Discrimination of SARS-Cov 2 and arboviruses (DENV, ZIKV and CHIKV) clinical features using machine learning techniques: a fast and inexpensive clinical screening for countries simultaneously affected by both diseases

João Daniel S. Castro\textsuperscript{a, b}\textsuperscript{*}
joaodasc@gmail.com

\textsuperscript{a} Universidade do Estado da Bahia, Departamento de Ciências Exatas e da Terra, Programa de Pós-Graduação em Química Aplicada, Salvador, BA, Brazil,

\textsuperscript{b} Secretária da Educação do Estado da Bahia, Centro Administrativo da Bahia - CAB, Salvador, BA, Brazil

Abstract

SARS-Cov-2 (Covid-19) has spread rapidly throughout the world, and especially in tropical countries already affected by outbreaks of arboviruses, such as Dengue, Zika and Chikungunya, and may lead these locations to a collapse of health systems. Thus, the present work aims to develop a methodology using a machine learning algorithm (Support Vector Machine) for the prediction and discrimination of patients affected by Covid-19 and arboviruses (DENV, ZIKV and CHIKV). Clinical data from 204 patients with both Covid-19 and arboviruses obtained from 23 scientific articles and 1 dataset were used. The developed model was able to predict 93.1\% of Covid-19 cases and 82.1\% of arbovirus cases, with an accuracy of 89.1\% and Area under Roc Curve of 95.6\%, proving to be effective in prediction and possible screening of these patients, especially those affected by Covid-19, allowing early isolation.

Keywords: Coronavirus; Covid-19; Arbovirus; Machine Learning; Clinical Features; Dengue; Zika; Chikungunya

Introduction

1 In December 2019, a series of pneumonia cases of unknown cause emerged among
2 visitors to a wet market in the city of Wuhan, Hubei (China). Genetic sequencing of
3 samples from the respiratory tract of these patients identified a new type of Coronavirus,
called (2019-nCoV) [1], which since the outbreak, has affected more than 1 million people and totals 60,000 deaths worldwide [2,3].

Covid-19 produces a systemic respiratory disease that can progress to severe pneumonia in approximately 16% of cases [4]. Sepsis is the most frequent complication, followed by respiratory failure, acute respiratory distress (ARDS), heart failure and septic shock [5].

Concomitantly with the Covid-19 outbreak, tropical countries have faced successive outbreaks of arboviruses, such as Dengue (DENV), Zika (ZIKV) and Chikungunya (CHIKV), presenting themselves as a challenge to the health systems of these countries. Arboviruses are diseases caused by viruses transmitted to humans through the bites of infected blood-sucking arthropods, predominantly Aedes mosquitoes [6], and their clinical manifestations can vary from mild and undifferentiated febrile disease to neurological, joint and hemorrhagic febrile syndromes [7]. In addition, it is observed that there is a low probability of predicting a possible outbreak [8], which may surprise health systems that, today, are saturated due to the Covid-19 pandemic.

The imminence of a syndemic state between Covid-19 and arboviruses (mainly Dengue) has been shown to be a concern of researchers from Latin American countries [9,10]. Even though it has not yet reported cases of co-infection, your treatment may be adversely affected, since the spectrum of drugs does not seem compatible [11]. Associated with this, researchers have reported a false positive diagnosis for Dengue in situations in which patients are affected by the SARS-CoV-2 virus [12,13], even though there was no genomic similarity between the viruses [14], which can make the diagnosis of these diseases even more difficult.

Brazil has historically faced cycles of arbovirus epidemics throughout its territory, with Dengue being the most prevalent arbovirus [15]. According to the most recent Brazilian epidemiological bulletin, until now (ie, April 06, 2020) approximately 500 thousand cases of arboviruses have been reported (DENV, ZIKV, CHIKV), 96% of which are associated with dengue, 2.73% Chikungunya and 0.33% to Zika [16], compared to 9506 cases of Covid-19 (until April 3) [17].

Thus, in order to optimize resources and care for patients, as well as to enable effective screening, allowing isolation and early management, biomarkers (obtained through blood counts at admission) can be effective in predicting the degree of severity of disease. Studies indicate that Covid-19 causes Lymphocytopenia [4] and Thrombocytopenia [18],
while arboviruses can cause Leukopenia, Lymphohistiocytosis and mild Lymphocytopenia [19–22].

Therefore, based on the clinical data of patients affected by Covid-19 or Arboviruses (DENV, ZIKV and CHIKV), this study aims to build a predictive model using machine learning algorithms, in order to screen and diagnose patients in case a possible syndemic state. It is also intended to share clinical data by the medical community about diseases of global circulation, in order to mitigate their impacts.

2 Materials and Methods

An electronic search was performed in the databases of Periódicos CAPES (portal of the Brazilian development agency CAPES, which aggregates numerous publishers of scientific journals), Google Scholar, Google Dataset and Science Direct, using the keywords, “Clinical features” OR “White blood cells count” OR “Haemogram” AND “coronavirus 2019” OR “COVID-19” OR “2019-nCoV” OR “SARS-CoV-2”. To search for clinical data related to arboviruses, the terms “Clinical features” OR “White blood cells count” OR “Haemogram” AND “Dengue” OR “Zika” OR “Chikungunya” were used. A Dataset containing clinical data from Brazilian patients Positive for SARS-CoV-2 was also added to the data set, made available by Hospital Israelita Albert Einstein [23].

The papers were chosen according to the criteria of: analysis of the abstract, title and body of the text, excluding works that presented only the average of the data obtained for a group of patients or that did not present the patients' blood count. Thus, 23 scientific papers were selected, all of which were written in English.

This study was carried out in accordance with the Declaration of Helsinki and in accordance with the terms of local legislation.

The predictive model was built using the open-source software Orange (v. 3.20.1) [24]. Before the implementation of the algorithm, the data were normalized (center by mean) and then submitted to the detection of outliers (abnormal data), using covariance assessment.

The input data for the training of the algorithm were the White Blood Cells count (WBC) and the Lymphocyte count, available in the articles, and the classes used for the prediction
were Covid-19 and Arboviruses. The type of algorithm used was SVM with Cost regression \( c = 1.50 \), Kernel type RBF \( g = 1.04 \) and the following optimization parameters: numerical tolerance = 0.005, without iteration limit. The model was trained using the data and validated using the cross-validation methodology, with \( k \)-fold =20. The validation method consists of dividing the data set into \( k \) parts, using \( k-1 \) parts for modeling (training) and the remaining part for testing. In this model the data set was tested twenty times, each one with a different fold from dataset. The idea behind this type of algorithm is to map the data, transposing it into a high dimensional space in which hyperplanes separate the data into subgroups (clusters) according to their characteristics, allowing the classification problem to be more easily solved [25]. The model developed is available in supplementary materials.

3. Results and discussion

3.1 Characteristics of the included studies

The detailed characteristics of this study are found in the Dataset available as supplementary material. Data from 204 patients were used in total, and after detecting outliers, 120 patients remained positive for SARS-CoV-2 and 67 patients with arboviruses (DENV, ZIKV and CHIKV), in a wide age group and of both sexes, totaling 187 patients.

The parameters used for the evaluation and screening was the existence / prevalence of Leukopenia and Lymphohistiocytosis / Lymphocytopenia, based on what was previously reported in the literature. Tables 1 and 2 summarize the data found in the literature. It is worth noting that the dataset provided by Hospital Israelita Albert Einstein [23] was already normalized, and therefore it is not shown in the data set in the tables below.

Table 1. Clinical features of patients affected by Covid-19

<table>
<thead>
<tr>
<th>Número de Pacientes</th>
<th>Disease</th>
<th>WBC ((10^9/L))</th>
<th>Lymphocites ((10^9/L))</th>
<th>Observation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Número de Pacientes</td>
<td>Disease</td>
<td>WBC (10⁹/L)</td>
<td>Lymphocites (10⁹/L)</td>
<td>Observation</td>
<td>Reference</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>1</td>
<td>Zika</td>
<td>2.28</td>
<td>0.83</td>
<td>Liver injury</td>
<td>[34]</td>
</tr>
<tr>
<td>1</td>
<td>Zika</td>
<td>6.10</td>
<td>1.0</td>
<td>-</td>
<td>[35]</td>
</tr>
<tr>
<td>1</td>
<td>Zika</td>
<td>4.90</td>
<td>1.42</td>
<td>-</td>
<td>[36]</td>
</tr>
<tr>
<td>12</td>
<td>Zika</td>
<td>5.96 (3.78 - 10.13)</td>
<td>2.64 (0.39 - 7.09)</td>
<td>-</td>
<td>[37]</td>
</tr>
<tr>
<td>1</td>
<td>Zika</td>
<td>3.65</td>
<td>0.47</td>
<td>-</td>
<td>[38]</td>
</tr>
<tr>
<td>1</td>
<td>Zika</td>
<td>3.10</td>
<td>1.55</td>
<td>-</td>
<td>[39]</td>
</tr>
<tr>
<td>1</td>
<td>Chikungunya</td>
<td>3.10</td>
<td>0.7</td>
<td>-</td>
<td>[40]</td>
</tr>
<tr>
<td>14</td>
<td>Chikungunya</td>
<td>7.32 (3.00 - 13.00)</td>
<td>3.44 (1.30 - 7.54)</td>
<td>Children</td>
<td>[41]</td>
</tr>
<tr>
<td>1</td>
<td>Chikungunya</td>
<td>15.7</td>
<td>6.09</td>
<td>-</td>
<td>[42]</td>
</tr>
<tr>
<td>1</td>
<td>Chikungunya</td>
<td>19.1</td>
<td>1.01</td>
<td>Erysipelas</td>
<td>[43]</td>
</tr>
<tr>
<td>3</td>
<td>Chikungunya</td>
<td>28.9 (21.8 - 35.1)</td>
<td>1.2 (0.7 - 1.2)</td>
<td>Fatal cases</td>
<td>[44]</td>
</tr>
<tr>
<td>1</td>
<td>Chikungunya</td>
<td>3.2</td>
<td>1.09</td>
<td>-</td>
<td>[45]</td>
</tr>
<tr>
<td>5</td>
<td>Zika/Dengue</td>
<td>3.65 (1.24 - 6.3)</td>
<td>0.96 (0.39 - 1.84)</td>
<td>Coinfection</td>
<td>[46]</td>
</tr>
<tr>
<td>1</td>
<td>Dengue</td>
<td>1.60</td>
<td>0.93</td>
<td>-</td>
<td>[47]</td>
</tr>
<tr>
<td>40</td>
<td>Dengue</td>
<td>7.16 (4.43 - 11.9)</td>
<td>4.87 (1.90 - 6.70)</td>
<td>-</td>
<td>[48]</td>
</tr>
<tr>
<td>Total = 84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Clinical features of patients affected by arboviruses (DENV, ZIKV and CHIKV)

3.2 Model assessment

![Confusion Matrix for classification test using an SVM machine learning algorithm](https://example.com/confusion_matrix.png)

Figure 1. Confusion Matrix for classification test using an SVM machine learning algorithm
The results of a classification can be represented as a matrix called a confusion matrix, a square matrix (G x G) whose rows and columns represent the experimental and predicted data respectively [49].

The matrix shows that the model was able to predict more than 90% of Covid-19 cases and 82.1% of cases of arboviruses, based on clinical data obtained in the literature. The developed model also has an Area under Roc Curve (AUC) of 0.956, an accuracy of 0.891 and an precision of 0.893.

Lymphocytopenia is a clinical condition observed in patients with SARS-CoV-2, regardless of gender, age and / or pre-existence of comorbidities [50–52]. This condition allowed the correct classification of most of the observed cases. The classification error of individuals affected by arboviruses is also reported by Yan and collaborators, when describing two clinical pictures of SARS-Cov-2 as false positives for Dengue (DENV) [12], showing the urgency in the development of more accurate methodologies.

Leukopenia presents to patients with arboviruses (DENV, ZIKV and CHIKV) in medium to intense degrees, according to the PAHO (Pan-American Health Organization) classification [53], and also presented in clinical studies [54,55]. Associated with this, some patients may experience lymphohistiocytosis [56–58]. These characteristics allowed the correct classification of 82.1% of the cases.

Cases misclassified totaling 10.93% of cases and are presented in Table 3:

<table>
<thead>
<tr>
<th>WBC (10^9/L)</th>
<th>Lymphocyte (10^9/L)</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.6</td>
<td>2.2</td>
<td>Covid -19</td>
<td>Arboviruses</td>
</tr>
<tr>
<td>8.24</td>
<td>2.46</td>
<td>Covid -19</td>
<td>Arboviruses</td>
</tr>
<tr>
<td>6.48</td>
<td>1.98</td>
<td>Covid -19</td>
<td>Arboviruses</td>
</tr>
<tr>
<td>4.2</td>
<td>0.7</td>
<td>Covid -19</td>
<td>Arboviruses</td>
</tr>
<tr>
<td>4.23</td>
<td>1.28</td>
<td>Covid -19</td>
<td>Arboviruses</td>
</tr>
<tr>
<td>6.5</td>
<td>2.8</td>
<td>Covid -19</td>
<td>Arboviruses</td>
</tr>
<tr>
<td>11.4</td>
<td>2.7</td>
<td>Covid -19</td>
<td>Arboviruses</td>
</tr>
</tbody>
</table>
From the analysis of table 3 and figure 2, we can find the factors that allow the erroneous classification of patients: half of the patients affected by Covid-19 who have WBC values (≅ 5) associated with low values of lymphocytes (> 2) are classified as having arboviruses; while the other half is classified according to the high value of lymphocytes.

Patients affected by arboviruses who have lymphocytopenia (lymphocytes ≤ 2) and without leukopenia are classified as having Covid-19.
Figure 2. WBC x Lymphocyte count plot for patients classified incorrectly. The caption of each patient indicates the disease to which it is affected, the symbols indicate the predicted illness.

4 Conclusion

The model developed was effective in discriminating based on clinical data between patients affected by Covid-19 and patients affected by arboviruses (DENV, ZIKV, CHIKV), being able to predict 93.1% of the cases of SARS-CoV-2 and 82.1% of cases of arboviruses. Even though it presented a not very satisfactory result for arboviruses, the model allows the screening of patients with Coronavirus (contagious), thus speeding up the isolation and hospitalization procedures.

Because it was developed using free software, this model can be used in public health systems in countries in the event of a simultaneous outbreak of arboviruses and Covid-19, minimizing the impacts of a possible collapse of the health system.
The author thanks the collaboration of doctors Marjorie Lobo and Tainá Cajazeira and the researcher Luane Barbosa for the review of this work.

References


[41] G.M. Oliveira, T.M.A. Marques, J.R. Dos Santos, E.D.F. Daher, R.D. Leite, E.S.


D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification


[53] Pan American Health Organization (PAHO), Tool for the diagnosis and care of patients with suspected arboviral diseases, 2017.

doi:10.1371/currents.outbreaks.0bf6aeb4d30824de63c4d5d745b217f5.


