PREDICTION OF COVID-19 FROM HEMOGRAM RESULTS AND AGE USING MACHINE LEARNING

Elena Caires Silveira

1 Medical Student at Multidisciplinary Institute for Health, Federal University of Bahia (Universidade Federal da Bahia), Brazil.

ABSTRACT

Article type: Research

Article History:
Received: 2020-08-17
Revised: 2020-08-19
Accepted: 2020-08-21

Introduction:
The rapid global dissemination of COVID-19 culminated in the mobilization of great technological efforts aimed at its better understanding and control. In this paper, Machine Learning gains notoriety, and its application has been widely documented for pathophysiological, diagnostic, therapeutic, prognostic and monitoring of COVID-19 purposes. The present paper aimed to build a model for the prediction of the diagnosis of COVID-19 based on blood count results and age of patients and to identify the main characteristics taken into account by the algorithm for the predictive decision.

Material and Methods:
Anonymous data from 1157 patients made available by the COVID-19 Data Sharing / BR repository were used. The work took place in two distinct stages: description and analysis of the data; and construction of the predictive model.

Results:
With the exception of hemoglobin measurement, mean corpusrus volume, red cell distribution width, mean platelet volume and neutrophil-lymphocyte ratio, there was a statistically significant association of all other hematological parameters assessed with COVID-19. The predictive model developed from the XGBoost classifier reached an accuracy of 80.0% with a sensitivity of 75.6% and specificity of 82.0%. The variables that had the greatest influence on the predictive decision were basophil, eosinophil and leukocyte measurements.

Conclusion:
The present paper confirms the potential of using blood count results, a widely available and accessible test, in the context of the diagnostic evaluation and pathophysiological investigation of COVID-19. Conclusion:
This work highlights the relevance of the systematization and dissemination of data related to COVID-19 for use in new research.

Keywords:

INTRODUCTION

In late 2019, a new type of coronavirus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes the pandemic disease COVID-19, emerged in Wuhan province, Hubei, China [1]. Coronaviruses belong to the Coronaviridae family and members of Group IV ((+) ssRNA) in the Baltimore classification and are a group of enveloped and spiked surface viruses capable of infecting humans and a wide variety of animals [2, 3]. SARS-CoV-2 belongs to line B of beta-corona viruses and has a high degree of homology with previous SARS-CoV and MERS-CoV [4, 5]. Such virus enters host cells via endocytosis after binding to the receptor ACE2, and has a variable incubation period of 1 to 14 days [6].

Viral transmission between individuals can occur through direct, close or indirect contact, through
infected secretions or droplets [7-9]. Such secretions and droplets can still contaminate surfaces and objects, thus enabling transmission by fomites [10, 11]. In addition, SARS-CoV-2 airborne transmission was observed during medical procedures that lead to aerosol production, and the possibility of this type of transmission in the absence of aerosol-generating procedures is being investigated [12-14].

COVID-19 has a wide clinical spectrum, ranging from asymptomatic patients to cases of pneumonia, severe acute respiratory syndrome and multiple organ dysfunction [15]. In symptomatic patients, the most common manifestations are fever, cough, dyspnea, fatigue, diarrhea, muscle pain and headache [15, 16]. Despite manifesting mainly as a respiratory infection, there is evidence that COVID-19 involves multiple systems, including the cardiovascular, respiratory, gastrointestinal, neurological, hematopoietic and immune systems [17].

The rapid global spread of COVID-19, recognized by the World Health Organization (WHO) as a pandemic in March 2020 [18], culminated in the mobilization of major technological efforts aimed at its better understanding and control. In this context, the modality of Artificial Intelligence known as Machine Learning gains notoriety, and its application has been widely documented for pathophysiological, diagnostic, therapeutic, prognostic and monitoring of COVID-19 purposes [19-23].

Machine Learning is a scientific discipline arising from the intersection between statistics and computer science that focuses on the development of algorithms capable of learning from data through pattern recognition [24, 25]. In supervised learning, algorithms are trained from data sets composed of instances and labels (training sets), so that they can later be used to predict unknown labels for new instances provided, this predictive ability being verified from test sets, where the labels are not provided to the algorithm [25].

In the last decade, Machine Learning has aroused the interest of researchers and physicians due to its great possibility of contributing to Medicine and, conversely, the large volume of data generated by medical activities has aroused the interest of computer scientists. As highlighted by Cabitza and Banfi [26], the potential for applying Machine Learning to laboratory data for diagnostic and prognostic purposes deserves attention.

Considering the above, the present paper proposed to build a prediction model using the supervised classifier XGBoost for the diagnosis of COVID-19 based on blood count results, age and sex of patients and to identify the main characteristics (instances) taken into account by the algorithm for the predictive decision. The XGBoost is an efficient, flexible and portable library which implements machine learning algorithms providing decision-tree-based ensembling under the Gradient Boosting framework [27]. The choice of hematological data as instances for the prediction is supported by the identification of the correlation between hematological parameters and the diagnosis of COVID-19 in recent studies, as well as in the wide availability of the blood count, characterized by being a practical and accessible exam.

**MATERIAL AND METHODS**

The present paper aimed to construct a classification algorithm oriented to the prediction of the diagnosis of COVID-19 based on parameters from the blood count, age and sex of the patients. For this purpose, anonymous data from 1157 patients made available by the repository COVID-19 Data Sharing/BR were used, referring to patients who underwent laboratory tests for the diagnosis of COVID-19 in the Fleury Network from May 29, 2020 to June 15, 2020.

The hematological parameters considered were: hemoglobin measurement (g/dL); erythrocytes (millions/mm3); hematocrit (%); mean corpuscular volume, or MCV (fL); mean corpuscular hemoglobin or MCH (pg); mean corpuscular hemoglobin concentration, or MCHC (g/dL); red cell distribution width, or RDW (%); global leukotic count (mm3); neutrophil count (mm3); lymphocyte count (mm3); eosinophil count (mm3); basophil count (mm3); monocyte count (mm3); platelet count (mm3); mean platelet volume, or MPV (fL); and neutrophil-lymphocyte ratio, or NLR. COVID-19 confirmed cases were those for which the result of the polymerase chain reaction test after reverse transcription (RT-PCR) for SARS-CoV-2 performed from a sample obtained from a nasopharyngeal swab was positive. Such a test is the gold standard method for diagnosing the disease. Only patients with data available for all evaluated variables were included in the study, with a maximum difference of 01 days between the collections of materials for both exams being admitted. After the collection, the work study place in two distinct stages: description and analysis of the data; and construction of the predictive model.

For the descriptive analysis of the data, the patients were grouped into positive and negative for COVID-19 according to the result of the RT-PCR test for SARS-CoV-2. Categorical variables were represented as proportions and continuous variables were represented as means and standard deviations within each group of patients. The proportions for categorical variables were compared between groups using the χ2 test. For continuous variables, after denying the hypothesis of normal distribution using the Shapiro-Wilk test, a comparison between groups was performed using the Mann-Whitney U test. In both cases, p-values less than 0.05 were considered significant. For the construction of the predictive
model, the data were segregated into two distinct sets: training set (which corresponded to 75% of the data), used during the making and improvement of the algorithm; and the independent test set (which corresponded to the remaining 25%), used to assess the model’s predictive capacity. Therefore, the instances (hematological data, sex and age) and the labels (positive or negative result of the RT-PCR test) of the training set were presented to the algorithm, so that it could later predict, from only the instances of the test set, their respective labels.

The supervised classifier XGBoost, a scalable machine learning system available as an open source package - was adopted as the predictor, given its good performance in characterizing patterns and the possibility to list the degree of importance of each variable for the predictions [28].

The evaluation of the model was based on the comparison between the real labels of each test patient (not presented to the algorithm) and the labels provided by the algorithm for such. Accuracy, specificity, sensitivity, F1-score and ROC-curve together with the value corresponding to the area under the curve (AUC) were used as evaluation parameters.

All steps of statistical analysis and development of the predictive model were performed in Python (version 3.6.9). The methodology adopted is schematized in Fig 1.

RESULTS

Among the 1157 patients analyzed, 349 presented positive results for the RT-PCR test (positive group), being therefore, for the purposes of the study, considered as cases of COVID-19, while 808 presented negative results (negative group). The positive group consisted of 164 (45.9%) males and 185 (54.1%) females, with an average age of 50.4 years (standard deviation = 16.8). The negative group had 342 (42.3%) males and 466 (57.7%) females, with an average age of 45.7 years (standard deviation = 17.8). The comparison of sex proportions using the χ² test revealed that there was no statistically significant difference (95% CI: -2.5947% to 9.8295%, p = 0.257) between the two groups. Regarding age, there was a statistically significant difference between the distributions in the two groups (p < 0.001) according to the Mann-Whitney U test.

The means and standard deviations of each hematological variable in the positive and negative groups are available in Table 1, as well as the p values from the Mann-Whitney U test for comparison between groups.

<table>
<thead>
<tr>
<th>Hematological parameter</th>
<th>Mean ± standard deviation</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>With COVID-19 (n=349)</td>
<td>Sem COVID-19 (n=808)</td>
<td></td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>13.87 ± 1.44</td>
<td>13.81 ± 1.50</td>
</tr>
<tr>
<td>Erythrocytes</td>
<td>4.76 ± 0.53</td>
<td>4.71 ± 0.52</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>41.68 ± 4.08</td>
<td>41.35 ± 4.14</td>
</tr>
<tr>
<td>MCV</td>
<td>87.89 ± 4.95</td>
<td>88.15 ± 5.16</td>
</tr>
<tr>
<td>MCH</td>
<td>29.23 ± 1.04</td>
<td>29.75 ± 9.76</td>
</tr>
<tr>
<td>MCHC</td>
<td>33.19 ± 1.95</td>
<td>33.39 ± 1.42</td>
</tr>
<tr>
<td>RDW</td>
<td>13.16 ± 1.20</td>
<td>13.08 ± 1.16</td>
</tr>
<tr>
<td>Leukocytes</td>
<td>5849.80 ± 6063.18</td>
<td>7214.44 ± 5184.46</td>
</tr>
<tr>
<td>Neutrophils</td>
<td>3190.69 ± 1772.83</td>
<td>4133.72 ± 2287.76</td>
</tr>
<tr>
<td>Lymphocytes</td>
<td>1673.09 ± 669.92</td>
<td>2172.97 ± 2422.24</td>
</tr>
<tr>
<td>Eosinophils</td>
<td>98.19 ± 168.75</td>
<td>171.57 ± 183.31</td>
</tr>
<tr>
<td>Basophils</td>
<td>21.83 ± 15.67</td>
<td>37.09 ± 21.82</td>
</tr>
<tr>
<td>Monocytes</td>
<td>566.56 ± 257.63</td>
<td>603.25 ± 329.23</td>
</tr>
<tr>
<td>Platelets</td>
<td>224982.81 ± 76408.50</td>
<td>253806.68 ± 66371.22</td>
</tr>
<tr>
<td>NLR</td>
<td>2.27 ± 1.96</td>
<td>2.43 ± 2.31</td>
</tr>
</tbody>
</table>

Data source: COVID-19 Data Sharing/BR repository. *p-values were determined using the Mann-Whitney U test. MCV = mean corpuscular volume; MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; RDW = red cell distribution width; NLR = neutrophil-lymphocyte ratio.
With the exception of hemoglobin measurement, mean corpuscular volume, red cell distribution width, mean platelet volume and neutrophil-lymphocyte ratio, there was a statistically significant difference for the values of all other hematological parameters analyzed between the two groups. The significance for the values of leukocytes, neutrophils, lymphocytes, eosinophils, basophils and platelets was more expressive, for which the p-value was less than 0.001.

The predictive model, an XGBoost classifier, was trained with data from 867 patients, among whom 259 belonged to the positive group and 608 belonged to the negative group. The values of hyperparameters that demonstrated greater balance in minimizing bias errors and variance errors, as well as in the relationship between sensitivity and specificity, were: learning rate (learning_rate) = 0.005; maximum depth of each decision tree (max_depth) = 3; number of decision trees (n_estimators) = 420; regularization value (reg_lambda) = 0.5.

In the independent test set, consisting of data from 290 patients (of which 90 belonged to the positive group and 200 belonged to the negative group), the model reached an accuracy of 80.0% with a sensitivity of 82.0% (CI: 75.6% to 88.4%), specificity of 80.0% (CI: 76.0% to 84.0%), positive predictive value of 65.38% (CI: 57.88% to 72.20%), negative predictive value of 88.17% (CI: 83.75% to 91.51%), F1-score of 70.1% and AUC value of 81.1%. The confusion matrix regarding model predictions and true results (RT-PCR test results for SARS-CoV-2 in a nasopharyngeal swab sample) is shown in Fig 2, and the ROC curve for the model is shown in Fig 3.

![Fig 2: Confusion matrix for the predictive model.](image)

![Fig 3: ROC curve and AUC-score for the predictive model.](image)

The variables that had the greatest influence on the predictive decision of the model were the basophil dosage (17.7%), the eosinophil dosage (11.0%) and the leukocyte dosage (7.5%). Those that exercised less influence were sex (which did not exert any influence), the mean platelet volume (2.1%) and the mean corpuscular volume (3.8%). The percentage importance of each instance for the prediction of the algorithm is shown in Table 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Predictive relevance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basophils</td>
<td>17.7</td>
</tr>
<tr>
<td>Eosinophils</td>
<td>11.0</td>
</tr>
<tr>
<td>Leukocytes</td>
<td>7.5</td>
</tr>
<tr>
<td>Lymphocytes</td>
<td>6.5</td>
</tr>
<tr>
<td>Erythrocytes</td>
<td>6.3</td>
</tr>
<tr>
<td>Platelets</td>
<td>6.0</td>
</tr>
<tr>
<td>MCHC</td>
<td>5.5</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>5.3</td>
</tr>
<tr>
<td>Age</td>
<td>5.3</td>
</tr>
<tr>
<td>Neutrophils</td>
<td>5.1</td>
</tr>
<tr>
<td>Monocytes</td>
<td>5.1</td>
</tr>
<tr>
<td>MCH</td>
<td>4.8</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>4.2</td>
</tr>
<tr>
<td>NLR</td>
<td>4.0</td>
</tr>
<tr>
<td>RDW</td>
<td>4.0</td>
</tr>
<tr>
<td>MCV</td>
<td>3.8</td>
</tr>
<tr>
<td>MPV</td>
<td>2.1</td>
</tr>
<tr>
<td>Sex</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**DISCUSSION**

The present study demonstrated the association of COVID-19 with the change in the values distribution of several hematological parameters, in particular the global leukocyte count and the counts of neutrophils, lymphocytes, eosinophils, basophils and platelets. For all of these parameters, lower averages were found in the group of patients with COVID-19. Such findings are consistent with those described by Ferrari et al. [29] and Dorn et al. [24].
The predictive model developed in the study was able to achieve a specificity of 82.0% and a sensitivity of 75.6% in the independent test set, demonstrating the potential of using blood count results, a widely available and accessible test, in the context of the clinical evaluation of COVID-19. In this context, it is valid to point out that the use of CBC data has already been proposed as a parameter of great utility in the diagnosis and management of viral pandemics [30]. Furthermore, the performance obtained by the model reflects the possibility of recognizing hematological patterns that point to the diagnosis of COVID-19, which reinforces the clinical significance of the change in the distribution of several hematological parameters associated with COVID-19 observed in the analytical stage of the study.

Stratification of data according to the predictive value of each of them is another major contribution of the model, given that it may not only guide a careful clinical observation of the parameters identified as most relevant (such as basophil and eosinophil measurements) in the face of suspicion of COVID-19, but also contribute to understanding the pathophysiological process involved in the disease. It has already been described in the literature that eosinophils and basophils can play a central role in the organism’s response to SARS-CoV-2 [31, 32], and the present study corroborates the relevance of investigating this role. This study presents as one of its main limitations the fact that the developed predictive model - as well as all supervised machine learning models - is totally data-driven, so that, in the face of new data with distinct characteristics patterns, performance may be compromised. In addition, the model would benefit from the availability of more data. Another limitation is the non-availability of information regarding the clinical context of symptoms and severity of patients in the repository used since COVID-19 has a wide spectrum of clinical presentations, it is inferable that it also has a wide laboratory spectrum – there are studies which describe different laboratory profiles for different classifications of disease severity –, making it difficult to recognize patterns in the absence of such information [15, 17, 31]. Reports of variable sensitivity of the RT-PCR test for SARS-CoV-2 must also be considered: since the test was the definer of the true labels in the study, the occurrence of false negatives in the training set and/or the test set could reflect inadequacies in the development and/or evaluation of the model, respectively [33-35].

CONCLUSION
This study, in line with the emerging technological efforts to contribute to the management of the international health crisis resulting from COVID-19, developed a predictive model capable of predicting with 80% accuracy the test result for the disease only from results of blood count and patient age. In this context, it is important to highlight the relevance of systematizing and disseminating data related to the disease so that they can be used in related research. It is also important that researchers from different areas of knowledge engage in such studies.

AKNOWLEDGEMENT
This work used data made available by the repository COVID-19 Data Sharing / BR, available at: https://repositoriodatasharingfapesp.uspdigital.usp.br/. I thank and congratulate the São Paulo Research Foundation and the collaborating/ participant institutions - University of São Paulo, Fleury Institute, Sírio-Libanês Hospital and Israëlia Albert Einstein Hospital – for the initiative, which makes a notable contribution to the development of research on this theme.

AUTHOR’S CONTRIBUTION
The author agree on this final form of the manuscript.

CONFLICTS OF INTEREST
The author declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE
No financial interests related to the material of this manuscript have been declared.

REFERENCES


