frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 13 August 2021
doi: 10.3389/frai.2021.579931

# Predicting the Disease Outcome in COVID-19 Positive Patients Through Machine Learning: A Retrospective Cohort Study With Brazilian Data

*Fernanda Sumika Hojo De Souza [1]*, Natália Satchiko Hojo-Souza [2], Edimilson Batista Dos Santos [1], Cristiano Maciel Da Silva [3] and Daniel Ludovico Guidoni [1]*

*[1]Department of Computer Science, Federal University of São João Del-Rei, São João Del-Rei, Brazil, [2]Laboratory of Immunopathology, Oswaldo Cruz Foundation–Minas, Belo Horizonte, Brazil, [3]Department of Technology, Federal University of São João Del-Rei, Ouro Branco, Brazil*

The first officially registered case of COVID-19 in Brazil was on February 26, 2020. Since then, the situation has worsened with more than 672, 000 confirmed cases and at least 36, 000 reported deaths by June 2020. Accurate diagnosis of patients with COVID-19 is extremely important to offer adequate treatment, and avoid overloading the healthcare system. Characteristics of patients such as age, comorbidities and varied clinical symptoms can help in classifying the level of infection severity, predict the disease outcome and the need for hospitalization. Here, we present a study to predict a poor prognosis in positive COVID-19 patients and possible outcomes using machine learning. The study dataset comprises information of 8, 443 patients concerning closed cases due to cure or death. Our experimental results show the disease outcome can be predicted with a Receiver Operating Characteristic AUC of 0.92, Sensitivity of 0.88 and Specificity of 0.82 for the best prediction model. This is a preliminary retrospective study which can be improved with the inclusion of further data. Conclusion: Machine learning techniques fed with demographic and clinical data along with comorbidities of the patients can assist in the prognostic prediction and physician decision-making, allowing a faster response and contributing to the non-overload of healthcare systems.

**Keywords: COVID-19, prediction model, disease outcome, machine learning, Brazil**

OPEN ACCESS

***Edited by:***
*Alejandro F. Frangi,*
*University of Leeds, United Kingdom*

***Reviewed by:***
*Yaojiang Huang,*
*Minzu University of China, China*
*Aarti Sathyanarayana,*
*Harvard University, United States*

****Correspondence:***
*Fernanda Sumika Hojo De Souza*
*fsumika@ufsj.edu.br*

***Specialty section:***
*This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence*

***Received:*** *18 August 2020*
***Accepted:*** *02 August 2021*
***Published:*** *13 August 2021*

***Citation:***
*De Souza FSH, Hojo-Souza NS, Dos Santos EB, Da Silva CM and Guidoni DL (2021) Predicting the Disease Outcome in COVID-19 Positive Patients Through Machine Learning: A Retrospective Cohort Study With Brazilian Data. Front. Artif. Intell. 4:579931. doi: 10.3389/frai.2021.579931*

## 1 INTRODUCTION

A new coronavirus with a high efficiency in infecting humans emerged in the Wuhan city (Hubei Province, China) in December 2019. The disease, named COVID-19, is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) being highly contagious (Jin et al., 2020; Sohrabi et al., 2020). The virus spread quickly outside China and the World Health Organization (WHO) recognized the outbreak as a pandemic in March 2020 (World health organization, 2020a). By June 2020, nearly seven million cases have been confirmed and more than 400, 000 deaths have been reported from SARS-CoV-2 infection, reaching 216 countries (World health organization, 2020b).

Despite high transmissibility, the disease spectrum is diverse, ranging from asymptomatic cases to extremely severe conditions. SARS-CoV-2 infection is characterized by fever, generalized weakness, dry cough, headache, dyspnoea, myalgia, as well as leukopenia, lymphocytopenia, neutrophilia, elevated levels of C-reactive protein, D-dimer, and inflammatory cytokines (Fu et al., 2020; Guan

et al., 2020; Huang et al., 2020) and loss of smell and taste in the early stage of infection (Menni et al., 2020). However, the status can quickly evolve to acute respiratory distress syndrome (ARDS), cytokine storm, coagulation dysfunction, acute cardiac injury, acute kidney injury, and multi-organ dysfunction if the disease is not resolved, resulting in patient death (Li et al., 2020; Zhou et al., 2020).

Elderly people and those with comorbidities such as diabetes and cardiovascular disease are more likely to progress to severe conditions (Fu et al., 2020). Obesity has also been linked to an increased likelihood of severe COVID-19 (Gao et al., 2020). In this context, a three-stage classification system was proposed according to the degree of severity of the disease (Siddiqi and Mehra, 2020). The initial stage is characterized by nonspecific clinical symptoms such as malaise, cough and fever. The diagnosis can be made in a respiratory sample to detect the presence of SARS-CoV-2 by RT-PCR and blood tests can reveal lymphopenia and neutrophilia. The second stage is characterized by viral lung disease and localized inflammation usually associated with hypoxia, requiring hospitalization and mechanical ventilation. However, there is a number of cases that progress to a more severe third stage of the disease, characterized by high levels of inflammatory biomarkers and severe systemic inflammation. In this phase, the occurrence of shock and generalized collapse of the organs is large and the prognosis for recovery is poor.

So far, there is still not enough vaccine for all or even specific therapeutic drugs for the treatment of COVID-19 (Sanders et al., 2020). Therefore, quarantine and social distancing have been recommended as a measure to reduce the rate of infection aiming not to exceed the capacity of health systems to provide care.

Currently, COVID-19 is on the rise in Latin American countries (Burki, 2020), whose health systems may not support the care of all seriously infected people. Lack of beds, ventilators in Intensive Care Units (ICUs) and Personal Protective Equipment (PPE) by health care workers restrain the treatment of severe cases. Faced with these challenges, identifying those patients with hospitalization priority is a crucial aspect in order to optimize care and promote a reduction of deaths.

As of early June 2020, more than 672, 000 confirmed cases and at least 36,000 deaths have been registered in Brazil (WHO, 2020). Brazil has become the epicenter of the pandemic, which expands to interior cities, most of which do not have ICUs.

The rapid spread of COVID-19 along with the slow vaccination process and the lack of specific therapeutic measures has accelerated the use of artificial intelligence and machine learning on different fronts such as viral dissemination pattern, rapid and accurate diagnosis, development of new therapeutic approaches and identification of people most susceptible to the disease (Alimadadi et al., 2020).

The aim of the present study is to make a prognosis or early identification of patients at increased risk of developing severe COVID-19 symptoms using an available database from the Espírito Santo Brazilian State. Espírito Santo has an estimated population of 4.06 million inhabitants (Brazilian Institute of Ge, 2020) and on May 30th, 2020 had registered 13, 690 confirmed cases of COVID-19. Using machine learning techniques, a classification problem can be solved aiming to predict the disease outcome for positive COVID-19 patients, based on individual information, in addition to comorbidities and varied clinical symptoms. We show that it is possible to predict the outcome of the individual's condition with a ROC AUC of 0.92, Sensitivity of 0.88 and Specificity of 0.82. This process can be of great importance in helping decision-making within hospitals, since resources are becoming limited every day. Patients classified as having a more severe condition can be prioritized in this case.

## 2 DATA AND METHODS

This is a retrospective cohort study that did not directly involve patients and does not require approval by an ethics committee. The database used is publicly available on the Espírito Santo state portal (Government of the state o, 2020). Two sets of data are used in our study, say the training cohort and the validation cohort. The database was downloaded twice, on May 23rd, 2020 and May 30th, 2020. Information regarding 13, 690 patients who tested positive for COVID-19 comprises the last database, along with the outcome of each case. As the main objective of the present work is to predict the disease outcome of patients infected by the virus, only closed cases (due to death or cure) are used, comprising 4, 826 and 3, 617 patients in the training cohort and validation cohort, respectively. Additional information on cleaning and preparing the data is provided below, followed by the machine learning methods employed. All analyses were performed using Python (version 3.6.8) and the machine learning package scikit-learn (version 0.22.2.post1) along with imblearn (version 0.4.3).

### 2.1 Data Cleaning and Preparation

The dataset includes individual basic information such as gender and age range, symptoms, comorbidities and a recent travelling history. A notification status of each entry in the database is said to be *closed* or *open*, since the data is updated daily as new information becomes available. Thus, only data whose status is *closed* were considered, as they are those that have the outcome of the case: *cure* or *death*. Cases whose outcome is unknown have been disregarded.

The city of origin of the patients and the neighborhood of residence are also available in the database. We considered that this information would not be very relevant to the problem under study, and we decided not to include such data in our datasets due to its high variability in values and possible noise generation in the experiments.

Therefore, based on the data available at the source, we end up our datasets with the following information: confirmation criteria, age range, gender, race/color, education, fever, respiratory distress, cough, runny nose, sore throat, diarrhea, headache, pulmonary comorbidity, cardiac comorbidity, kidney comorbidity, diabetes comorbidity, smoking comorbidity, obesity comorbidity, hospitalization, travel in Brazil and international travel. All of these variables comprise categorical variables:

**TABLE 1 |** Demographic data and clinical characteristics of the study populationTraining Dataset (Feb 29th - May 23rd).

| | All n (%) | Cure n (%) | Death n (%) |
|---|---|---|---|
| Confirmation Criteria | 4,826 (100.00) | 4,430 (91.79) | 396 (8.21) |
| Laboratorial | 4,676 (96.89) | 4,287 (96.77) | 389 (98.23) |
| Clinical Ep | 138 (2.86) | 132 (2.98) | 6 (1.52) |
| Clinical | 4 (0.08) | 4 (0.09) | 0 (0.00) |
| Unknown | 8 (0.17) | 7 (0.16) | 1 (0.25) |
| Basic Information | 4,826 (100.00) | 4,430 (91.79) | 396 (8.21) |
| Age range | | | |
| 0–4 years old | 38 (0.79) | 38 (0.86) | 0 (0.00) |
| 5–9 years old | 22 (0.46) | 22 (0.50) | 0 (0.00) |
| 10–19 years old | 99 (2.05) | 99 (2.23) | 0 (0.00) |
| 20–29 years old | 692 (14.34) | 687 (15.51) | 5 (1.26) |
| 30–39 years old | 1,368 (28.35) | 1,355 (30.59) | 13 (3.28) |
| 40–49 years old | 1,117 (23.15) | 1,088 (24.56) | 29 (7.32) |
| 50–59 years old | 756 (15.67) | 699 (15.78) | 57 (14.39) |
| 60–69 years old | 372 (7.71) | 283 (6.39) | 89 (22.47) |
| 70–79 years old | 206 (4.27) | 109 (2.46) | 97 (24.49) |
| 80–89 years old | 112 (2.32) | 35 (0.79) | 77 (19.44) |
| 90 years old or more | 44 (0.91) | 15 (0.34) | 29 (7.32) |
| Gender | | | |
| Male | 2,232 (46.25) | 2005 (45.26) | 227 (57.32) |
| Female | 2,594 (53.75) | 2,425 (54.74) | 169 (42.68) |
| Race/Color | | | |
| Asian | 214 (4.43) | 185 (4.18) | 29 (7.32) |
| White | 1,543 (31.97) | 1,438 (32.46) | 105 (26.52) |
| Unknown | 1,346 (27.89) | 1,250 (28.22) | 96 (24.24) |
| Indigenous | 4 (0.08) | 3 (0.07) | 1 (0.25) |
| Brown | 1,456 (30.17) | 1,326 (29.93) | 130 (32.83) |
| Black | 263 (5.45) | 228 (5.15) | 35 (8.84) |
| Education[a] | | | |
| Illiterate | 55 (1.14) | 29 (0.65) | 26 (6.57) |
| iES (1–4 grade) | 134 (2.78) | 98 (2.21) | 36 (9.09) |
| cES (4 grade) | 94 (1.95) | 76 (1.72) | 18 (4.55) |
| iES (5–8 grade) | 168 (3.48) | 128 (2.89) | 40 (10.10) |
| cES (8 grade) | 165 (3.42) | 139 (3.14) | 26 (6.57) |
| iHS | 157 (3.25) | 141 (3.18) | 16 (4.04) |
| cHS | 1,321 (27.37) | 1,289 (29.10) | 32 (8.08) |
| iHE | 115 (2.38) | 115 (2.60) | 0 (0.00) |
| cHE | 887 (18.38) | 880 (19.86) | 7 (1.77) |
| Unknown | 1730 (35.85) | 1,535 (34.65) | 195 (49.24) |
| Hospitalization | 4,826 (100.00) | 4,430 (91.79) | 396 (8.21) |
| Yes | 479 (9.93) | 249 (5.62) | 230 (58.08) |
| No | 3,249 (67.32) | 3,128 (70.61) | 121 (30.56) |
| unknown | 1,098 (22.75) | 1,053 (23.77) | 45 (11.36) |
| Travelling History | 4,826 (100.00) | 4,430 (91.79) | 396 (8.21) |
| Brazil | | | |
| Yes | 333 (6.90) | 325 (7.34) | 8 (2.02) |
| No | 3,300 (68.38) | 3,003 (67.79) | 297 (75.00) |
| unknown | 1,193 (24.72) | 1,102 (24.88) | 91 (22.98) |
| International | | | |
| Yes | 33 (0.68) | 32 (0.72) | 1 (0.25) |
| No | 3,153 (65.33) | 2,878 (64.97) | 275 (69.44) |
| unknown | 1,640 (33.98) | 1,520 (34.31) | 120 (30.30) |
| | **All n (%)** | **Cure n (%)** | **Death n (%)** |
| Sympton | 4,826 (100.00) | 4,430 (91.79) | 396 (8.21) |
| Fever | | | |
| Yes | 2,859 (59.24) | 2,612 (58.96) | 247 (62.37) |
| No | 1867 (38.69) | 1726 (38.96) | 141 (35.61) |
| unknown | 100 (2.07) | 92 (2.08) | 8 (2.02) |
| Respiratory Distress | | | |
| Yes | 1,243 (25.76) | 966 (21.81) | 277 (69.95) |
| No | 3,492 (72.36) | 3,376 (76.21) | 116 (29.29) |
| unknown | 91 (1.89) | 88 (1.99) | 3 (0.76) |
| Cough | | | |
| Yes | 3,104 (64.32) | 2,849 (64.31) | 255 (64.39) |
| No | 1,625 (33.67) | 1,492 (33.68) | 133 (33.59) |
| unknown | 97 (2.01) | 89 (2.01) | 8 (2.02) |
| Runny nose | | | |
| Yes | 1839 (38.11) | 1768 (39.91) | 71 (17.93) |
| No | 2,890 (59.88) | 2,572 (58.06) | 318 (80.30) |
| unknown | 97 (2.01) | 90 (2.03) | 7 (1.77) |
| Sore Throat | | | |
| Yes | 1,372 (28.43) | 1,332 (30.07) | 40 (10.10) |
| No | 3,354 (69.50) | 3,007 (67.88) | 347 (87.63) |
| unknown | 100 (2.07) | 91 (2.05) | 9 (2.27) |
| Diarrhea | | | |
| Yes | 593 (12.29) | 561 (12.66) | 32 (8.08) |
| No | 4,131 (85.60) | 3,777 (85.26) | 354 (89.39) |
| unknown | 102 (2.11) | 92 (2.08) | 10 (2.53) |
| Headache | | | |
| Yes | 2,201 (45.61) | 2,136 (48.22) | 65 (16.41) |
| No | 2,523 (52.28) | 2,202 (49.71) | 321 (81.06) |
| unknown | 102 (2.11) | 92 (2.08) | 10 (2.53) |
| Comorbidity | 4,826 (100.00) | 4,430 (91.79) | 396 (8.21) |
| Pulmonary | | | |
| Yes | 214 (4.43) | 166 (3.75) | 48 (12.12) |
| No | 4,509 (93.43) | 4,168 (94.09) | 341 (86.11) |
| unknown | 103 (2.13) | 96 (2.17) | 7 (1.77) |
| Cardiac | | | |
| Yes | 895 (18.55) | 683 (15.42) | 212 (53.54) |
| No | 3,831 (79.38) | 3,656 (82.53) | 175 (44.19) |
| unknown | 100 (2.07) | 91 (2.05) | 9 (2.27) |
| Kidney | | | |
| Yes | 44 (0.91) | 21 (0.47) | 23 (5.81) |
| No | 4,683 (97.04) | 4,318 (97.47) | 365 (92.17) |
| unknown | 99 (2.05) | 91 (2.05) | 8 (2.02) |
| Diabetes | | | |
| Yes | 381 (7.89) | 255 (5.76) | 126 (31.82) |
| No | 4,341 (89.95) | 4,079 (92.08) | 262 (66.16) |
| unknown | 104 (2.15) | 96 (2.17) | 8 (2.02) |
| Smoking | | | |
| Yes | 82 (1.70) | 45 (1.02) | 37 (9.34) |
| No | 4,640 (96.15) | 4,290 (96.84) | 350 (88.38) |
| unknown | 104 (2.15) | 95 (2.14) | 9 (2.27) |
| Obesity | | | |
| Yes | 248 (5.14) | 210 (4.74) | 38 (9.60) |
| No | 4,435 (91.90) | 4,089 (92.30) | 346 (87.37) |
| unknown | 143 (2.96) | 131 (2.96) | 12 (3.03) |

(Continued in next column)

**TABLE 1 |** (*Continued*) Demographic data and clinical characteristics of the study populationTraining Dataset (Feb 29th - May 23rd).

[a]*iES = incomplete Elementary school; cES = complete Elementary school; iHS = incomplete High school; cHS = complete High School; iHE = incomplete Higher Education; cHE = complete Higher Education.*

confirmation criteria, age range, gender, race/color, education, taking a value from a finite discrete set. On the other hand, symptoms, comorbidities, hospitalization and travelling history

**TABLE 2 |** Demographic data and clinical characteristics of the study populationValidation Dataset (May 24th - May 30th).

| | All n (%) | Cure n (%) | Death n (%) |
|---|---|---|---|
| Confirmation Criteria | 3,617 (100.00) | 3,396 (93.89) | 221 (6.11) |
| Laboratorial | 3,479 (96.18) | 3,259 (95.97) | 220 (99.55) |
| Clinical Ep | 108 (2.99) | 108 (3.18) | 0 (0.00) |
| Clinical | 12 (0.33) | 11 (0.32) | 1 (0.45) |
| Unknown | 18 (0.50) | 18 (0.53) | 0 (0.00) |
| Basic Information | 3,617 (100.00) | 3,396 (93.89) | 221 (6.11) |
| Age range | | | |
| 0-4 years old | 30 (0.83) | 29 (0.85) | 1 (0.45) |
| 5-9 years old | 23 (0.64) | 23 (0.68) | 0 (0.00) |
| 10-19 years old | 67 (1.85) | 67 (1.97) | 0 (0.00) |
| 20-29 years old | 483 (13.35) | 481 (14.16) | 2 (0.90) |
| 30-39 years old | 943 (26.07) | 937 (27.59) | 6 (2.71) |
| 40-49 years old | 825 (22.81) | 810 (23.85) | 15 (6.79) |
| 50-59 years old | 597 (16.51) | 575 (16.93) | 22 (9.95) |
| 60-69 years old | 328 (9.07) | 284 (8.36) | 44 (19.91) |
| 70-79 years old | 181 (5.00) | 130 (3.83) | 51 (23.08) |
| 80-89 years old | 114 (3.15) | 50 (1.47) | 64 (28.96) |
| 90 years old or more | 26 (0.72) | 10 (0.29) | 16 (7.24) |
| Gender | | | |
| Male | 1701 (47.03) | 1,580 (46.53) | 121 (54.75) |
| Female | 1916 (52.97) | 1816 (53.47) | 100 (45.25) |
| Race/Color | | | |
| Asian | 233 (6.44) | 204 (6.01) | 29 (13.12) |
| White | 1,176 (32.51) | 1,113 (32.77) | 63 (28.51) |
| Unknown | 699 (19.33) | 660 (19.43) | 39 (17.65) |
| Indigenous | 5 (0.14) | 5 (0.15) | 0 (0.00) |
| Brown | 1,249 (34.53) | 1,176 (34.63) | 73 (33.03) |
| Black | 255 (7.05) | 238 (7.01) | 17 (7.69) |
| Education[a] | | | |
| Illiterate | 65 (1.80) | 47 (1.38) | 18 (8.14) |
| iES (1–4 grade) | 135 (3.73) | 117 (3.45) | 18 (8.14) |
| cES (4 grade) | 75 (2.07) | 69 (2.03) | 6 (2.71) |
| iES (5–8 grade) | 165 (4.56) | 150 (4.42) | 15 (6.79) |
| cES (8 grade) | 173 (4.78) | 160 (4.71) | 13 (5.88) |
| iHS | 146 (4.04) | 136 (4.00) | 10 (4.52) |
| cHS | 971 (26.85) | 947 (27.89) | 24 (10.86) |
| iHE | 88 (2.43) | 87 (2.56) | 1 (0.45) |
| cHE | 604 (16.70) | 599 (17.64) | 5 (2.26) |
| Unknown | 1,195 (33.04) | 1,084 (31.92) | 111 (50.23) |
| Hospitalization | 3,617 (100.00) | 3,396 (93.89) | 221 (6.11) |
| Yes | 306 (8.46) | 210 (6.18) | 96 (43.44) |
| No | 2,496 (69.01) | 2,388 (70.32) | 108 (48.87) |
| unknown | 815 (22.53) | 798 (23.50) | 17 (7.69) |
| Travelling History | 3,617 (100.00) | 3,396 (93.89) | 221 (6.11) |
| Brazil | | | |
| Yes | 236 (6.52) | 231 (6.80) | 5 (2.26) |
| No | 2,500 (69.12) | 2,322 (68.37) | 178 (80.54) |
| unknown | 881 (24.36) | 843 (24.82) | 38 (17.19) |
| International | | | |
| Yes | 2 (0.06) | 2 (0.06) | 0 (0.00) |
| No | 2,435 (67.32) | 2,269 (66.81) | 166 (75.11) |
| unknown | 1,180 (32.62) | 1,125 (33.13) | 55 (24.89) |
| | **All n (%)** | **Cure n (%)** | **Death n (%)** |
| Sympton | 3,617 (100.00) | 3,396 (93.89) | 221 (6.11) |
| Fever | | | |
| Yes | 2077 (57.42) | 1928 (56.77) | 149 (67.42) |
| No | 1,478 (40.86) | 1,406 (41.40) | 72 (32.58) |
| unknown | 62 (1.71) | 62 (1.83) | 0 (0.00) |
| Respiratory Distress | | | |
| Yes | 956 (26.43) | 814 (23.97) | 142 (64.25) |
| No | 2,601 (71.91) | 2,522 (74.26) | 79 (35.75) |
| unknown | 60 (1.66) | 60 (1.77) | 0 (0.00) |
| Cough | | | |
| Yes | 2,237 (61.85) | 2096 (61.72) | 141 (63.80) |
| No | 1,317 (36.41) | 1,238 (36.45) | 79 (35.75) |
| unknown | 63 (1.74) | 62 (1.83) | 1 (0.45) |
| Runny nose | | | |
| Yes | 1,249 (34.53) | 1,212 (35.69) | 37 (16.74) |
| No | 2,305 (63.73) | 2,122 (62.49) | 183 (82.81) |
| unknown | 63 (1.74) | 62 (1.83) | 1 (0.45) |
| Sore Throat | | | |
| Yes | 975 (26.96) | 955 (28.12) | 20 (9.05) |
| No | 2,578 (71.27) | 2,378 (70.02) | 200 (90.50) |
| unknown | 64 (1.77) | 63 (1.86) | 1 (0.45) |
| Diarrhea | | | |
| Yes | 465 (12.86) | 450 (13.25) | 15 (6.79) |
| No | 3,087 (85.35) | 2,882 (84.86) | 205 (92.76) |
| unknown | 65 (1.80) | 64 (1.88) | 1 (0.45) |
| Headache | | | |
| Yes | 1,628 (45.01) | 1,588 (46.76) | 40 (18.10) |
| No | 1925 (53.22) | 1745 (51.38) | 180 (81.45) |
| unknown | 64 (1.77) | 63 (1.86) | 1 (0.45) |
| Comorbidity | 3,617 (100.00) | 3,396 (93.89) | 221 (6.11) |
| Pulmonary | | | |
| Yes | 168 (4.64) | 145 (4.27) | 23 (10.41) |
| No | 3,385 (93.59) | 3,189 (93.90) | 196 (88.69) |
| unknown | 64 (1.77) | 62 (1.83) | 2 (0.90) |
| Cardiac | | | |
| Yes | 726 (20.07) | 602 (17.73) | 124 (56.11) |
| No | 2,830 (78.24) | 2,734 (80.51) | 96 (43.44) |
| unknown | 61 (1.69) | 60 (1.77) | 1 (0.45) |
| Kidney | | | |
| Yes | 43 (1.19) | 28 (0.82) | 15 (6.79) |
| No | 3,511 (97.07) | 3,306 (97.35) | 205 (92.76) |
| unknown | 63 (1.74) | 62 (1.83) | 1 (0.45) |
| Diabetes | | | |
| Yes | 345 (9.54) | 276 (8.13) | 69 (31.22) |
| No | 3,209 (88.72) | 3,058 (90.05) | 151 (68.33) |
| unknown | 63 (1.74) | 62 (1.83) | 1 (0.45) |
| Smoking | | | |
| Yes | 60 (1.66) | 50 (1.47) | 10 (4.52) |
| No | 3,491 (96.52) | 3,281 (96.61) | 210 (95.02) |
| unknown | 66 (1.82) | 65 (1.91) | 1 (0.45) |
| Obesity | | | |
| Yes | 156 (4.31) | 141 (4.15) | 15 (6.79) |
| No | 3,378 (93.39) | 3,174 (93.46) | 204 (92.31) |
| unknown | 83 (2.29) | 81 (2.39) | 2 (0.90) |

(Continued in next column)

**TABLE 2 |** (*Continued*) Demographic data and clinical characteristics of the study populationValidation Dataset (May 24th - May 30th).

[a]*IES = incomplete Elementary school; cES = complete Elementary school; iHS = incomplete High school; cHS = complete High School; iHE = incomplete Higher Education; cHE = complete Higher Education.*

are categorized among yes/no/unknown. **Tables 1**, **2** details the dataset variables for all training and validation patients, respectively, showing their distribution among the categories,

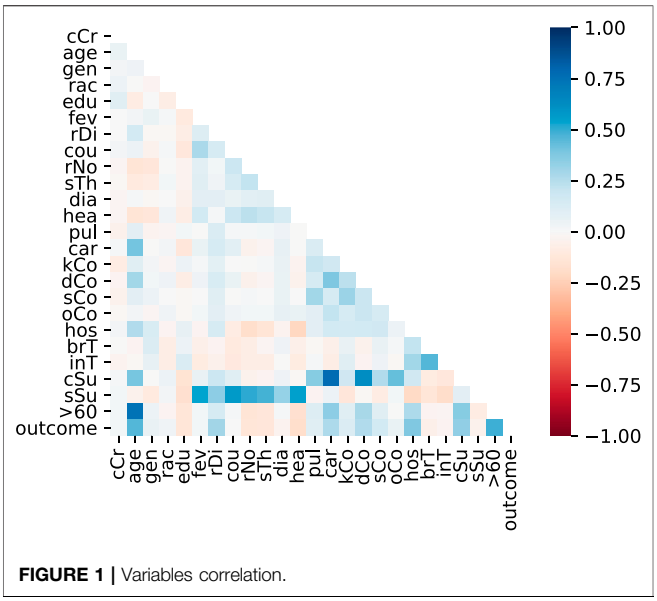


**FIGURE 1 |** Variables correlation.

as well as separated by the outcome (target variable), i.e., cure or death.

Some of the variables have unknown values due to lack of information. Instances with such a characteristic were kept in the dataset as an *unknown* category, as there was no decrease in the performance of the models due to their presence.

According to recent studies related to COVID-19, older age and the presence of comorbidities are aggravating factors that can contribute to the disease severity. In addition, the presence of two or more clinical symptoms was considered important in the COVID-19 severity (Wang et al., 2020a). Thus, in order to add more knowledge to the dataset, additional variables were developed, namely: 1) sum of the comorbidities presented by the patient, 2) sum of the symptoms presented by the patient and 3) indicative if the patient has more than 60 years old. These new variables provide information that can contribute to predict the outcome of a new COVID-19 patient. They are calculated based on already existing variables from **Tables 1**, **2**. Our final datasets contains 24 independent variables and the target variable, represented by the disease outcome[1]: cure or death.

**Tables 1**, **2** also present the distribution of the two classes, i.e., cure and death. It can be seen that we have imbalanced data, as the number of deaths corresponds only to 8.21 and 6.11% of the samples in the training and validation datasets, respectively. This difference can be a problem for machine learning models, making it difficult to predict samples of the minority class. Strategies to deal with this situation are often used, such as weighting and resampling (Santos et al., 2018). We employed an oversampling strategy, increasing the number of death samples in order to obtain a balanced dataset. A simple procedure based on randomly picking samples with replacement was performed.

**Figure 1** shows the correlation heatmap for the training dataset variables. It can be observed in the last line that some of the variables have a high correlation with the target variable, i.e., the disease outcome. They include age, respiratory distress, sum of commorbidities, hospitalization and age greater equal 60 years old. Similar correlations were found by Pourhomayoun and Shakibi (2020) regarding age and chronic diseases.

## 2.2 Machine Learning Models

Machine Learning (Mitchell, 1997) is a research area which is concerned with the question of how to construct computer programs that automatically improve with experience. Recently, many successful machine learning applications have been developed. Machine learning algorithms have proven to be of great practical value in a variety of application domains such as medical domain. They are especially useful in problems where databases may contain valuable implicit regularities that can be discovered automatically, e.g., to analyze outcomes of medical treatments from patient databases. A classification problem consists of identifying to which of a set of categories a new instance belongs, given a historical data used for training, which contains instances whose category membership is known. This type of problem is solved through supervised learning. In this paper, some supervised machine learning algorithms have been applied to a dataset having information from patients who tested positive for COVID-19 aiming to create computational models able to predict their disease outcome.

### 2.2.1 Logistic Regression (LR)

Logistic Regression (also called Logit Regression) is commonly used to estimate the probability that an instance belongs to a certain class. If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (positive class), or else it does not (negative class). This turns LR model into a binary classifier, however it can be generalized to support multiple classes (Géron, 2017). A LR model calculates a weighted sum of the input features (plus a polarization term), but instead of generating the result directly, a sigmoid (or logistic) function is applied. The sigmoid function (S-format) shows a number between 0 and 1. Once the LR model has estimated the probability that instance x belongs to the positive class, it can easily make its prediction.

### 2.2.2 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a technique for calculating a linear transformation. It takes into account class information for calculating the transformation so that the separation between classes is maximum in the new coordinate space. LDA is also known as Fisher's Linear Discriminant (Duda et al., 2000), due to the work of R. Fisher. The transformation of LDA is calculated so that the new coordinate system produces data with maximum variance between classes and minimum intraclass variance. LDA can be very effective in identifying large linearly separable problems.

### 2.2.3 Naive Bayes (NB)

The Naive Bayes is an example of an induced classifier based on strong and unrealistic assumption: all the variables are considered to be conditionally independent given the value of the class variable. Consequently, a NB classifier is automatically

[1]The disease outcome is also referred to as a "class" throughout this text.

achieved by only inducing the numerical parameters of the model. To this end, only information about the variables and their corresponding values are needed to estimate probabilities, leading to a computational time complexity that is linear with respect to the amount of training instances. NB is also space efficient, requiring only the information provided by two-dimensional tables, in which each entry corresponds to a probability estimated for a given value of a particular variable. According to Friedman et al. (1997), NB has provided good results on several domains.

### 2.2.4 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors algorithm is based on the concept of neighborhood, in which the neighbors are similar. Thus, it is possible to classify the elements of an n-dimensional space into $K$ sets. This parameter $K$ represents the number of neighbors and it is defined by the user in order to obtain a better classification. Classification is calculated based on a vote of the $K$-neighbors closest to each point (each instance of data or training example is viewed as points in space). According to Peter Norvig (2013), the classifier can get good results when there is lot of data in a low dimension (domains with few variables). However, in large dimensional spaces, usually the closest neighbors are distant.

### 2.2.5 Decision Trees (DT)

Classification and Regression Tree (CART) is an algorithm to train Decision Trees (DT) (Géron, 2017). A decision tree returns a response after executing a test sequence and it is considered one of the most successful methods of machine learning (Peter Norvig, 2013). The CART algorithm works by first splitting the training set into two subsets using a single feature $k$ and a threshold $t_k$. It searches for the pair ($k$, $t_k$) that produces the purest subsets (weighted by their size). Once the CART algorithm has successfully split the training set in two, it splits the subsets using the same logic, then the sub-subsets, and so on, recursively. It stops recursing once it reaches the maximum depth or if it cannot find a split that will reduce impurity.

### 2.2.6 XGBOOST (XGB)

XGBoost (eXtreme Gradient Boosting) is an implementation of stochastic gradient boosting. This implementation is computationally efficient with many options and is available as a package for the main data science software languages (Bruce and Bruce, 2017). The XGB library implements the gradient boosting decision tree algorithm. It was designed to be highly efficient, flexible and portable. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. XGB provides a parallel tree boosting that solves many data science problems in a fast and accurate way. This approach supports both regression and classification predictive modeling problems.

### 2.2.7 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning method which became popular in some years ago for solving problems in classification, regression, and novelty detection (Bishop, 2006). The SVM approaches the problem of finding one solution that separate the classes exactly from the training data set through the concept of the margin, which is defined to be the smallest distance between the decision boundary and any of the samples. SVM constructs a decision boundary (maximum margin separator) with the greatest possible distance to example points. The idea of SVM is to focus on points more important than others that lead to the best generalization. For this, a linear separation in hyperplane is created, even if the data are not separable linearly in the original input space, because they can incorporate the data in a space of superior dimension, using kernel trick. The linear dimension separator is actually nonlinear in the original space.

## 2.3 Evaluation Metrics

In this study, we evaluate the performance of each of the learning models in terms of accuracy, Receiver Operating Characteristic curve and area under the curve, precision, recall, Precision-Recall curve and area under the curve, F1-score and finally the confusion matrix. These metrics are detailed in the following.

1) Confusion Matrix: in a binary classification, the result on a test set is often displayed as a two-dimensional *confusion matrix* with a row and column for each class. Each matrix element shows the number of test examples for which the actual class is the row and the predicted class is the column. Good results correspond to large numbers down the main diagonal and small, ideally zero, off-diagonal elements (Witten et al., 2011). The scheme of a confusion matrix is illustrated below.

| | | predicted class | |
|---|---|---|---|
| | | **Cure** | **Death** |
| actual class | Cure | TN | FP |
| | death | FN | TP |

TN = true negative, FP = false positive.
FN = false negative, TP = true positive

2) Accuracy: it is the ratio of the number of correct predictions to the total number of samples. It works well when there are equal number of samples belonging to each class. However, accuracy is misleading for skewed class distribution since correct predictions for the minority class can fully ignored. It can be given by:

$$accuracy = \frac{TN + TP}{TN + FP + TP + FN} \tag{1}$$

3) Precision: also known as the positive predictive value, precision is defined as the proportion of positive examples that are truly positive. A precise model will only predict the positive class in cases very likely to be positive. This metric can be calculated by following formula:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

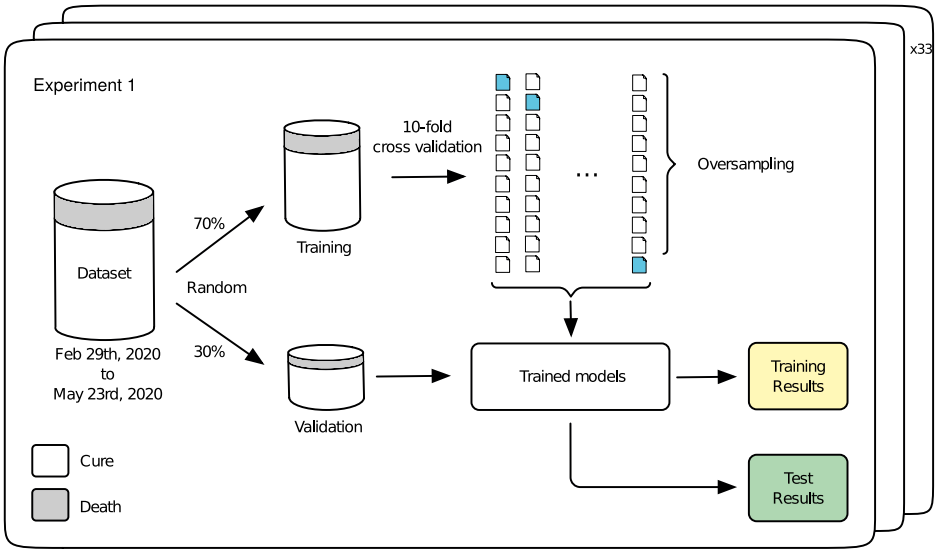


**FIGURE 2 |** Experiment 1.

4) Recall: also known as sensitivity, is a measure of how complete the results are. A model with high recall captures a large portion of the positives examples, meaning that it has wide breadth. It is calculated as:

$$recall = \frac{TP}{TP + FN} \tag{3}$$

5) F1-score: this metric seeks a balance between precision and recall and represents an interesting metric when there is an uneven class distribution. It is given by the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

6) Receiver Operating Characteristic (ROC) curve: ROC curves are a graphical technique for evaluating the performance of a binary classifier at different classification thresholds. These curves depict the performance of a classifier without regard to class distribution or error costs. They plot the TP rate on the vertical axis against the FP rate on the horizontal axis.
7) Precision-Recall (PR) curve: a precision-recall curve shows the relationship between precision (positive predictive value) and recall (sensitivity) for every possible cut-off. It represents an alternative to a ROC curve and shows a good choice to compare models when facing imbalanced datasets. The main difference between ROC curves and PR curves is that the number of TN results is not used to make a PR curve.
8) Area Under the Curve (AUC): AUC measures the entire two dimensional area underneath an entire curve. Therefore, it gives an aggregate measure of performance in a single value. AUC ranges from 0.0 to 1.0; a model with predictions 100% correct has an AUC of 1.0 while one whose predictions are 100% wrong gives an AUC of 0.0. We use AUC values for ROC and PR curves in our experiments.

# 3 RESULTS

Our experimental design involves two main parts. The first, named *Experiment 1*, consists of a series of repeated tests using the training dataset, while the second, *Experiment 2*, performs a final test by using both the training and validation datasets.

The first experiment includes 4, 826 patients (46.25% male and 53.75% female), distributed in different age groups, educational level, race/color, hospitalization and travelling history. Of the total number of patients, 91.79% cured and 8.21% deceased. **Table 1** shows seven symptoms and six comorbidities present/absent among those who cured or died. The second experiment includes a new dataset containing 3, 617 patients (47.03% male and 52.97% female) of which 93.89% cured and 6.11% died (**Table 2**).

Both datasets present similar distributions among the different variables' categories. However, it is possible to observe the increase in the number of confirmed (closed) cases, since the validation dataset comprises only 7 days and has only 25% fewer samples than the training dataset that corresponds to a period of 75 days. These numbers reflect the rapid progress of cases in Brazil.

## 3.1 Experiment 1

Experiment 1 is developed to evaluate the performance of the different prediction models under a series of repeated tests using different partitions of the training dataset (**Table 1**). The idea underlying this experiment is illustrated in **Figure 2**. A 70-30 split is performed in the dataset through a random, but stratified procedure. The 70-part is used for training by a 10-fold cross

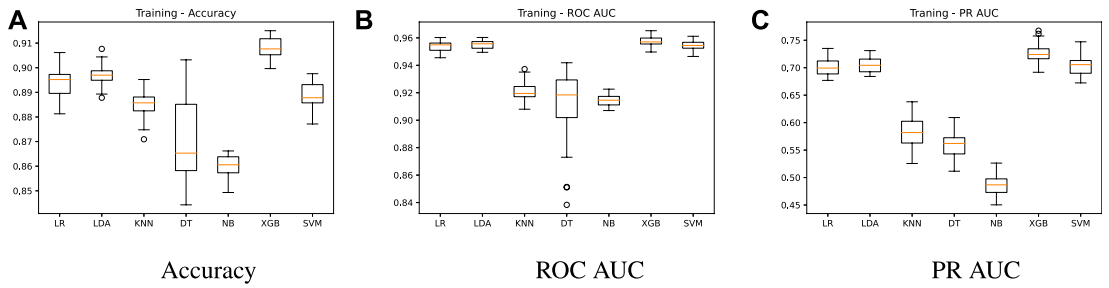


**FIGURE 3 |** Training results - algorithms comparison.

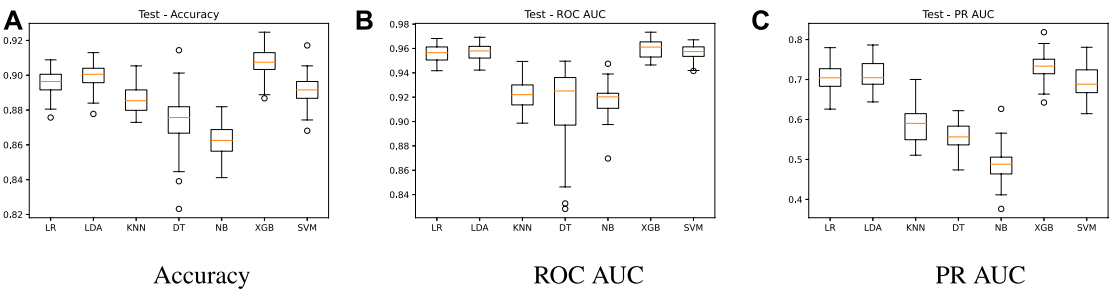


**FIGURE 4 |** Test results - algorithms comparison.

validation, with oversampling applied only in the training folds (9 of them), generating the training results, i.e., an estimate of the performance of the models. Once we have the trained models, the 30-part is used for validation, leading to the test results. This procedure is repeated 33 times and the results are reported in **Figures 3**, **4**. Grid search was used in order to find the best hyper-parameters for the models.

**Figure 3** shows summarized training results for the seven prediction models regarding accuracy, ROC AUC and PR AUC metrics. A higher performance for the three metrics is achieved by LR, LDA, XBG and SVM models, with ROC AUC mean values of 0.95, 0.95, 0.96 and 0.95, respectively (**Figure 3B**). On the other hand, models KNN, DT and NB present a ROC AUC of 0.92, 0.91 and 0.91, respectively. In **Figure 3C**, we present the precision-recall AUC, which does not consider the true negative results, giving a higher importance to the minority class. The PR AUC mean value is around 0.7 for the best models (LR = 0.70, LDA = 0.70, XGB = 0.72 and SVM = 0.70), while models KNN, DT and NB have 0.58, 0.55 and 0.48 mean values. As mentioned before, in this study the accuracy is not the best metric to compare the different models. According to **Figure 3A**, LR, LDA, XBG and SVM models achieve accuracy mean values of 0.89, 0.90, 0.91, 0.89, while for KNN, DT and NB we have 0.88, 0.87 and 0.86, respectively. As we can see, although the accuracy values do not present significant differences among the models, the other two metrics (ROC AUC and PR AUC) make performance differences more evident. The DT model has a lower robustness due to a higher dispersion while the NB model presents the worse performance.

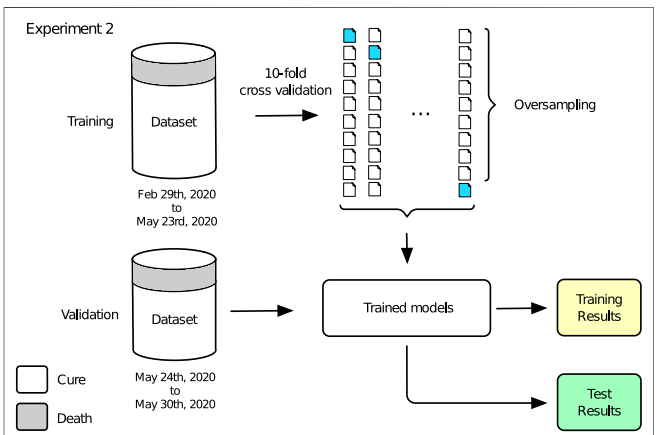


**FIGURE 5 |** Experiment 2.

In **Figure 4**, we show the summarized test results using the 30-part from the split and the trained models. This evaluation aims to check whether the models were able to learn from the data and achieve good predictions with unseen data. A similar behavior is observed for all metrics, keeping LR, LDA, XGB and SVM as the best models. The accuracy remains high, with mean values above 0.89 for these models (**Figure 4A**), while the ROC AUC is around 0.95 (**Figure 4B**). A PR AUC of 0.7 (**Figure 4C**) on average is also achieved by the test results. These results verify the learning capacity of the models and the absence of overfitting, as there was no significant drop in performance in the test results.

**TABLE 3 |** Experiment 2 TrainingModels performance.

| Model | Accuracy | ROC AUC | PR AUC | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| LR | 0.88 | 0.96 | 0.73 | 0.46 | 0.88 | 0.59 |
| LDA | 0.89 | 0.96 | 0.74 | 0.47 | 0.87 | 0.59 |
| KNN | 0.87 | 0.92 | 0.58 | 0.41 | 0.83 | 0.54 |
| DT | 0.80 | 0.93 | 0.59 | 0.38 | 0.90 | 0.52 |
| NB | 0.85 | 0.92 | 0.52 | 0.39 | 0.83 | 0.51 |
| XGB | 0.81 | 0.95 | 0.69 | 0.39 | 0.90 | 0.52 |
| SVM | 0.88 | 0.96 | 0.74 | 0.45 | 0.89 | 0.58 |

**TABLE 4 |** Experiment 2 TestModels performance.

| Model | Accuracy | ROC AUC | PR AUC | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| LR | 0.86 | 0.92 | 0.50 | 0.28 | 0.86 | 0.42 |
| LDA | 0.86 | 0.92 | 0.51 | 0.28 | 0.85 | 0.43 |
| KNN | 0.85 | 0.88 | 0.45 | 0.27 | 0.80 | 0.40 |
| DT | 0.82 | 0.90 | 0.50 | 0.24 | 0.87 | 0.38 |
| NB | 0.82 | 0.89 | 0.32 | 0.23 | 0.82 | 0.37 |
| XGB | 0.82 | 0.92 | 0.51 | 0.24 | 0.88 | 0.38 |
| SVM | 0.85 | 0.92 | 0.50 | 0.27 | 0.85 | 0.41 |

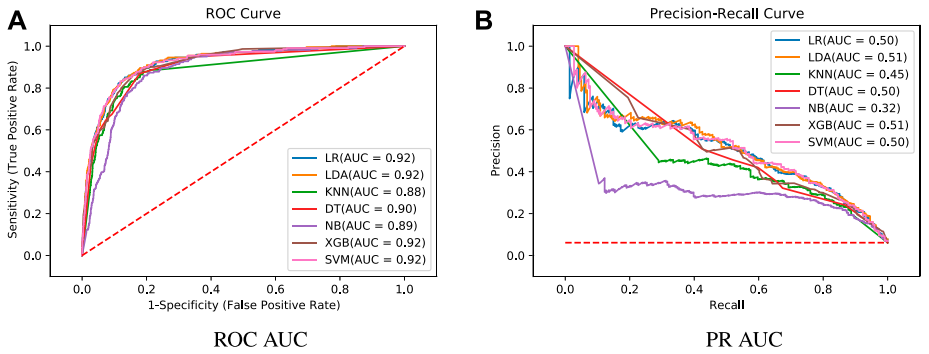


**FIGURE 6 |** Test results - algorithms comparison.

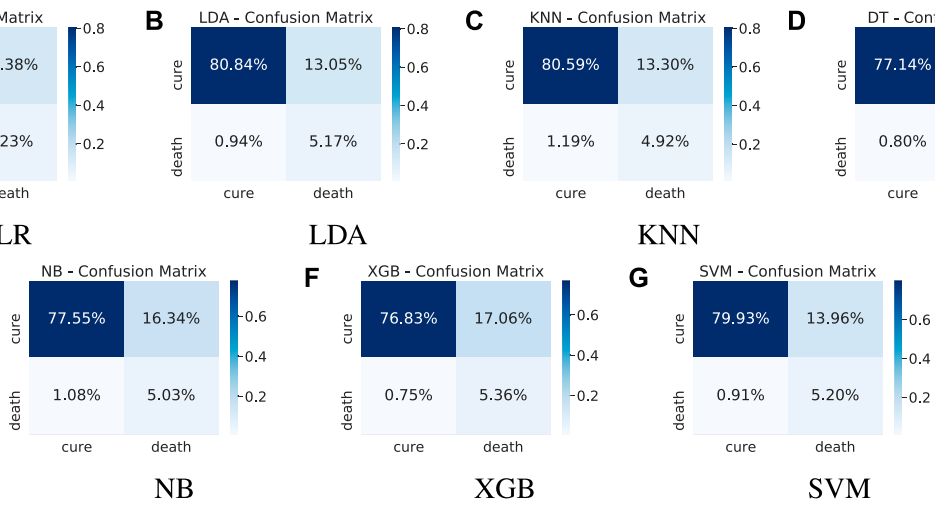


**FIGURE 7 |** Test results - confusion matrices.

## 3.2 Experiment 2

The second part of our experimental design concerns a validation test using new data from **Table 2**. For that, algorithms were trained using all data from the training dataset, i.e., closed cases included in the database until May 23rd, 2020 while prediction was performed for the validation dataset. The main difference of experiments 1 and 2 rely on the validation data being sequential samples for *Experiment 2*, while in *Experiment 1* validation sets correspond to samples randomly selected from the whole dataset. This temporal aspect may show importance since future predictions will be made for new patients (sequential samples) being tested positive for COVID-19.

The general conception of *Experiment 2* is shown in **Figure 5**. A training step is performed using the whole training dataset through 10-fold cross validation and oversampling, producing the training results. With the trained models in hands, a

**TABLE 5 |** Hyperparameter tuning in validation tests.

| Model | Parameters |
|---|---|
| LR | C = 100, penalty = l2, solver = liblinear, multi_class = ovr |
| LDA | solver = svd |
| KNN | n_neighbors = 19, weights = distance, metric = euclidean |
| DT | max_depth = 3 |
| NB | - |
| XGB | max_depth = 3, n_estimators = 100, learning_rate = 0.01, subsample = 1 |
| SVM | kernel = rbf, C = 10, gamma = 0.0001 |

validation step is developed, making predictions for the whole validation dataset and leading to the final test results. Training results are reported on **Table 3** while final tests results are shown through **Figures 6**, **7**. Again, grid search was used in order to find the best hyper-parameters for the models.

The training results of *Experiment 2* are detailed in **Table 3**, presenting the following metrics: accuracy, ROC AUC, PR AUC, Precision, Recall and F1-score. All metric results are consistent with the training results of the former experiment. Again, the best models regarding ROC AUC and PR AUC are LR, LDA, XBG and SVM. It is important to mention that the minority class has high importance in our application, leaving the accuracy as a secondary metric. One can note that although KNN and NB models have high accuracy values, they present the worst recall values.

On the analysis of the test results, it can be seen in **Figure 6** the representation of the ROC curve (**Figure 6A**) and PR curve (**Figure 6B**). Detailed results for the whole set of metrics are shown in **Table 4**. **Table 5** presents the best parameters used in these final experiments. Comparing the ROC AUC values from **Table 3** and **Figure 6A**, a slight drop in values can be noted for all models, with a decrease of 0.037 on average. This behaviour is expected since we are using completely new data, but no significant difference that indicates overfitting or a poor learning step. Most of the models have a very close performance, making it difficult to select one as the best model. From **Figure 6B**, a greater difference among the models can be noted for the PR AUC metric. Models KNN and NB show a clearly underperformance compared to the other models. In general, this metric shows an inferior performance in the tests in relation to the training step for all models. This can be due to the difference on the number of samples in the minority class. While the training dataset has 8.21% of the samples in death class, the validation dataset is even more imbalanced, with only 6.11% of the samples belonging to that class.

Moreover, it is important to mention that unlike ROC AUC, whose baseline is 0.5 (random guess classifier), in PR AUC the baseline is dependent on the application itself. In the case of this work, a random estimator would have a PR AUC of 0.06 (6.11% in death class, see the horizontal line in **Figure 6B**) and therefore, values around 0.5 are definitely a substantial increase.

From **Table 4** and **Figure 6B**, we can observe that although the results show a high recall, they present a relatively low precision. This means the minority class (death) is well detected but the model also include points of the other class (cure) in it. This fact is confirmed by the confusion matrices introduced in **Figure 7**. We can note an expressive amount of false positives samples, represented in the upper right corner. False positives concern patients who cured but were wrongly classified as deaths by the models. It is possible to note that all models present a similar behavior regarding this wrong prediction. Aiming to find an explanation for this behavior, we decided to analyse the characteristics of these samples looking for similarities.

Analysis of the characteristics of false positive patients shows that such patients had at least one of the following critical situations: > 60 years old, respiratory distress, some comorbidity and hospitalization (see **Table 6**). In an attempt to show that such patients may have been critical cases, **Table 6** shows the percentage of patients who have the characteristics most related to the chance of death according to the calculation of odds ratio from **Tables 7**, **8**.

**Table 7** shows that the death chance is greater among COVID-19 patients over 60 years old. Respiratory distress and comorbidities such as kidney disease, diabetes, cardiac disease and obesity, as well as smoking, increase the likelihood of death from COVID-19. On the other hand, runny nose, sore throat, diarrhea and headache were less likely to occur in patients who deceased. The validation cohort (**Table 8**) showed similar results, with the exception of the fever symptom, which was more likely to occur among patients who died compared to those who cured.

A brief review of other literature works is presented in the next section along with a discussion on how this work can help in the current scenario of COVID-19 in Brazil.

## 4 DISCUSSION

Prediction models of the prognosis for a given disease have the main objective of supporting the physician's decision-making about what is the best measure of patient referral, assisting in the

**TABLE 6 |** Critical conditions present in false positive patients.

| Condition | LR | LDA | KNN | DT | NB | XGB | SVM | AVG |
|---|---|---|---|---|---|---|---|---|
| >60 years old | 64.05 | 74.36 | 61.95 | 78.22 | 55.84 | 76.82 | 70.30 | 68.79 |
| Respiratory Distress | 59.30 | 51.69 | 54.68 | 36.30 | 35.70 | 37.44 | 51.49 | 46.66 |
| Comorbidity | 60.74 | 61.23 | 62.16 | 52.48 | 75.30 | 53.32 | 59.41 | 60.66 |
| Hospitalization | 34.30 | 33.05 | 29.94 | 34.65 | 20.64 | 34.04 | 34.46 | 31.58 |
| At least one condition | 100.00 | 100.00 | 98.34 | 100.00 | 90.52 | 100.00 | 100.00 | 98.41 |

**TABLE 7 |** Odds ratio for training dataset.

| Condition | OR | 95% CI | *p* |
|---|---|---|---|
| >60 years old | 25.30 | (19.83–32.34) | <0.0001 |
| Fever | 1.16 | (0.93–1.14) | = 0.1835 |
| Respiratory Distress | 8.34 | (6.64–10.48) | <0.0001 |
| Cough | 1.00 | (0.81–1.25) | = 0.9710 |
| Runny nose | 0.27 | (0.21–0.35) | <0.0001 |
| Sore Throat | 0.26 | 0.18–0.36 () | <0.0001 |
| Diarrhea | 0.61 | (0.41–0.88) | = 0.0090 |
| Headache | 0.21 | (0.16–0.27) | <0.0001 |
| Pulmonary Disease | 3.53 | (2.51–4.96) | <0.0001 |
| Cardiac Disease | 6.48 | (5.22–8.05) | <0.0001 |
| Kidney | 12.96 | (7.10–23.60) | <0.0001 |
| Diabetes | 7.70 | (6.01–9.85) | <0.0001 |
| Smoking | 10.08 | (6.40–15.80) | <0.0001 |
| Obesity | 2.14 | (1.49–3.07) | <0.0001 |
| Hospitalization | 23.88 | (18.50–30.82) | <0.0001 |

**TABLE 8 |** Odds ratio for validation dataset.

| Condition | OR | 95% CI | *p* |
|---|---|---|---|
| >60 years old | 23.45 | (16.71–32.91) | <0.0001 |
| Fever | 1.51 | (1.13–2.02) | = 0.0053 |
| Respiratory Distress | 5.57 | (4.18–7.41) | <0.0001 |
| Cough | 1.05 | (0.79–1.40) | = 0.7159 |
| Runny nose | 0.35 | (0.25–0.51) | < 0.0001 |
| Sore Throat | 0.25 | (0.16–0.40) | < 0.0001 |
| Diarrhea | 0.47 | (0.27–0.80) | = 0.0054 |
| Headache | 0.24 | (0.17–0.35) | < 0.0001 |
| Pulmonary Disease | 2.58 | (1.62–4.10) | = 0.0001 |
| Cardiac Disease | 5.87 | (4.43–7.77) | < 0.0001 |
| Kidney | 8.63 | (4.54–16.43) | < 0.0001 |
| Diabetes | 5.06 | (3.71–6.90) | < 0.0001 |
| Smoking | 3.12 | (1.56–6.25) | = 0.0013 |
| Obesity | 1.65 | (0.95–2.87) | = 0.0730 |
| Hospitalization | 10.11 | (7.42–13.77) | < 0.0001 |

screening of patients at high risk of progressing to severe disease. Artificial intelligence models aiming to identify risk factors for prognostic prediction of severe COVID-19 have been developed using age, clinical characteristics, laboratory tests and chest imaging (Wang et al., 2020b; Chen et al., 2020; Gong et al., 2020; Jiang et al., 2020; Xie et al., 2020; Yan et al., 2020).

A study using age, hypertension history and coronary heart disease showed good discriminatory power (AUC = 0.83) between COVID-19 surviving and non-surviving patients. The inclusion of biochemical data increased (AUC = 0.88) the discriminatory power (Wang et al., 2020b). Those results refer to the validation cohort consisting of 44 patients.

Blood parameters were also used to select predictive biomarkers of mortality through machine learning. Lactate dehydrogenase (LDH), lymphocytes and high-sensitivity C-reactive protein (hsCRP) proved to be good indicators for early predicting the degree of COVID-19 severity, with >90% accuracy (Yan et al., 2020).

A machine learning study involving a cohort of 214 non-severe and 148 severe patients with COVID-19 found >90% prediction accuracy for disease severity using symptom/comorbidities data. The addition of biochemical data to symptoms/comorbidities achieved >99% predictive accuracy. Therefore, it was suggested that symptoms and comorbidities can be used in an initial screening and the biochemical data inclusion could predict the severity degree and assist in the development of treatment plans (Chen et al., 2020).

Importantly, in relation to our study, the sample size in the aforementioned studies was limited, since they were carried out with data of the beginning of the pandemic. In fact, the sample size can influence the robustness of the models performance. Larger datasets provide a better training stage, potentially leading to better performance in prediction.

Although some studies have pointed out changes in blood parameters such as lymphopenia, neutrophilia and increased lactate dehydrogenase concentration (Huang et al., 2020; Kermali et al., 2020), as well as changes in the chest images (Shi et al., 2020) as good indicators of the disease severity, these data are not publicly available and were not included in our study due to lack of access. In future studies we intend to include such data and check if there is an improvement in models performance. In view of the costs and difficulties of performing laboratory and chest imaging exams for an alarmingly increasing number of patients, our study proves to be important in that it is able to differentiate those critically ill patients who need ICU care using less complex approaches, that is, age, symptoms and comorbidities at the time of screening.

Moreover, the deeper analysis of the characteristics of false positive patients shows that such patients had at least one critical condition related to a more severe disease. It is tempting to speculate that, in a way, this percentage of patients predicted in the model could be those with the most critical condition, but that due to early and effective care were cured. This could even be a positive aspect of the prediction models, since it is important to identify severe cases which deserve special care. Unfortunately, it is not possible to confirm such a hypothesis since the database does not provide information to differentiate mild and severe cases from those who have been cured. Additionally, odds ratio results are similar to those reported by Chen et al. (2020) in severe COVID-19 patients compared to non-severe patients, emphasizing a high probability of complication in patients with comorbidities.

Brazil has an unified health system, namely Sistema Único de Saúde (SUS), that allows for almost universal health coverage across the country, despite regional inequalities. With the growing number of COVID-19 cases, the TeleSUS system (Ministry of Health Minist, 2020) was implemented on May 2020 as a strategy to perform remote preclinical health care and avoid unnecessary travel and the exhaustion of on-site health services. In this context, our study could also assist in screening those who may need early care or hospitalization solely through reports of personal data, symptoms and comorbidities. This model can be applied in other localities that have overloaded healthcare systems. Moreover, this model can also help in understanding the upcoming demand for ICU beds, staff and other critical resources.

Finally, it is important to highlight that this study was based only on a database from a State (Espírito Santo) of Brazil, requiring application in other States, since regional variations can occur in a

country with continental characteristics such as Brazil. As future works, we intend to evaluate other machine learning models such as deep learning and other ensemble learning methods.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://coronavirus.es.gov.br/painel-covid-19-es.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

FS and NH-S conceived the original idea. CS and DG prepared the data. FS and ES designed and performed the experiments, derived the models and analysed the data. FS and NH-S wrote the manuscript with input from all authors. All authors discussed the results and contributed to the final manuscript.

## FUNDING

This work was partially supported by CNPq and FAPEMIG.

## ACKNOWLEDGMENTS

We would like to thank the Health Department of Espírito Santo for placing the COVID-19 database open access. This manuscript has been released as a pre-print at https://doi.org/10.1101/2020.06.26.20140764, (Souza et al., 2020).

## REFERENCES

Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., and Cheng, X. (2020). Artificial Intelligence and Machine Learning to Fight Covid-19. *Physiol. Genomics* 52, 200–202. doi:10.1152/physiolgenomics.00029.2020

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Brazilian Institute of Geography and Statistics (2020). Instituto brasileiro de geografia e estatística (ibge). espírito santo. Available at: https://cidades.ibge.gov.br/brasil/es/panorama. (Accessed June 01, 2020).

Bruce, P., and Bruce, A. (2017). *Practical Statistics for Data Scientists*. Sebastopol, CA: O'Reilly Media.

Burki, T. (2020). Covid-19 in Latin america. *Lancet Infect. Dis.* 20, 547–548. doi:10.1016/s1473-3099(20)30303-0

Chen, Y., Ouyang, L., Bao, S., Li, Q., Han, L., Zhang, H., et al. An Interpretable Machine Learning Framework for Accurate Severe vs Non-severe Covid-19 Clinical Type Classification. medRxiv (2020). doi:10.1101/2020.05.18.20105841

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. 2nd Edition. Wiley-Interscience.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learn.* 29, 131–163. doi:10.1023/a:1007465528199

Fu, L., Wang, B., Yuan, T., Chen, X., Ao, Y., Fitzpatrick, T., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 (Covid-19) in china: A Systematic Review and Meta-Analysis. *J. Infect.* 80, 656–665. doi:10.1016/j.jinf.2020.03.041

Gao, F., Zheng, K. I., Wang, X.-B., Sun, Q.-F., Pan, K.-H., Wang, T.-Y., et al. (2020). Obesity Is a Risk Factor for Greater Covid-19 Severity. *Dia Care* 43, e72–e74. doi:10.2337/dc20-0682

Géron, A. (2017). *Hands-on Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA: O'Reilly Media.

Gong, J., Ou, J., Qiu, X., Jie, Y., Chen, Y., Yuan, L., et al. (2020). A Tool for Early Prediction of Severe Coronavirus Disease 2019 (COVID-19): A Multicenter Study Using the Risk Nomogram in Wuhan and Guangdong, China. *Clin. Infect. Dis.* 71, 833–840. doi:10.1093/cid/ciaa443.Ciaa443

Government of the state of the holy spirit (2020). *Governo Do Estado Do Espírito Santo. Covid-19 - Painel Covid-19 - Estado Do Espírito Santo* (Accessed June 01, 2020).

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in china. *N. Engl. J. Med.* 382, 1708–1720. doi:10.1056/NEJMoa2002032

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, china. *The Lancet* 395, 497–506. doi:10.1016/S0140-6736(20)30183-5

Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., et al. (2020). Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity. *Comput. Mater. Continua* 62, 537–551. doi:10.32604/cmc.2020.01069110.32604/cmc.2020.010691

Jin, Y., Yang, H., Ji, W., Wu, W., Chen, S., Zhang, W., et al. (2020). Virology, Epidemiology, Pathogenesis, and Control of Covid-19. *Viruses* 12, 372. doi:10.3390/v12040372

Kermali, M., Khalsa, R. K., Pillai, K., Ismail, Z., and Harky, A. (2020). The Role of Biomarkers in Diagnosis of COVID-19 - A Systematic Review. *Life Sci.* 254, 117788. doi:10.1016/j.lfs.2020.117788

Li, H., Liu, S.-M., Yu, X.-H., Tang, S.-L., and Tang, C.-K. (2020). Coronavirus Disease 2019 (Covid-19): Current Status and Future Perspectives. *Int. J. Antimicrob. Agents* 55, 105951. doi:10.1016/j.ijantimicag.2020.105951

Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., et al. (2020). Real-time Tracking of Self-Reported Symptoms to Predict Potential Covid-19. *Nat. Med.* 26, 1037–1040. doi:10.1038/s41591-020-0916-2

Ministry of Health (2020). Ministério da saúde. telesus. Available at: https://coronavirus.saude.gov.br/telesus. (Accessed June 08, 2020).

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Education.

Peter Norvig, S. R. (2013). *Inteligência Artificial*. third edn. Elsevier Editora.

Pourhomayoun, M., and Shakibi, M. Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making. medRxiv (2020). doi:10.1101/2020.03.30.20047308

Sanders, J. M., Monogue, M. L., Jodlowski, T. Z., and Cutrell, J. B. (2020). Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19). *JAMA* 323, 1824–1836. doi:10.1001/jama.2020.6019

Santos, M. S., Soares, J. P., Abreu, P. H., Araújo, H., and Santos, J. (2018). Cross-validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [research Frontier]. *IEEE Comput. Intell. Mag.* 13, 59–76. doi:10.1109/mci.2018.2866730

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., et al. (2020). Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for Covid-19. *IEEE Rev. Biomed. Eng.*

Siddiqi, H. K., and Mehra, M. R. (2020). Covid-19 Illness in Native and Immunosuppressed States: A Clinical-Therapeutic Staging Proposal. *J. Heart Lung Transplant.* 39, 405–407. doi:10.1016/j.healun.2020.03.012

Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., et al. (2020). World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (Covid-19). *Int. J. Surg.* 76, 71–76. doi:10.1016/j.ijsu.2020.02.034

Souza, F. S. H., Hojo-Souza, N. S., Santos, E. B., Silva, C. M., and Guidoni, D. L. Predicting the Disease Outcome in Covid-19 Positive Patients through Machine Learning: a Retrospective Cohort Study with Brazilian Data. medRxiv (2020). doi:10.1101/2020.06.26.20140764

Wang, B., Li, R., Lu, Z., and Huang, Y. (2020). Does Comorbidity Increase the Risk of Patients with Covid-19: Evidence from Meta-Analysis. *Aging (Albany NY)* 12, 6049–6057. doi:10.18632/aging.10300010.18632/aging.103000

Wang, K., Zuo, P., Liu, Y., Zhang, M., Zhao, X., Xie, S., et al. (2020). Clinical and Laboratory Predictors of In-Hospital Mortality in Patients with Coronavirus Disease-2019: A Cohort Study in Wuhan, China. *Clin. Infect. Dis.* 71, 2079–2088. doi:10.1093/cid/ciaa538.Ciaa538

Witten, I. H., Frank, E., and Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools And Techniques. *Morgan Kaufmann Series in Data Management Systems*. 3 edn. Amsterdam: Morgan Kaufmann).

World health organization (2020). Coronavirus Disease (Covid-19) Outbreak Situation. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. (Accessed June 08, 2020).

World health organization (2020). Who Announces Covid-19 Outbreak a Pandemic. Available at: http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic. (Accessed May 22, 2020).

Xie, J., Hungerford, D., Chen, H., Abrams, S. T., Li, S., Wang, G., et al. Development and External Validation of a Prognostic Multivariable Model on Admission for Hospitalized Patients with Covid-19. medRxiv (2020). doi:10.1101/2020.03.28.20045997

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020). An Interpretable Mortality Prediction Model for Covid-19 Patients. *Nat. Mach Intell.* 2, 283–288. doi:10.1038/s42256-020-0180-7

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., et al. (2020). Clinical Course and Risk Factors for Mortality of Adult Inpatients with Covid-19 in Wuhan, china: a Retrospective Cohort Study. *The Lancet* 395, 1054–1062. doi:10.1016/S0140-6736(20)30566-310.1016/s0140-6736(20)30566-3

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.