Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Feature selection reveal peripheral blood parameter's changes between COVID-19 infections patients from Brazil and Ecuador

Bruno César Feltes, Igor Araújo Vieira, Jorge Parraga-Alava, Jaime Meza, Edy Portman, Luis Teran, Márcio Dorn

PII: S1567-1348(22)00025-9
DOI: https://doi.org/10.1016/j.meegid.2022.105228
Reference: MEEGID 105228
To appear in: Infection, Genetics and Evolution
Received date: 24 December 2021
Accepted date: 23 January 2022

Please cite this article as: B.C. Feltes, I.A. Vieira, J. Parraga-Alava, et al., Feature selection reveal peripheral blood parameter's changes between COVID-19 infections patients from Brazil and Ecuador, Infection, Genetics and Evolution (2021), https://doi.org/10.1016/j.meegid.2022.105228

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier B.V.
Feature selection reveal peripheral blood parameter’s changes between COVID-19 infections patients from Brazil and Ecuador

Bruno César Feltes
Department of Genetics, Institute of Bioscience
Department of Biophysics, Institute of Bioscience
Federal University of Rio Grande do Sul, Porto Alegre, RS - Brazil

Igor Araújo Vieira
Genomic Medicine Laboratory, Experimental Research Center
Hospital de Clínicas de Porto Alegre, Porto Alegre, RS - Brazil

Jorge Parraga-Alava
Facultad de Ciencias Informáticas
Universidad Técnica de Manabí, Portoviejo, Manabí - Ecuador

Jaime Meza
Facultad de Ciencias Informáticas
Universidad Técnica de Manabí, Portoviejo, Manabí - Ecuador

Edy Portman
Department of Informatics
University of Fribourg, Fribourg - Switzerland
Luis Teran
Department of Informatics
University of Fribourg, Fribourg - Switzerland

Márcio Dorn
Institute of Informatics
Center of Biotechnology
Federal University of Rio Grande do Sul, RS, Brazil
National Institute of Science and Technology - Forensic Science, Porto Alegre, RS - Brazil

December 24, 2021

Abstract

The investigation of conventional complete blood-count (CBC) data for classifying the SARS-CoV-2 infection status became a topic of interest, particularly as a complementary laboratory tool in developing and third-world countries that financially struggled to test their population. Although hematological parameters in COVID-19-affected individuals from Asian and USA populations are available, there are no descriptions of comparative analyses of CBC findings between COVID-19 positive and negative cases from Latin American countries. In this sense, machine learning techniques have been employed to examine CBC data and aid in screening patients suspected of SARS-CoV-2 infection. In this work, we used machine learning to compare CBC data between two highly genetically distinguished Latin American countries: Brazil and Ecuador. We notice a clear distribution pattern of positive and negative cases between the two countries. Interestingly, almost all red blood cell count parameters were divergent. For males,
neutrophils and lymphocytes are distinct between Brazil and Ecuador, while eosinophils are distinguished for females. Finally, neutrophils, lymphocytes, and monocytes displayed a particular distribution for both genders. Therefore, our findings demonstrate that the same set of CBC features relevant to one population is unlikely to apply to another. This is the first study to compare CBC data from two genetically distinct Latin American countries.

1 INTRODUCTION

Since the end of 2019, the Coronavirus disease (COVID-19) has impacted nearly all branches of society, especially the global economy and health services [39, 41]. This reality was particularly noticeable in developing and third-world countries, which struggled with financial resources and administration problems, leading to inefficient population testing and poorly devised strategies to manage the crisis [24].

As expected and heavily warned, SARS-CoV-2 mutated, and now a myriad of variants, including rapidly expanding virus lineages since December 2020, designated variants of concern (VOCs) were identified in distinct populations [51]. In this sense, some VOCs became more threatening than others due to their higher infectability, such as Alpha (initially identified in the UK), Beta (South Africa), Gamma (Brazil), and the Delta (India) variants [51, 48, 40, 15]. In addition to the high genetic variability of the virus, the extensively diverse genetic background of the human host also contributes to the complexity of COVID-19 clinical outcomes. Thus, comparing clinical and laboratory data between affected and non-affected individuals from extensive admixture-Latin American populations is essential to better understand the effect of the interaction between host genetic ancestry and exposure to specific SARS-CoV-2 VOCs on the susceptibility to infection and disease severity [23, 17]. In particular, Brazilian and Ecuadorian
populations have heterogeneous genetic constitutions with a predominant three-hybrid composition (European, African, and Native American). Nonetheless, the contribution of each ancestry component is markedly different between these countries [45, 49, 59, 47].

Unfortunately, the current scenario is that numerous countries are still struggling with high SARS-CoV-2-associated infection and mortality rates, mainly because of the new VOCs that can infect fully vaccinated individuals [20]. Thus, more efforts should be made to properly recognize the differences between the populations to understand the distinct hematological and/or immunological reactions to SARS-CoV-2 infection. The hematological manifestations of COVID-19, mainly due to the exacerbated release of inflammatory mediators (“cytokine storm”), include blood cell count alterations, particularly lymphopenia, neutrophilia, and thrombocytopenia exhibiting potential for early diagnosis of the disease and prognostic significance [2, 32, 52, 60]. Interestingly, despite several studies describing the profile of hematological parameters in COVID-19-affected cohorts predominantly from Asian and USA populations [13, 29, 50], no works compared these laboratory data from COVID-19 patients between countries in Latin America. Given that population-specific genetic ancestry influences on hematopoiesis and in turn on Complete Blood Count (CBC)-associated traits [55, 10], as well as the high genetic variability observed in Latin America [11], it is imperative to learn the differences in hematological profiles to SARS-CoV-2 infection between individuals from distinct Latin countries. Being this a challenging task, employing advanced computational techniques that could automatically detect patterns buried within the large amount of available data obtained from clinical data is gradually gaining more ground in the Medical field.

In this sense, Machine Learning (ML) techniques have been employed to investigate CBC data and aid in the screening of patients suspected of SARS-CoV-2 infection [57, 58, 19, 3, 7, 9,
Briefly ML algorithms perform computational tasks by relying on learning patterns from data samples to automate inferences. In classification tasks, the aim is to induce a model that can distinguish pathological from non-pathological (healthy) samples from the population. Some variables used in the induction of the classification model are more relevant than others and may reveal important characteristics about the set of samples analyzed. These variables can change from one cohort to another, revealing relevant characteristics and comparing populations.

In this work, we used ML to compare CBC data between two highly genetically distinct Latin American countries: Brazil and Ecuador. We found striking differences in CBC profiles from the two countries, especially in red-blood-cell counts. Moreover, white-blood-cell counts displayed distinguished patterns, specific for males and females, positive and negative cases. Therefore, our data showed that features relevant to one population are unlikely to apply to another, which agrees with the genetic ancestry background of both countries.

2 MATERIAL AND METHODS

Datasets

Datasets were obtained from an open repository, the COVID-19 Data Sharing/BR initiative [14, 35], of COVID-19-related cases in Brazil, which is comprised by nearly 177,000 clinical cases. Datasets originate from the Albert Einstein Hospital[1]. The data from these datasets were collected from February 26th, 2020, to June 30th, 2020, and the control data (individuals without COVID-19) was collected from November 1st, 2019, to June 30th, 2020. Patient data is anonymized, demographic data (including sex, year of birth, and residence zip code) is also

[1] https://www.einstein.br
described. Other clinical results are also available, such as pulmonary function tests, urinalysis parameters, and pulmonary imaging results. COVID-19 detection by RT-PCR tests is described for all patients, and serology diagnosis (IgG and IgM antibody detection) is provided for some samples. Information is not complete for all patients, with a distinct mixture of results offered individually.

Twenty distinct CBC test parameters were obtained from the dataset, including hematocrit (%), hemoglobin (g/dl), platelets (×10^3 μl), mean platelet volume (MPV) (fl), red blood cells (RBCs) (×10^6 μl), white blood cells or leukocytes (×10^3 μl), lymphocytes (×10^3 μl), basophils (×10^3 μl), eosinophils (×10^3 μl), monocytes (×10^3 μl), neutrophils (×10^3 μl), mean corpuscular volume (MCV) (fl), mean corpuscular hemoglobin (MCH) (pg), mean corpuscular hemoglobin concentration (MCHC) (g/dl), red blood cell distribution width (RDW) (%), % Basophils, % Eosinophils, % Lymphocytes % Monocytes, and % Neutrophils. Since elevated neutrophil/lymphocyte ratio (NLR) has been described as a hematologic biomarker in COVID-19 positive patients [4], it was calculated from the result of absolute counts of these respective types of leukocytes from each patient of the datasets. Overall baseline characteristics can be found on the complete dataset, available at the FAPESP COVID-19 Data Sharing/BR. Patients with incomplete (missing data) or lacking the above parameters were not included. A unique CBC test was used for patients with multiple test results, with the selection based on the blood test date. In this sense, same-day, or the day closest to the test, results to the PCR-test collection date were utilized as a reference.

For Ecuador, the CBC data was obtained between March 3rd, 2021, to August 9th, 2021.

2 https://repositoriodatasharingfapesp.uspdigital.usp.br
Exams were carried out by Segurilab\(^3\) and Previne Salub\(^4\) laboratories. The data set is comprised by nearly 400 clinical cases and each patient were agree to participate in the study by sharing your PCR test results along with a CBC. The features in the CBC data from Ecuador are the same as the Brazilian one. We used three datasets in our experiments (Table 1): Br-v1 contains data from 3,108 patients from Brazil; Ecu-v1 contains data from 251 patients from Ecuador, and Br-v2 contains a reduced version of the Br-v1 dataset considering the same number of samples as in Ecu-v1. By constructing the Br-v2 dataset, we ensured that the age and sex of the patients were compatible with the data from Ecu-v1 (Table 2), thus avoiding any bias in the subsequent analyses. All datasets are available at https://github.com/sbcblab/CBCBrazilEcuador.

The number of negative samples for all datasets exceeds the positive samples. This is expected from disease data since the number of infections will be small compared to the entire population. Class imbalance is common in many real-world applications and affects the quality and reliability of ML approaches [31, 26, 33]. We use Shannon’s entropy to measure the imbalance of each dataset (Table 5). On a dataset of \(n\) instances, \(k\) classes of size \(c_i\) we compute the entropy using Equation 1, and then the balance is measured of each dataset using Equation 2 getting 0 for an unbalanced dataset and 1 for a balanced dataset.

Finally, the datasets do not provide viral sequencing data to investigate which SARS-CoV-2 variant was identified in positive cases (i.e., specific VOC/viral lineage of each patient, or dominant transmission lineage in a subset of samples, were not available).

\[
H = -\sum_{i=1}^{k} \frac{c_i}{n} \log \frac{c_i}{n}
\]

\(^3\) https://segurilab.ec
\(^4\) https://previenesalud.com
Feature Selection and Feature Extraction

Feature selection (FS) is part of the tasks involving dimensionality reduction, one of the fundamental data pre-processing steps for building ML models [5]. ML models assume that all features are relevant for the task at hand. However, the computational cost of inducing the model increases with the number of features. Irrelevant attributes can also impair the predictive ability of the model, as there is evidence that the same learning algorithm achieves better performance by inducing on a subset of the same data where some of the original features are discarded [53, 12, 11]. In most problems, the relevant attributes are unknown, and FS can discover knowledge from the collected data. In different tasks, the collection, maintenance, and provision of input data present an economic cost to be minimized and more stringent performance requirements. In CBC data, FS is performed as a selection of the most representative variables in the dataset, taking into account the values of each feature in the different samples. With this, it can assist in biological pattern discovery, class prediction, and recognition of data that are not directly associated or known to the problem but may play a key role [6, 37].

FS methods are grouped into four types [5, 30]: (i) Filter, (ii) Wrapper, (iii) Embedded and (iv) Hybrid. While the evaluation of the filter-based methods is independent of any classifier, the wrapper and embedded methods use the accuracy of the classifier itself. However, they also require strategies to search the feature space to perform the selection [30]. In Wrapper type methods, the selection is done using some optimization algorithm (particularly Metaheuristics) and then wrapping a classifier around the selected features, using their accuracy as an evaluation
criterion. The set of the most discriminative features is found by minimizing the classification error, which often generates better results than filters. In this paper, we use a wrapper FS algorithm based on Recursive Feature Elimination (RFE) and Support Vector Machines (SVM-RFE) [22], in which SVM is the classifier employed to assess the quality of inputs subsets.

The main group of algorithms for dimensionality reduction is Feature Extraction (FE), a set of methods that transforms the original feature space into different spaces with a new set of axis by combining their features and finding the ones that preserve the original information [54]. In our experiments, we used two features extractors to compare the population of Brazil and Ecuador: Principal Component Analysis (PCA) [27] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [34]. FS techniques are used to filter irrelevant or redundant features from the datasets. FE methods are used for dimensionality reduction, in which an initial set of raw data is reduced into more manageable groups. The main difference between these approaches is that FS keeps a subset of the original features, while FE creates new ones.

**Statistical analysis**

The Mann-Whitney U (MWU) statistical test (3) was used to obtain the statistical significance between COVID-19 positive and negative groups for each component of the CBC tests (Table 3). The MWU test is a non-parametric statistical technique employed to analyze differences between the medians of two distinct datasets [36]. A p-value < 0.05 was considered statistically significant.

\[
U_i = n_1n_2 + \frac{n_1(n_1 + 1)}{2} - R_i
\]

(3)

where \( n_1 \) and \( n_2 \) are the number of the score in the 1st and 2nd condition, respectively.
\( R_i \) is the sum of the ranks in the 1st condition.

3 RESULTS

CBC differences between Brazilian and Ecuadorian populations

Intra- and inter-populational variations were spotted in our PCA and t-SNE analyzes, with a clear division observed in the PCA between positive and negative cases of both genders between Brazilian and Ecuadorian populations.

Positive and negative cases in male and female individuals, when looked at individually, were distinctively distributed in the Ecuadorian population. In contrast, the Brazilian cases were more mixed (Fig.1A-B). However, when comparing male and female positive and negative cases together from both populations, an explicit division was observed (Fig.1C). The same tendency is noticed in the t-SNE (Fig.2C). The results indicate that samples from male and female individuals might show a mixed distribution when observed individually in the context of their own countries. However, when analyzed together, Brazilian and Ecuadorian positive and negative cases have a distinct distribution, suggesting particularities in CBC patterns.

Fig.3 and Fig.4 display the CBC data of positive and negative cases for female and male individuals, respectively, between the two Southern American populations. The figures are also divided into white blood cell-derived parameters (WBCP) and red blood cell-derived parameters (RBCP), along with two platelet-derived variables (MPV and platelet count).

Regarding inter-populational differences for female negative cases, a remarkable distinction was observed between basophils, eosinophils, and all RBCP, except for MCH and RDW. The same pattern was evident for female positive cases; besides that, only MCH exhibited
similar distribution **Fig.3**. In contrast, inter-populational distinctions between positive male patients, basophils, neutrophils, lymphocytes, and all RBCP, except MCH, can be noted. The same pattern was sustained in negative cases **Fig.4**. Hence, CBC patterns are characteristic of each country, especially the red blood cell ones. Likewise, basophils appear as significant features in male and female individuals regardless of SARS-CoV-2 infection status.

When mixing male and female positive and negative cases from Brazilian (Brazil-v2) and Ecuadorian datasets **Fig.5**, we also discerned a similar pattern, where RBCP remain as key separation feature, including an intriguing double profile of RDW found only in Brazilian negative cases. However, neutrophils, lymphocytes, and monocytes display particular distributions in negative and positive patients, especially in the Brazilian population. In addition, no recognizable intra- or inter-population patterns of platelet-derived variables was observed between positive and negative cases of both genders.

**Distinct CBC parameters classify SARS-CoV-2 infection status in Brazil and Ecuador**

Lastly, a FS approach (RFE-SVM) was used to identify the most relevant parameters (features) that could better classify the CBC profiles between positive and negative cases for COVID-19 of the different genders and population groups analyzed (**Fig.6**). Briefly, a smaller number of CBC features were found to explain the CBC differences that better classified positive and negative cases for COVID-19 in the Ecuadorian population (n features= 2-3), except for the analysis involving only female individuals from the same population (n features= 12), in comparison with the Brazilian dataset (BR-V2, n= 11-16 features depending on gender) (**Fig.6A and B**). Concerning the analysis of the Ecuadorian population involving male individuals and both genders
together, the selected features were the lymphocyte and neutrophil counts and two specific RBCP. In general, focusing on the Brazilian dataset, monocyte and lymphocyte counts exhibited a more central representative role. The results from ML/FS approaches were supported by the statistical analysis (Supp. Material Table 2), which also showed that despite the statistical significance of several CBC parameters in both males and females from both populations, they still present heterogeneous patterns (Supp. Material Table 2). For instance, some CBC features were significant only for a specific gender in one of the populations, such as monocyte counts, which were exclusively relevant in females from the BR-v2 (Brazil) dataset. It must be noted that the MWU statistical test (table 3) provides a uni-dimensional selection of the most relevant CBC features for discriminating COVID-19 positive and negative cases from both populations. At the same time, the FS approach is a multidimensional analysis considering the global set of features included in the study.

Moreover, when considering both populations together in FS analyses (Fig.6C), lymphocyte and neutrophil counts remained the most relevant features when comparing male and female individuals separately. These findings suggest that the number and types of CBC parameters conferring decisive differences to distinguish SARS-CoV-2 infection status tend to be population-specific. Hence, the same set of features (CBC variables) relevant to one population is unlikely to apply to another.

It is also interesting to highlight that three features were deemed enough to classify individuals with COVID-19 in the Ecuadorian population (Fig.6B), excluding analysis restricted to females. In contrast, in the Brazilian one, at least eleven features were needed (Fig.6A). This explains why, when combining the datasets, the classifier required numerous features to classify the inter-populational CBC data correctly (Fig.6C).
4 DISCUSSION

Recently, given the scenario of limited resources for large-scale testing of individuals for COVID-19 using molecular biology techniques (mainly RT-PCR), a wide range of studies have explored the importance of identifying blood cell alterations from conventional CBC data as a complementary laboratory tool for the classification of the SARS-CoV-2 infection status in different populations [57, 58, 19, 3, 7, 9, 25, 56]. Nonetheless, there are no descriptions in the literature of comparative analyses of CBC findings between COVID-19 positive and negative cases from certain Latin American countries, such as Ecuador, as well as a scarcity of comparisons of CBC profiles from these countries concerning other populations with different genetic background on the same continent.

In the present study, ML/FS approaches were applied to identify intra- and inter-populational CBC alterations in datasets from Brazil and Ecuador, two Latin American countries with marked differences in their genetic ancestry composition. In this sense, we notice a clear distribution pattern in the PCA and t-SNE analysis of positive and negative cases between the two countries. Such cases were distinctly grouped, with only a mild overlap. Almost all RBCP were strikingly divergent in the two populations; however, for WBCP, the results were mixed, with few patterns emerging. Nevertheless, basophils surfaced an important differential feature for both males and females in both populations. For males, neutrophil and lymphocyte counts were distinct between Brazil and Ecuador, while eosinophils profiles were distinguished for females. Finally, when combining male and female samples, neutrophils, lymphocytes, and monocytes displayed a particular distribution. Previous findings support our results. For instance, WBCP alterations, specifically of neutrophils and lymphocytes counts, have been reported in numerous
studies from different populations [2, 52, 60]. Remarkably, a recent work using ML and Artificial Intelligence approaches to predict SARS-CoV-2 positive patients from the full CBC dataset originated from the Albert Einstein Hospital, the same dataset from the Brazilian population queried here, also identified WBCP alterations in basophils, lymphocytes, eosinophils, and monocytes, as well as in the count of RBCs [8]. These similar results for the WBCP group validate the computational strategy applied in the present study and support our comparison of CBC patterns between the Brazilian population and the Ecuadorian one. Nonetheless, when comparing affected and unaffected Brazilian individuals in our analyses, almost all RBCP had divergent patterns, not exclusively the RBCs count reported in the previous study. Lastly, unlike the study conducted by Banerjee and colleagues (2020) [8], we did not identify a clear, recognizable pattern of decrease in platelets in Brazilian individuals with COVID-19.

Although Brazilian and Ecuadorian populations have a similar three-hybrid genetic composition (European, African, and Native American), the fraction of each ancestry component significantly differs between the two countries. Briefly, the European and the "Mestizo" ancestry (an admixture of Spanish and Indigenous American ancestry) are the most predominant contributions in Brazil and Ecuador, respectively [45, 49, 59, 47]. These distinct genetic ancestries’ profiles may explain, at least in part, some differences found in the present study comparing CBC patterns between affected and unaffected individuals. In this context, our results from the FS approach (RFE-SVM), employed to identify the most relevant CBC variables that could better distinguish the SARS-CoV-2 infection status, indicated that only three CBC features successfully classify Ecuadorian individuals with COVID-19. In comparison, more than ten features were required for the same classification purpose in the Brazilian population. Of note, all these CBC features’ differences derived from RFE-SVM analyses were statistically significant.
using the MWU test (table 3), reinforcing the findings obtained through FS approach. We hypothesize that the more significant number of CBC features needed to distinguish positive from negative cases for COVID-19 in Brazil may be explained by the fact that this country has one of the most heterogeneous genetic constitutions in the world with a highly extensive admixture [46, 49, 42]. Two additional factors should also be pointed out. Firstly is the exposure to different predominant VOCs or viral lineages of SARS-CoV-2 in the pandemic epidemiological week period in which the CBC data was collected in each country. For example, the dominant transmission lineage in Ecuador throughout 2020 was B.1.1.74 (previously designated as an Alpha VOC, one of the lineages descended from B.1.1, which became one of the most dominant lineages during the early phase of the pandemic in Europe and North America) [21]. This lineage differs from those more commonly identified as circulating in most regions of Brazil in the same year (B.1.1.28 and B.1.1.33, as well as P.1 and P.2 lineages, previously designated as Alpha and Gamma VOCs, respectively) [43, 15, 40, 16]. Secondly is the influence of altitude differences (Ecuadorians are located at a higher altitude than Brazilians) on each population’s CBC constitutive profiles (baseline or healthy condition). It has been widely described that the rate of erythropoiesis is altitude-dependent [44, 18], as well as the occurrence of human genetic adaptations to higher altitudes for maintaining an adequate efficiency of oxygen delivery [38, 28], which ultimately can explain different RBCP patterns between the two populations studied.

The results of the current study must be interpreted in the context of the following limitations: (i) small sample number in the Ecuador dataset (Ecu-v1, n= 375); (ii) lack of data regarding the results of other clinical and laboratory tests in the Ecuadorian sampling, in addition to the CBC-associated hematological parameters analyzed here, such as biochemical and immunological tests, pulmonary function and imaging tests; and (iii) temporal difference in the
period of the pandemic in which the datasets from the two countries were collected, being those of
the Brazilian population obtained at an earlier time of the pandemic (November 1st, 2019 to June
30th, 2020) and those of Ecuador most recently collected (March 3rd, 2021 to August 9th, 2021).
On the other hand, a positive aspect of our study is the importance of the CBC profiles’
comparison between COVID-19 positive and negative individuals from Latin American countries,
providing additional evidence that the same hematological parameters considered in a specific
population context cannot be used in the complementary differential diagnosis of SARS-CoV-2
infection in another population with a distinct genetic ancestry profile. Importantly, this is the first
study to explore CBC patterns among infected and uninfected individuals from Ecuador.

5 CONCLUSIONS

In conclusion, the current study highlights the importance of applying ML/FS approaches to
identify changes in peripheral blood parameters derived from CBC in positive and negative
individuals for SARS-CoV-2 infection and prioritize more relevant parameters of prediction
infection status in different populations. In summary, there was a distinct RBCP pattern between
the two populations in almost all parameters for both male and female individuals. However,
WBCP had no clear patterns. This result strengthens the need to evaluate and find these population
differences since it directly impacts the understanding of treatment options, outcomes, and
preventive measures.

Additionally, the interpretation of our results was in agreement with the genetic ancestry
background behind the CBC data of Brazil and Ecuador. Our findings demonstrate that the same
CBC parameters filtered by ML/FS approaches that allow distinguishing positive from negative
cases for COVID-19 in a given population will not necessarily be able to discriminate infection
status in another population. These approaches should be employed in a population-specific context, considering the differential aspects of genetic ancestry, evolutionary adaptations (e.g., high altitudes and oxygenation levels), and/or exposure to different environmental factors (e.g., infection by predominant SARS-CoV-2 genetic lineages in specific periods of the pandemic), which make these CBC patterns as complementary laboratory biomarkers for COVID-19 diagnosis that are unique to each population.

ACKNOWLEDGMENTS

This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul - FAPERGS [19/2551-0001906-8], Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq [311611/2018-4], and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - STICAMS'ID [88881.522073 / 2020-01]. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Authors thanks the Universitäti St.Gallen (HSG) through the Seed Money Grants 2020 Program under the project “A hemogram-data-based machine learning approach to optimize the use of COVID-19 diagnostic tests”.

References

[1] Kaustubh Adhikari, Javier Mendoza-Revilla, Juan Camilo Chacón-Duque, Macarena Fuentes-Guajardo, and Andrés Ruiz-Linares. Admixture in latin america. Current opinion in genetics & development, 41:106–114, 2016.

[2] Charles Agbuduwe and Supratik Basu. Haematological manifestations of covid-19: From
cytopenia to coagulopathy. *European journal of haematology*, 105(5):540–546, 2020.

[3] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B Munroe, Bina Joe, and Xi Cheng. Artificial intelligence and machine learning to fight covid-19, 2020.

[4] Ahmed Abdelaal Ahmed Mahmoud M Alkhatip, Mohamed Gomaa Kamel, Mohamed Khaled Hamza, Ehab Mohamed Farag, Hany Mahmoud Yassin, Mohamed Elayashy, Amr Ahmed Naguib, Mohamed Wagih, Fatma Abd-Elshahed Abd-Elhay, Haytham Zien Algameel, et al. The diagnostic and prognostic role of neutrophil-to-lymphocyte ratio in covid-19: a systematic review and meta-analysis. *Expert review of molecular diagnostics*, pages 1–10, 2021.

[5] J C Ang, A Mirzal, H Haron, et al. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5):971–989, 2016.

[6] Jun Chin Ang, Andri Mirzal, Habibollah Haron, and Haza Nuzly Abdull Hamed. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13:971–989, 2015.

[7] Eduardo Avila, Alessandro Kahmann, Clarice Alho, and Marcio Dorn. Hemogram data as a tool for decision-making in covid-19 management: applications to resource scarcity scenarios. *PeerJ*, 8:e9482, 2020.

[8] Abhirup Banerjee, Surajit Ray, Bart Vorselaars, Joanne Kitson, Michail Mamalakis, Simonne Weeks, Mark Baker, and Louise S Mackenzie. Use of machine learning and artificial intelligence to predict sars-cov-2 infection from full blood counts in a population. *International immunopharmacology*, 86:106705, 2020.
[9] Davide Brinati, Andrea Campagner, Davide Ferrari, Massimo Locatelli, Giuseppe Banfi, and Federico Cabitza. Detection of covid-19 infection from routine blood exams with machine learning: a feasibility study. Journal of medical systems, 44(8):1–12, 2020.

[10] Ming-Huei Chen, Laura M Raffield, Abdou Mousas, Saori Sakaue, Jennifer E Huffman, Arden Moscati, Bhavi Trivedi, Tao Jiang, Parsa Akbari, Dragana Vuckovic, et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell, 182(5):1198–1213, 2020.

[11] N. Cilia, C. De Stefano, F. Fontanella, S. Raimondo, and A. di Freca. An experimental comparison of feature-selection and classification methods for microarray datasets. Information (Switzerland), 2019.

[12] M. Dash and H. Liu. Feature selection for classification, in intelligent data analysis. Elsevier Intelligent Data Analysis, 1997.

[13] Rami M Elshazli, Eman A Toraih, Abdelaziz Elgaml, Mohammed El-Mowafy, Mohamed El-Mesery, Mohamed N Amin, Mohammad H Hussein, Mary T Killackey, Manal S Fawzy, and Emad Kardil. Diagnostic and prognostic value of hematological and immunological markers in covid-19 infection: A meta-analysis of 6320 patients. PloS one, 15(8):e0238160, 2020.

[14] FAPESP. FAPESP COVID-19 Data Sharing/BR, 2020. https://repositoriodatasharingfapesp.uspdigital.usp.br.

[15] Nuno R Faria, Thomas A Mellan, Charles Whittaker, Ingra M Claro, Darlan da S Candido, Swapnil Mishra, Myuki AE Crispim, Flavia CS Sales, Iwona Hawryluk, John T McCrone, et al. Genomics and epidemiology of the p. 1 sars-cov-2 lineage in manaus, brazil. Science, 372(6544):815–821, 2021.
[16] Vincius Bonetti Franceschi, Patrícia Aline Gröhs Ferrareze, Ricardo Ariel Zimerman, Gabriela Bettella Cybis, and Claudia Elizabeth Thompson. Mutation hotspots and spatiotemporal distribution of sars-cov-2 lineages in brazil, february 2020-2021. *Virus Research*, 304:198532, 2021.

[17] Ingrid Fricke-Galindo and Ramcés Falfán-Valencia. Genetics insight for covid-19 susceptibility and severity: A review. *Frontiers in Immunology*, 12:1057, 2021.

[18] Ri-Li Ge, Sarah Witkowski, Yu Zhang, Clarence Alfrey, Mark Sivieri, Trine Karlsen, Geir K Resaland, Matthew Harber, James Stray-Gundersen, and BD Levine. Determinants of erythropoietin release in response to short-term hypobaric hypoxia. *Journal of Applied Physiology*, 92(6):2361–2367, 2002.

[19] Jiao Gong, Jingyi Ou, Xueping Qiu, Yusheng Jie, Yaqiong Chen, Lianxiong Yuan, Jing Cao, Mingkai Tan, Wenxiong Xu, Fang Zheng, et al. A tool for early prediction of severe coronavirus disease 2019 (covid-19): a multicenter study using the risk nomogram in wuhan and guangdong, china. *Clinical infectious diseases*, 71(15):833–840, 2020.

[20] Ravindra K Gupta. Will sars-cov-2 variants of concern affect the promise of vaccines? *Nature Reviews Immunology*, 21(6):340–341, 2021.

[21] Bernardo Gutierrez, Sully Márquez, Belén Prado-Vivar, Mónica Becerra-Wong, Juan José Guadalupe, Darlan da Silva Candido, Juan Carlos Fernandez-Cadena, Gabriel Morey-Leon, Rubén Armas-Gonzalez, Derly Madeleiny Andrade-Molina, et al. Genomic epidemiology of sars-cov-2 transmission lineages in ecuador. *medRxiv*, 2021.

[22] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389–422, 2002.
[23] William T Harvey, Alessandro M Carabelli, Ben Jackson, Ravindra K Gupta, Emma C Thomson, Ewan M Harrison, Catherine Ludden, Richard Reeve, Andrew Rambaut, Sharon J Peacock, et al. Sars-cov-2 variants, spike mutations and immune escape. Nature Reviews Microbiology, 19(7):409–424, 2021.
[24] Peter J Hotez, Jorge A Huete-Perez, and Maria Elena Bottazzi. Covid-19 in the americas and the erosion of human rights for the poor, 2020.
[25] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, Charles N John, MD Iftikhar Hussain, and Muhammad Nabeel. Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app. Informatics in Medicine Unlocked, 20:100378, 2020.
[26] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. Journal of Big Data, 6(1):27, 2019.
[27] I T Jolliffe and J Cadima. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065):20150202, 2016.
[28] Colleen G Julian and Lorna G Moore. Human genetic adaptation to high altitude: evidence from the andes. Genes, 10(2):150, 2019.
[29] Esin Merve Erol Koç, Rahime Bedir Fındık, Hatice Akkaya, İşılay Karadağ, Eda Özden Tokaloğlu, and Özlem Moraloğlu Tekin. Comparison of hematological parameters and perinatal outcomes between covid-19 pregnancies and healthy pregnancy cohort. Journal of Perinatal Medicine, 49(2):141–147, 2021.
[30] Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM
Transactions on Computational Biology and Bioinformatics, 9(4):1106–1119, 2012.

[31] Joffrey L. Leevy, Taghi M. Khoshgoftaar, Richard A. Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1):42, 2018.

[32] Qiubai Li, Yulin Cao, Lei Chen, Di Wu, Jianming Yu, Hongxiang Wang, Wenjuan He, Li Chen, Fang Dong, Weiquan Chen, et al. Hematological features of persons with covid-19. Leukemia, 34(8):2163–2172, 2020.

[33] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250:113–141, 2013.

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(Nov):2579–2605, 2008.

[35] Luiz E. Mello, Andrea Suman, Claudia Bauzer Medeiros, Claudio Almeida Prado, Edgar Gil Rizzatti, Fatima L. S. Nunes, Gabriela F. Barnabé, João Eduardo Ferreira, José Sá, Luiz F. L. Reis, Luiz Vicente Pizzó, Luzia Sarno, Raphael de Lamonica, Rui Monteiro de Barros Maciel, Roberto Marcondes Cesar-Jr, and Rodrigo Carvalho. Opening Brazilian COVID-19 patient data to support world research on pandemics, 2020.

[36] Zivorad M Milenovic. Application of mann-whitney u test in research of professional training of primary school teachers. Metodicki obzori, 6(1):73–79, 2011.

[37] Bilal Mirza, Wei Wang, Jie Wang, Howard Choi, Neo Christopher Chung, and Peipei Ping. Machine learning and integrative analysis of biomedical big data. Genes, 10:87, 2019.

[38] Andrew J Murray, Hugh E Montgomery, Martin Feelisch, Michael PW Grocott, and Daniel S Martin. Metabolic adjustment to high-altitude hypoxia: from genetic signals to physiological implications. Biochemical Society Transactions, 46(3):599–607, 2018.
[39] Maria Nicola, Zaid Alsafi, Catrin Sohrabi, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, Maliha Agha, and Riaz Agha. The socio-economic implications of the coronavirus pandemic (covid-19): A review. *International journal of surgery*, 78:185–193, 2020.

[40] Carolina Kymie Vasques Nonaka, Tiago Gräf, Camila Araújo de Lorenzo Barcia, Vanessa Ferreira Costa, Janderson Lopes de Oliveira, Rogério da Hora Passos, Iasmin Nogueira Bastos, Maria Clara Brito de Santana, Ian Marinho Santos, Karoline Almeida Felix de Sousa, et al. Sars-cov-2 variant of concern p. I (gamma) infection in young and middle-aged patients admitted to the intensive care units of a single hospital in Salvador, northeast Brazil, February 2021. *International Journal of Infectious Diseases*, 111:47–54, 2021.

[41] Anton Pak, Oyelola A Adegboyce, Adeshina I Adekunle, Kazi M Rahman, Emma S McBryde, and Damon P Eisen. Economic consequences of the covid-19 outbreak: the need for epidemic preparedness. *Frontiers in public health*, 8:241, 2020.

[42] Sergio DJ Pena, Fabrício R Santos, and Eduardo Tarazona-Santos. Genetic admixture in Brazil. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 184, pages 928–938. Wiley Online Library, 2020.

[43] Paola Cristina Resende, Edson Delatorre, Tiago Gräf, Daiana Mir, Fernando Couto Motta, Luciana Reis Appolinario, Anna Carolina Dias da Paixão, Ana Carolina da Fonseca Mendonça, Maria Ogrzewalska, Braulia Caetano, et al. Evolutionary dynamics and dissemination pattern of the sars-cov-2 lineage b. 1.1. 33 during the early pandemic phase in Brazil. *Frontiers in microbiology*, 11, 2020.

[44] Paul Robach, Yvonne Fulla, Klaas R Westerterp, and Jean-Paul Richalet. Comparative
response of epo and soluble transferrin receptor at high altitude. *Medicine and science in sports and exercise*, 36(9):1493–8, 2004.

[45] Ronald Rodrigues de Moura, Antonio Víctor Campos Coelho, Valdir de Queiroz Balbino, Sergio Crovella, and Lucas André Cavalcanti Brandão. Meta-analysis of brazilian genetic admixture and comparison with other latin america countries. *American Journal of Human Biology*, 27(5):674–680, 2015.

[46] Andrés Ruiz-Linares, Kaustubh Adhikari, Víctor Acuña-Alonso, Mirsha Quinto-Sanchez, Claudia Jaramillo, William Arias, Macarena Fuentes, María Pizarro, Paola Everardo, Francisco De Avila, et al. Admixture in latin america: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS genetics*, 10(9):e1004572, 2014.

[47] Roberta Santangelo, Fabricio González-Andrade, Claus Børsting, Antonio Torroni, Vania Pereira, and Niels Morling. Analysis of ancestry informative markers in three main ethnic groups from ecuador supports a trihybrid origin of ecuadorians. *Forensic Science International: Genetics*, 31:29–33, 2017.

[48] Jasdeep Singh, Syed Asad Rahman, Nasreen Z Ehtesham, Subhash Hira, and Seyed E Hasnain. Sars-cov-2 variants of concern are emerging in india. *Nature medicine*, pages 1–3, 2021.

[49] Aracele Maria de Souza, Sarah Stela Resende, Taís Nóbrega de Sousa, and Cristiana Ferreira Alves de Brito. A systematic scoping review of the genetic ancestry of the brazilian population. *Genetics and molecular biology*, 42:495–508, 2019.

[50] Inge Stegeman, Eleanor A Ochodo, Fatuma Guleid, Gea A Holtman, Bada Yang, Clare Davenport, Jonathan J Deeks, Jacqueline Dinnes, Sabine Dittrich, Devy Emperador, et al.
Routine laboratory testing to determine if a patient has covid-19. *Cochrane Database of Systematic Reviews*, (11), 2020.

[51] Kaiming Tao, Philip L Tzou, Janin Nouhin, Ravindra K Gupta, Tulio de Oliveira, Sergei L Kosakovsky Pond, Daniela Fera, and Robert W Shafer. The biological and clinical significance of emerging sars-cov-2 variants. *Nature Reviews Genetics*, pages 1–17, 2021.

[52] Evangelos Terpos, Ioannis Ntanasis-Stathopoulos, Ismail Elalamy, Efstathios Kastriotis, Theodoros N Sergentanis, Marianna Politou, Theodora Psaltopoulou, Grigoris Gerotziafas, and Meletios A Dimopoulos. Hematological findings and complications of covid-19. *American journal of hematology*, 95(7):834–847, 2020.

[53] J. Utans, J. Moddy, and S. Rehfuss. Input variable selection for neural networks: application to predicting the u.s. business cycle. *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFER)*, 1995.

[54] R Varshavsky, A Gottlieb, M Linial, et al. Novel unsupervised feature filtering of biological data. *Bioinformatics*, 22(14):e507–e513, 2006.

[55] Dragana Vuckovic, Erik L Bao, Parsa Akbari, Caleb A Lareau, Abdou Mousas, Tao Jiang, Ming-Huei Chen, Louis M Rafffield, Manuel Tardaguila, Jennifer E Huffman, et al. The polygenic and monogenic basis of blood traits and diseases. *Cell*, 182(5):1214–1231, 2020.

[56] Guangyao Wu, Pei Yang, Yuanliang Xie, Henry C Woodruff, Xiangang Rao, Julien Guiot, Anne-Noelle Frix, Renaud Louis, Michel Moutschen, Jiawei Li, et al. Development of a clinical decision support system for severity risk prediction and triage of covid-19 patients at hospital admission: an international multicentre study. *European Respiratory Journal*, 56(2), 2020.

[57] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan
Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. An interpretable mortality prediction model for covid-19 patients. *Nature machine intelligence*, 2(5):283–288, 2020.

[58] Haochen Yao, Nan Zhang, Ruochi Zhang, Meiyu Duan, Tianqi Xie, Jiahui Pan, Ejun Peng, Juanjuan Huang, Yingli Zhang, Xiaoming Xu, et al. Severity detection for the coronavirus disease 2019 (covid-19) patients using a machine learning model based on the blood and urine tests. *Frontiers in cell and developmental biology*, 8:683, 2020.

[59] Ana Karina Zambrano, Aníbal Gaviria, Santiago Cobos-Navarrete, Carmen Gruezo, Cristina Rodríguez-Pollit, Isaac Armendáriz-Castilla, Jennyfer M García-Cárdenas, Santiago Guerrero, Andrés López-Cortés, Paola E Reone, et al. The three-hybrid genetic composition of an ecuadorian population using snps-indels compared with autosomes, mitochondrial dna and y chromosome data. *Scientific reports*, 9(1):1–8, 2019.

[60] Yihua Zhu, XingJian Cao, Yonghui Lu, Dongsheng Xu, Renfei Lu, and Xinling Li. Lymphocyte cell population as a potential hematological index for early diagnosis of covid-19. *Cellular and Molecular Biology*, 66(7):202–206, 2020.

| Dataset | Samples | PCR Positive | PCR Negative | Shannon Entropy |
|---------|---------|--------------|--------------|-----------------|
|         | Selected | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| Br-v1   | 4473    | 744  | 621   | 1365  | 1440 | 1668   | 3108  | 0.93 | 0.84   | 0.89  |
| Ecu-v1  | 375     | 75   | 49    | 124   | 160  | 91     | 251   | 0.90 | 0.93   | 0.92  |
| Br-v2   | 375     | 75   | 49    | 124   | 160  | 91     | 251   | 0.90 | 0.93   | 0.92  |

Table 2: Dataset preprocessing, displaying the mean age of males and females individuals in positive and negative samples.
| Dataset | PCR Positive          | PCR Negative          |
|---------|-----------------------|-----------------------|
|         | Male                  | Female                | Total                | Male                  | Female                | Total                |
| Br-v1   | 53.27 ± 15.91         | 52.07 ± 17.28         | 52.72 ± 16.55        | 46.60 ± 19.16         | 45.44 ± 17.65         | 45.98 ± 18.37        |
| Ecu-v1  | 38.40 ± 14.20         | 37.10 ± 14.30         | 37.06 ± 14.31        | 37.10 ± 11.30         | 36.20 ± 13.50         | 36.80 ± 12.10        |
| Br-v2   | 38.40 ± 14.25         | 35.02 ± 14.29         | 37.09 ± 11.26        | 36.16 ± 13.48         | 36.76 ± 12.09         |

Figure 1: PCA analysis, displaying the distribution of negative and positive cases, divided by male (M) and female (F) individuals, and by population. A) Distribution for M, F and MF individuals, respectively, in Ecuador. B) Distribution for M, F and MF individuals, respectively, in Brazil, after data treatment; and C) Distribution for M, F and MF, respectively, individuals in Brazil and Ecuador.

Figure 2: T-SNE analysis, displaying the distribution of negative and positive cases, divided by male (M) and female (F) individuals, and by population. A) Distribution for M, F and MF individuals, respectively, in Ecuador. B) Distribution for M, F and MF individuals, respectively, in Brazil, after data treatment; and C) Distribution for M, F and MF, respectively, individuals in Brazil and Ecuador.

Table 3: Statistical significance (Mann-Whitney U test) of the CBC parameters’ differences between COVID-19 positive and negative individuals considering gender and country.
| Cell Type  | U Value | p Value |
|-----------|---------|---------|
| Lymphocytes | (U=2238.0) | p<0.554) |
|           | (U=4874.5) | p<0.012) |
|           | (U=13868.5) | p<0.013) |
|           | (U=2068.0) | p<0.990) |
|           | (U=4796.5) | p<0.045) |
|           | (U=13224.0) | p<0.086) |
| Monocytes | (U=2303.0) | p<0.972) |
|           | (U=6123.0) | p<0.020) |
|           | (U=15707.5) | p<0.087) |
|           | (U=2978.5) | p<0.018) |
|           | (U=5768.0) | p<0.013) |
|           | (U=17052.0) | p<0.132) |
| Eosinophils | (U=1197.5) | p<0.000) |
|           | (U=3331.5) | p<0.000) |
|           | (U=8477.5) | p<0.000) |
|           | (U=1301.0) | p<0.039) |
|           | (U=4998.5) | p<0.000) |
|           | (U=11594.0) | p<0.000) |
| Basophils | (U=1470.0) | p<0.001) |
|           | (U=3910.0) | p<0.000) |
|           | (U=10174.5) | p<0.000) |
|           | (U=1541.0) | p<0.000) |
|           | (U=4435.0) | p<0.000) |
|           | (U=11269.5) | p<0.000) |
| Neutrophils | (U=1724.0) | p<0.027) |
|           | (U=5606.0) | p<0.419) |
|           | (U=13425.0) | p<0.000) |
|           | (U=1140.0) | p<0.000) |
|           | (U=5280.0) | p<0.000) |
|           | (U=11550.0) | p<0.000) |
| Lymphocytes | (U=1277.0) | p<0.000) |
|           | (U=2142.5) | p<0.000) |
|           | (U=6835.0) | p<0.000) |
|           | (U=721.5) | p<0.000) |
|           | (U=3406.0) | p<0.000) |
|           | (U=7469.5) | p<0.000) |
| NLR | (U=2360.5) | p<0.571) |
|           | (U=7502.5) | p<0.002) |
|           | (U=18377.0) | p<0.004) |
|           | (U=2337.0) | p<0.000) |
|           | (U=7007.0) | p<0.038) |
|           | (U=17538.0) | p<0.045) |
| Monocytes | (U=1627.5) | p<0.008) |
|           | (U=4708.5) | p<0.008) |
|           | (U=11818.0) | p<0.000) |
|           | (U=1467.0) | p<0.000) |
|           | (U=4510.0) | p<0.002) |
|           | (U=10929.5) | p<0.000) |
| Eosinophils | (U=819.5) | p<0.000) |
|           | (U=2505.0) | p<0.000) |
|           | (U=6223.5) | p<0.000) |
|           | (U=993.5) | p<0.000) |
|           | (U=4276.0) | p<0.000) |
|           | (U=9694.5) | p<0.000) |
| Basophils | (U=1384.5) | p<0.000) |
|           | (U=3733.0) | p<0.000) |
|           | (U=9654.0) | p<0.000) |
|           | (U=744.0) | p<0.000) |
|           | (U=3756.0) | p<0.000) |
|           | (U=8269.5) | p<0.000) |
| RBCs | (U=2756.0) | p<0.021) |
|           | (U=7481.0) | p<0.002) |
|           | (U=18081.0) | p<0.011) |
|           | (U=2370.0) | p<0.542) |
|           | (U=5081.0) | p<0.059) |
|           | (U=13944.5) | p<0.102) |
|     | MCV   | MCH   | MCHC  | RDW   | Hemoglobin | Hematocrit | Platelets | MPV   |
|-----|-------|-------|-------|-------|------------|------------|-----------|-------|
|     | (U=2527.0) | (U=2074.0) | (U=1882.5) | (U=2042.0) | (U=2648.5) | (U=2769.5) | (U=1365.5) | (U=2591.0) |
|     | p<0.195 | p<0.500 | p<0.131 | p<0.415 | p<0.068 | p<0.018 | p<0.000 | p<0.115 |
|     | (U=6843.5) | (U=4330.0) | (U=3944.0) | (U=4767.5) | (U=6539.5) | (U=7649.5) | (U=2945.0) | (U=391.0) |
|     | p<0.083 | p<0.001 | p<0.000 | p<0.011 | p<0.268 | p<0.001 | p<0.050 | p<0.422 |
|     | (U=17815.5) | (U=12445.5) | (U=11411.5) | (U=13157.0) | (U=16643.0) | (U=18603.5) | (U=8551.0) | (U=17147.0) |
|     | p<0.022 | p<0.002 | p<0.000 | p<0.015 | p<0.274 | p<0.002 | p<0.000 | p<0.109 |
|     | (U=2056.0) | (U=1918.5) | (U=2100.0) | (U=2592.5) | (U=2239.0) | (U=2164.5) | (U=1298.5) | (U=2102.0) |
|     | p<0.451 | p<0.176 | p<0.574 | p<0.114 | p<0.969 | p<0.781 | p<0.000 | p<0.580 |
|     | (U=6362.5) | (U=5899.5) | (U=5378.0) | (U=6420.5) | (U=4707.0) | (U=5113.0) | (U=4352.0) | (U=7118.5) |
|     | p<0.458 | p<0.838 | p<0.201 | p<0.389 | p<0.008 | p<0.068 | p<0.001 | p<0.021 |
|     | (U=15729.5) | (U=14695.0) | (U=14256.5) | (U=17302.5) | (U=12953.5) | (U=13327.5) | (U=10301.5) | (U=17038.0) |
|     | p<0.866 | p<0.381 | p<0.187 | p<0.078 | p<0.008 | p<0.024 | p<0.000 | p<0.135 |

Figure 3: Distributions of white and red blood cells as well as platelets related variables for positive (red) and negative (green) classes of the datasets: Brazil-V2 and Ecuador - Gender Feminine. The central white dot is the median. A) Violin charts for white blood cell-derived parameters. B) Violin charts for red blood cell-derived parameters and platelet-derived variables.

Figure 4: Distributions of white and red blood cells as well as platelets related variables for positive (red) and negative (green) classes of the datasets: Brazil-V2 and Ecuador - Gender Masculine. The central white dot is the median. A) Violin charts for white blood cell-derived
parameters. B) Violin charts for red blood cell-derived parameters and platelet-derived variables.

Figure 5: Distributions of white and red blood cells as well as platelets related variables for positive (red) and negative (green) classes of the datasets: Brazil-V2 and Ecuador - Gender Masculine and Feminine. The central white dot is the median. A) Violin charts for white blood cell-derived parameters. B) Violin charts for red blood cell-derived parameters and platelet-derived variables.

Figure 6: Feature Selection (RFE-SVM) for each dataset. A) FS results for Brazil (BR-V2); B) FS results for Ecuador (ECU); and C) FS results for Ecuador + Brazil (ECU+BR-v2). The results consider gender (M,F) and the population (M+F).
Highlights (for review)

- We used Machine Learning to compare CBC data between two highly genetically distinct Latin American countries: Brazil and Ecuador.

- The study highlights the importance of applying ML/FS approaches to identify changes in peripheral blood parameters derived from CBC in positive and negative individuals for SARS-CoV-2 infection.

- The same CBC parameters filtered by ML/FS approaches that allow distinguishing positive from negative cases for COVID-19 in a given population will not necessarily be able to discriminate infection status in another population.

- We found striking differences in CBC profiles from the two countries, especially in red-blood-cell counts.
Figure 1
Figure 2
Figure 3
Figure 5
Figure 6

A

BR-v2-M

n_features = 16
score = 0.758

BR-v2-F

n_features = 16
score = 0.906

BR-v2-MF

n_features = 11
score = 0.797

B

ECU-M

n_features = 2
score = 0.770

ECU-F

n_features = 12
score = 0.779

ECU-MF

n_features = 3
score = 0.792

C

BR-V2-ECU-M

n_features = 15
score = 0.743

BR-V2-ECU-F

n_features = 12
score = 0.794

BR-V2-ECU-MF

n_features = 15
score = 0.791