76-0.90) for predicting the need for mechanical ventilation. These values demonstrate that we are able to accurately predict the need for intensive care and ventilation from data acquired at the time of admission. The performance of the RF predictor is similar to results reported in studies from China[7] and the Netherlands[8] (AUC of 0.88 and 0.77, respectively). We note that these studies differ from ours due to the regional differences in the population and the viral strain. Further, these studies

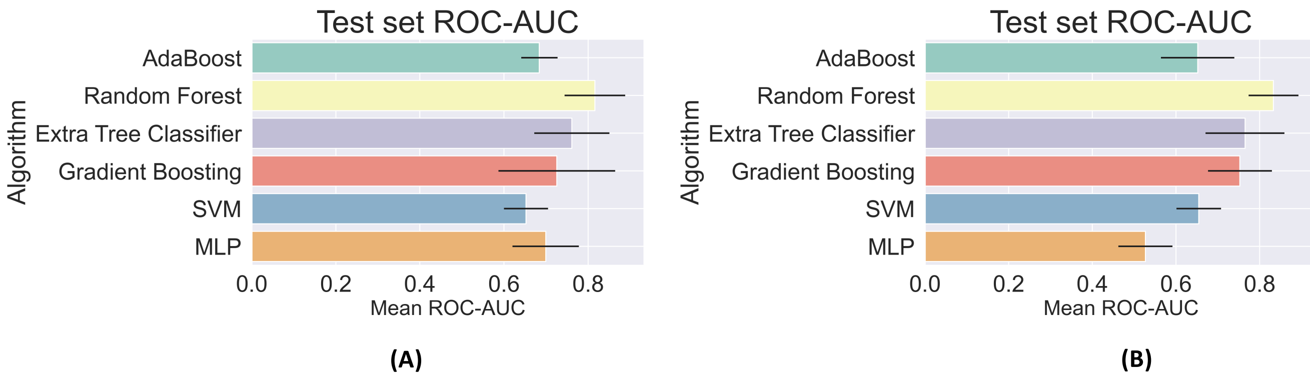

Figure 1. Area under the curve (AUC) for the classifiers considered in the study for predicting the need for ICU (A) and mechanical ventilation (B).

also included chest x-ray imaging features and tested a single type of ML algorithm (logistic regression). Deep learning models were also developed based on a cohort from China[9], and these report an AUC   0.89 for a coarse measure of disease severity that clubs together patients receiving ICU care or mechanical ventilation, and those ultimately succumbing to the disease.

When only socio-demographic and presenting clinical data was used as input, the AUC value for predicting ICU need dropped to 0.68, 95% CI (0.60-0.75), and that for predicting ventilation dropped to 0.70, 95% CI (0.61-0.79). This indicates that the lab marker data provides significant additional information and is important in improving the accuracy of these predictions. A recent comprehensive survey of laboratory markers concluded that many of the markers that are included in this study are correlated with COVID-19 severity and should therefore be used in models for predicting disease severity[10]. However, our results also indicate that it is possible to make moderately accurate predictions with only socio-demographic and presenting clinical data. This is particularly useful when quick decisions are required and the time or resources necessary for acquiring lab marker data are not available in a timely manner.

The top ten features with the strongest correlation to ICU admission are shown in Figure 2A, and the most important features for the RF classifier for ICU need are shown in Figure 2B. Similarly, the top ten features with the strongest correlation to the need for mechanical ventilation are shown in Figure 3A, and the most important features for the RF classifier for mechanical ventilation need are shown in Figure 3B.

Taken together, this set represents features that strongly influence the likelihood of ICU admission and mechanical ventilation. We note that they belong to all three categories – socio-demographic data, presenting clinical data, and blood panel profile data – showing that all these type of data are necessary in making an accurate assessment of disease severity. Several of these features have been implicated in determining the severity of COVID-19 by other researchers[11–17]; however, there are few studies that have considered them together and determined their relative importance.

In Figure 4, we plot the distribution of some of the most important input features, including lab markers, presenting symptoms and socio-demographic data for two sets of patients: those who require ICU care and whose who do not. We observe that the distribution of Creatinine (indicator of kidney function), C-reactive Protein (measure of inflammatory response), D-Dimer (measure of blood clot formation and breakdown) and Procalcitonin (elevated during infection and sepsis) among patients who require ICU care is spread over a larger range and has a higher average value. A similar trend is observed in the distribution for the respiratory rate. For SpO2 levels also we observe a distribution spread over a wider range for patients admitted to the ICU; however, in this case this group has a lower average value. We also note that

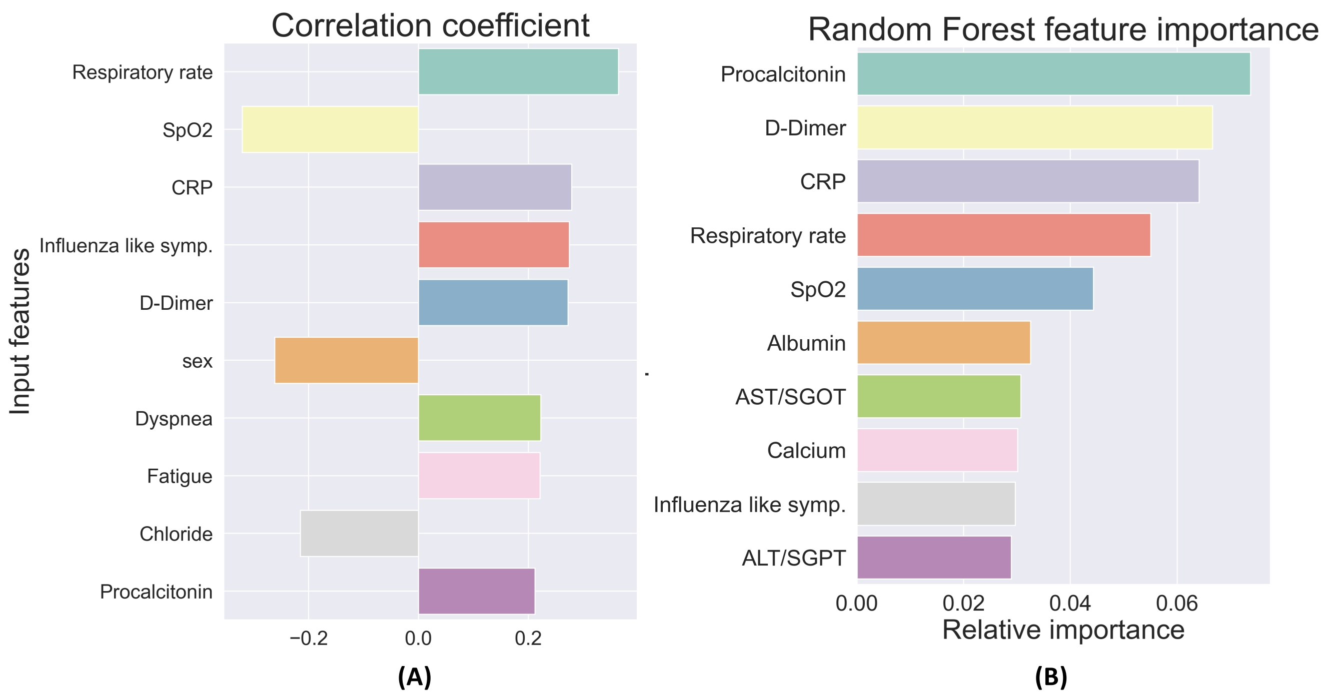

Figure 2. (A) Ten most highly correlated features with the need for ICU care. (B) Ten features with the highest relative importance for predicting the need for ICU care.

the presence of the influenza-like symptoms roughly doubles the likelihood of requiring ICU care (from around 25% to 52%). Further, the percentage of males who are admitted to the ICU is much higher than the percentage of females (46% to 20%).

Discussion

The results presented in this study demonstrate that data acquired at or around the time of admission of a COVID-19 patient to a care facility can be used to make an accurate assessment of their need for critical care and mechanical ventilation. Further, the important features in this data belong to three different sets, namely, socio-demographic data, presenting clinical data, and blood panel profile data. We also report that in cases where the blood panel data is not available, useful prediction might still be made, albeit with some loss of accuracy. This would be relevant to situations where the time or resources to acquire this type of data are limited. The list of important features identified in our study is also indicative of a disease that affects multiple systems in the body including the respiratory, the circulatory system and the immune system.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. *The COVID Tracking Project* (2020 (accessed September 25, 2020)).

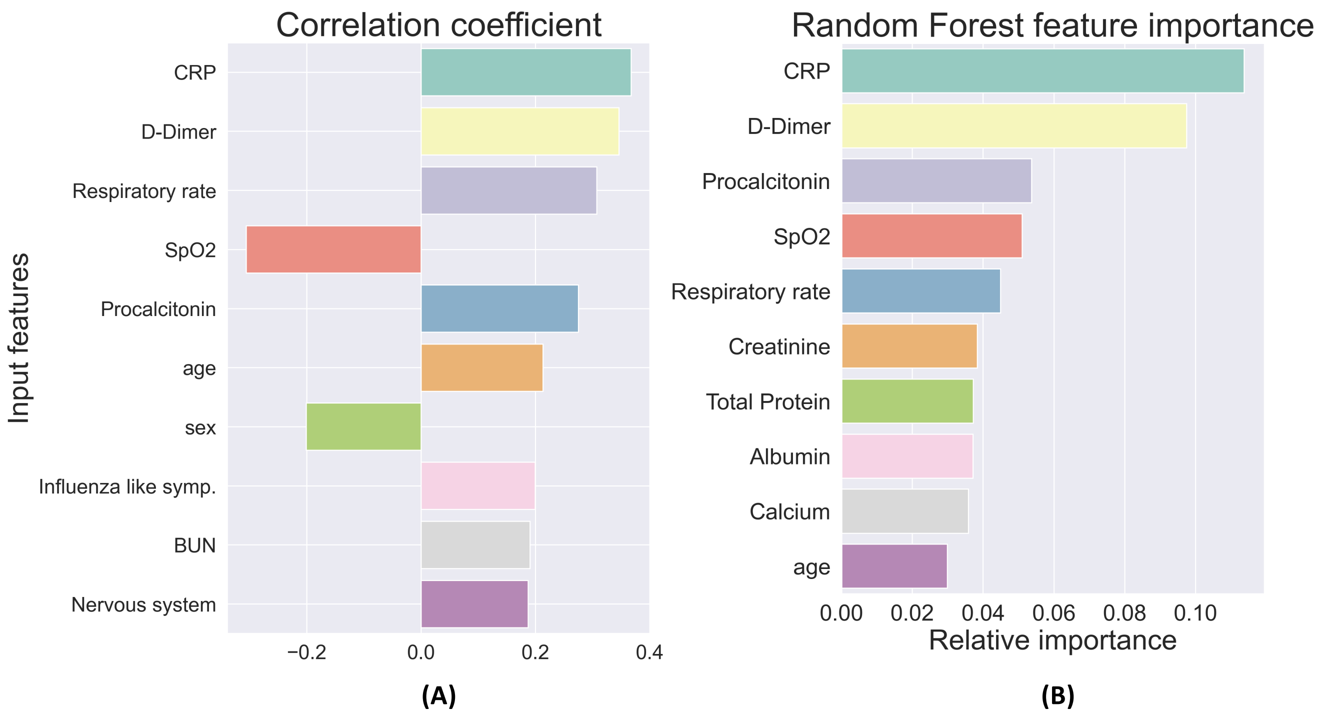

Figure 3. (A) Ten most highly correlated features with the need for mechanical ventilation. (B) Ten features with the highest relative importance for predicting the need for mechanical ventilation.

2. *IHME COVID-19 Projections* (2020 (accessed September 25, 2020)).

3. *AHA COVID-19 Bed Occupancy Projection Tool* (2020 (accessed September 25, 2020)).

4. Harris, P. A. *et al.* Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J. biomedical informatics* **42**, 377–381 (2009).

5. *Scikit Iterative Imputer* (2020 (accessed September 25, 2020)).

6. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

7. Liang, W. *et al.* Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19. *JAMA Intern. Medicine* (2020).

8. Schalekamp, S. *et al.* Model-based prediction of critical illness in hospitalized patients with covid-19. *Radiology* 202723 (2020).

9. Liang, W. *et al.* Early triage of critically ill covid-19 patients using deep learning. *Nat. communications* **11**, 1–7 (2020).

10. Skevaki, C., Fragkou, P. C., Cheng, C., Xie, M. & Renz, H. Laboratory characteristics of patients infected with the novel sars-cov-2 virus. *J. Infect.* (2020).

11. Liu, F. *et al.* Prognostic value of interleukin-6, c-reactive protein, and procalcitonin in patients with covid-19. *J. Clin. Virol.* 104370 (2020).

12. Lippi, G. & Plebani, M. Procalcitonin in patients with severe coronavirus disease 2019 (covid-19): a meta-analysis. *Clin. chimica acta; international journal clinical chemistry* **505**, 190 (2020).

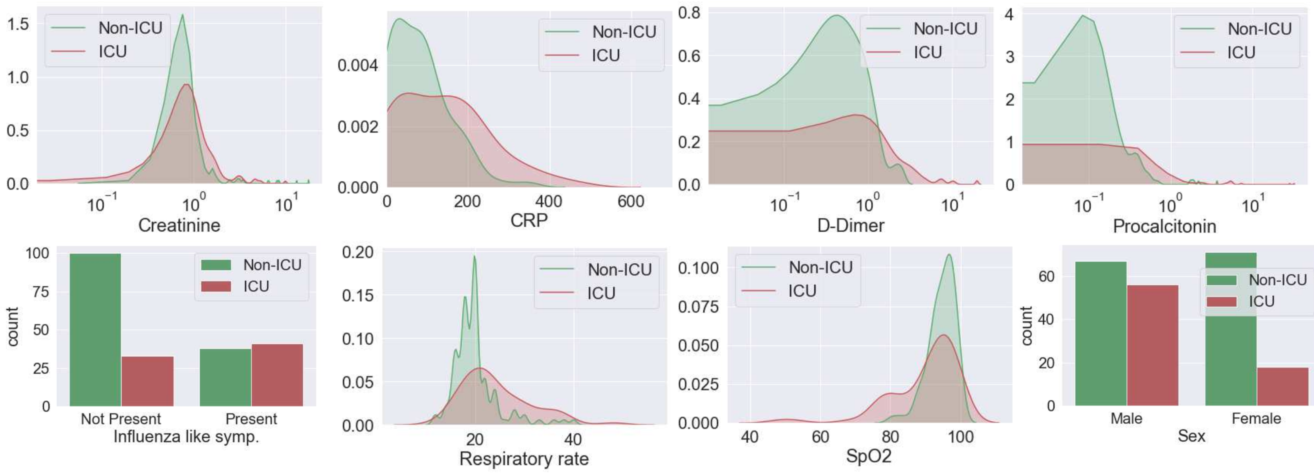

Figure 4. Distribution of (from top left to bottom right) Creatinine, C-reactive Protein (CRP), D-Dimer, Procalcitonin, influenza-like symptoms, respiratory rate, SpO2 level, and sex for patients admitted to ICU and those who are not.

13. Xie, J. *et al.* Association between hypoxemia and mortality in patients with covid-19. In *Mayo Clinic Proceedings* (Elsevier, 2020).

14. He, F. *et al.* Clinical features and risk factors for icu admission in covid-19 patients with cardiovascular diseases. *Aging disease* **11**, 763 (2020).

15. Zhang, J. *et al.* Risk factors for disease severity, unimprovement, and mortality of covid-19 patients in wuhan, china. *Clin. Microbiol. Infect.* (2020).

16. Liu, X. *et al.* Risk factors associated with disease severity and length of hospital stay in covid-19 patients. *J. Infect.* **81**, e95–e97 (2020).

17. Li, K. *et al.* The clinical and chest ct features associated with severe and critical covid-19 pneumonia. *Investig. radiology* (2020).

Author contributions statement

D.P. and V.K. performed the ML analysis. S.C. performed the statistical analysis. B.D, X.L., A.G. and B.V. organized and curated patient data. N.N. and V.D. provided the epidemiological and clinical insight and context to the study. A.A.O. conceived and guided the ML aspects of the study. All authors reviewed the manuscript.

Additional information

Competing interests The author(s) declare no competing interests.

Socio-Demographic Features	Type	Median	Min	Max
Age	Numerical	53	12	93
Sex	Categorical	0	0	1
Pregnant	Categorical	0	0	2
Race	Categorical	7	2	7
Ethnicity	Categorical	1	1	3
BMI	Numerical	29	0	84
Travel	Categorical	0	0	1
Primary Contact	Categorical	1	0	2
Secondary Contact	Categorical	1	0	2
Other Contact	Categorical	1	0	2
Work Contact	Categorical	0	0	1

Table 1. Socio-demographic features used as input.

Clinical Features	Type	Median	Min	Max
Immuno-compromised	Categorical	0	0	1
Cardiac history	Categorical	0	0	1
Diabetes Mellitus	Categorical	0	0	1
COPD	Categorical	0	0	1
Asthma	Categorical	0	0	1
Interstitial Lung Disease	Categorical	0	0	1
Obesity	Categorical	0	0	1
Auto-immune disease	Categorical	0	0	1
Hypertension	Categorical	0	0	1
Other Morbidity	Categorical	0	0	1
Fever	Categorical	1	0	1
Chills	Categorical	0	0	1
Shortness of breath or dyspnea	Categorical	1	0	1
Chest pain	Categorical	0	0	1
Cough	Categorical	1	0	1
Loss of smell	Categorical	0	0	1
Loss of taste	Categorical	0	0	1
Body ache / Myalgia	Categorical	0	0	1
Fatigue	Categorical	0	0	1
Throat Pain	Categorical	0	0	1
Abdominal pain	Categorical	0	0	1
Diarrhea	Categorical	0	0	1
Influenza like illness symptoms	Categorical	0	0	1
Other Symptom	Categorical	1	0	1
Days since symptoms presented	Numerical	5	1	29
General Appearance	Categorical	1	1	3
Head	Categorical	1	1	3
Eyes	Categorical	1	1	3
Ears	Categorical	1	1	3
Nose	Categorical	1	1	3
Throat	Categorical	1	1	3
Chest and lungs	Categorical	2	1	3
Heart	Categorical	1	1	3
Abdomen	Categorical	1	1	3
Extremities	Categorical	1	1	3
Nervous system	Categorical	1	1	3
Skin	Categorical	1	1	3
Systolic blood pressure	Numerical	129	54	228
Diastolic blood pressure	Numerical	75	34	116
Heart Rate	Numerical	106	53	156
Respiratory rate	Numerical	20	12	48
Body Temperature	Numerical	37	35	39
SpO2	Numerical	95	48	100

Table 2. Input features from presenting clinical data and the results of an initial physical examination.

Blood Panel Variables	Type	Median	Min	Max
Glucose	Numerical	131	53	575
Calcium	Numerical	8	6	11
Albumin	Numerical	3	0	4
Total Protein	Numerical	7	0	9
Sodium	Numerical	136	124	154
Potassium	Numerical	4	2	6
Bicarbonate (Total CO2)	Numerical	23	11	37
Chloride	Numerical	98	84	114
Blood urea nitrogen (BUN)	Numerical	13	0	137
Creatinine	Numerical	0	0	17
Alkaline Phosphatase (ALP)	Numerical	80	29	417
Alanine Amino Transferase (ALT/SGPT)	Numerical	35	5	247
Aspartate Amino Transferase (AST/SGOT)	Numerical	47	13	355
Bilirubin	Numerical	0	0	20
C-Reactive Protein (CRP)	Numerical	91	0	470
D-Dimer	Numerical	0	0	20
Procalcitonin	Numerical	0	0	31

Table 3. Input features from blood panel profile.

Figures

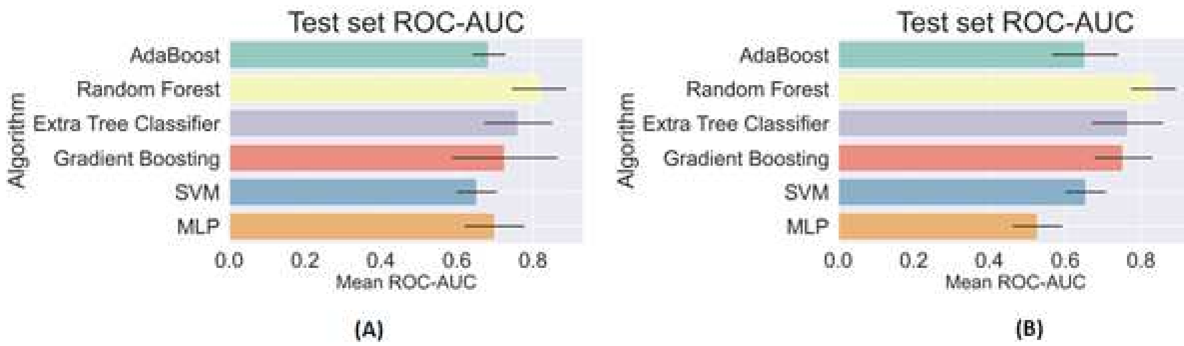

Figure 1

Area under the curve (AUC) for the classifiers considered in the study for predicting the need for ICU (A) and mechanical ventilation (B).

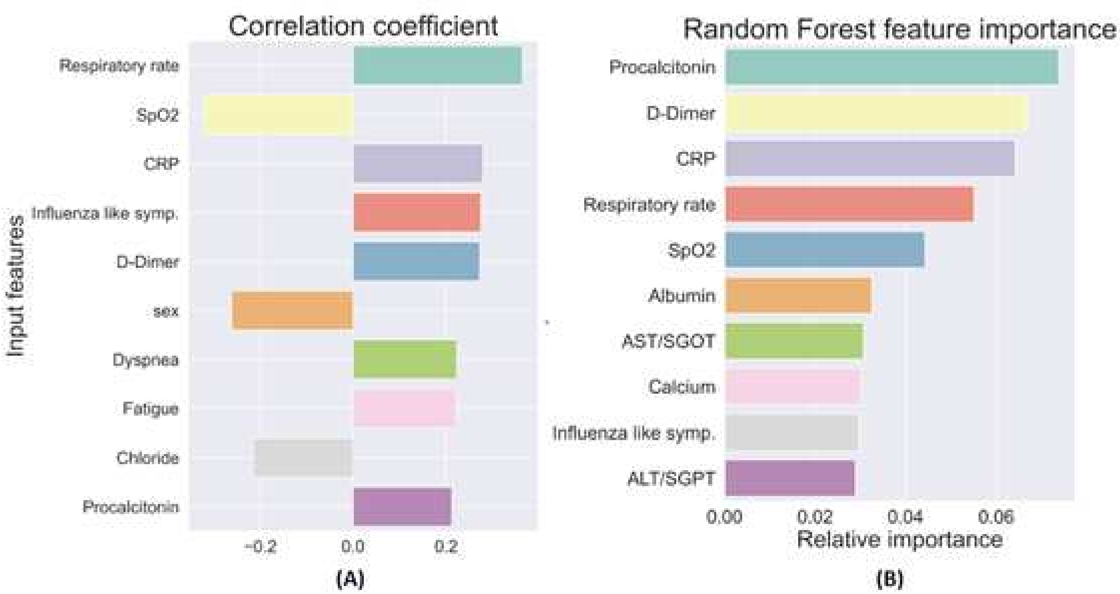

Figure 2

(A) Ten most highly correlated features with the need for ICU care. (B) Ten features with the highest relative importance for predicting the need for ICU care.

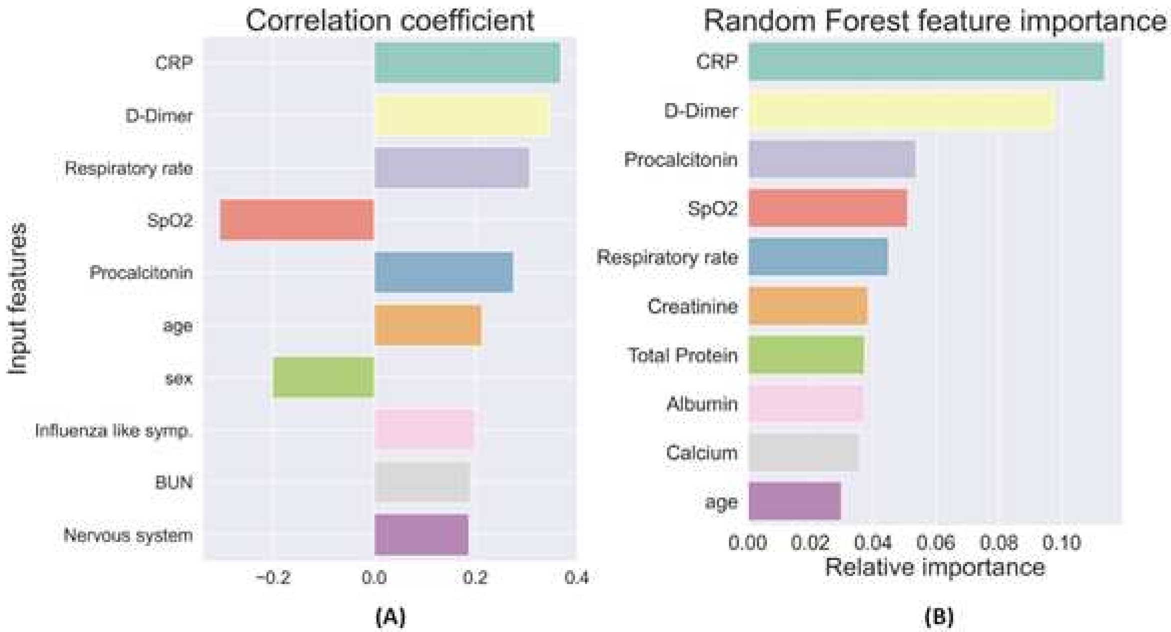

Figure 3

(A) Ten most highly correlated features with the need for mechanical ventilation. (B) Ten features with the highest relative importance for predicting the need for mechanical ventilation.

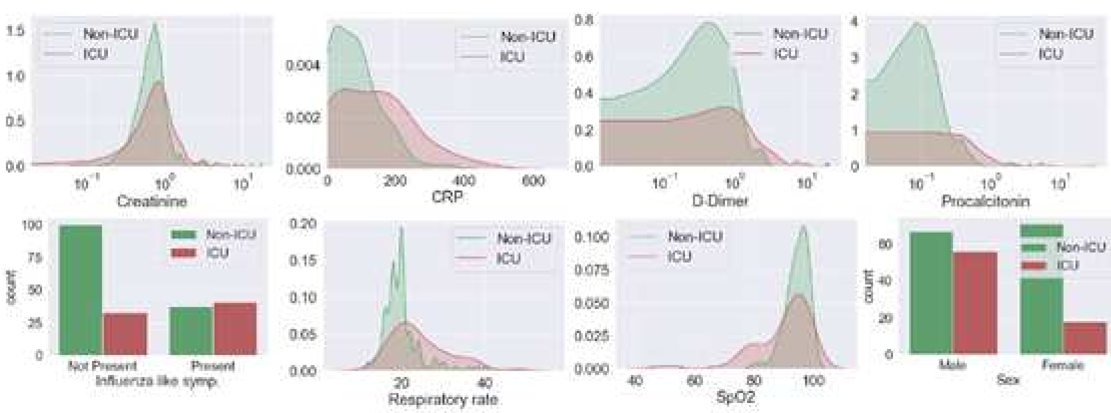

Figure 4

Distribution of (from top left to bottom right) Creatinine, C-reactive Protein (CRP), D-Dimer, Procalcitonin, influenza-like symptoms, respiratory rate, SpO2 level, and sex for patients admitted to ICU and those who are not.