Exploratory Analysis of COVID-19 Patients Usingprincipal Component Analysis, Feature Selection and Predictive Algorithms

Danilo Carlotti (danilopcarlotti@usp.br)
Universidade de São Paulo
Lewis Buss
Universidade de São Paulo
Gabriel Leite
Universidade de São Paulo
Anna Levin
Universidade de São Paulo
Ester Sabino
Universidade de São Paulo
Fátima Nunes
Universidade de São Paulo
João Ferreira
Universidade de São Paulo

Research Article

Keywords: SARS-Cov-2, tests, prediction, Open data, Open Science, COVID-19, RT-PCR, machine learning

Posted Date: January 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-132785/v1

Danilo Panzeri Nogueira Carlotti (corresponding author)
Post doc researcher
Faculty of Computer Science
University of São Paulo
danilopcarlotti@usp.br
Lewis Buss
Master candidate
Faculty of Medicine
University of São Paulo
lewisbuss@gmail.com
Gabriel Leite
Master candidate
Faculty of Medicine
University of São Paulo
gabriel.fialkovitz@hc.fm.usp.br
Anna Sarah Levin
Associate Professor
Faculty of Medicine
University of São Paulo
anna@usp.br
Ester Sabino
Associate Professor
Faculty of Medicine
University of São Paulo
sabinoec@gmail.com
Fátima Nunes
Full professor
The School of Arts, Sciences and Humanities
University of São Paulo
fatima.nunes@usp.br
João Eduardo Ferreira
Full professor
Faculty of Computer Science
University of São Paulo
jef@ime.usp.br

Exploratory analysis of COVID-19 patients using principal component analysis, feature selection and predictive algorithms

Exploratory analysis of COVID-19 patients using principal component analysis, feature selection and predictive algorithms

Abstract

A. Background

The novel coronavirus disease (COVID-19) emerged in late 2019 has shown that research done with open data could be the cornerstone for overcoming the need for collaborative, optimized and urgent analysis. Although several articles have been published, identification of variables that can have correlation with positive PCR results is still a challenge. In this paper we show a concrete example of open data analysis from 910 patients attended in the hospital undergoing SARS-CoV-2 RT-PCR in three private institutions in São Paulo, Brazil.

B. Results

We performed an exploratory analysis using principal component analysis, feature selection and predictive algorithms to test for associations between a number of laboratory test abnormalities and the SARS-CoV-2 RT-PCR result. More concretely, we found a set of 18 variables that showed some association with a positive PCR result.

C. Conclusion

Among these variables elevated lactic dehydrogenase (LDH) and d-dimer were the most correlated with a positive RT-PCR. We developed a classifier that achieved 76% mean accuracy, 77% mean precision and 92% mean sensitivity to identify individuals with COVID-19.

Index Terms

SARS-Cov-2, tests, prediction, Open data, Open Science, COVID-19, RT-PCR, machine learning

I. Declarations

A. Ethics approval and consent to participate

Not applicable

B. Consent for publication

Not applicable

C. Availability of data and materials

The datasets analysed during the current study are available in the FAPESP COVID-19 DataSharing/BR Repository, available at https://repositoriodatasharingfapesp.uspdigital.usp.br/.

D. Competing interests

The authors declare that they have no competing interests.

E. Funding

Not applicable

F. Authors' contributions

DPNC implemented and helped define the algorithms for the classification process. FN and JEF defined and chose the analytical methods used in the paper. The writing of the methods and computational analysis was done by DPNC, FN and JEF. LB, GL, AL and ES wrote about the medical background and literature, the criterion used to select the patients, as well as the interpretation of the results from a medical point of view. All authors read and approved the final manuscript.

G. Acknowledgements

Not applicable

II. Background

The novel coronavirus disease (COVID-19) emerged in late 2019 in China. It has since become a pandemic, with almost 70 million cases and over 1.5 million deaths at the time of writing according to World Health Organization [1]. In an effort to accelerate understanding of the newly-emerged SARS-CoV-2 virus, some institutions have perused a policy of rapid data sharing [2]. For example, the São Paulo Research Foundation (FAPESP – Portuguese abbreviation), one of the main Brazilian research funding agencies, has led an initiative, called the COVID-19 Data Sharing/BR repository, to publish open data on Brazilian COVID-19 patients [3]. The published data include laboratory tests and clinical outcomes [4]. A relational database model has been released to support the use of these data [5].

Real-time reverse-transcriptase polymerase chain reaction (RT-PCR) is the standard method to diagnose current infection with SARS-CoV-2, the causative agent of COVID-19. The sensitivity (true-positive rate) of RT-PCR varies widely depending on when the test is performed in relation to symptom onset [6]. Test performance also varies between RT-PCR kits [7]. The potentially high false-negative rate justifies the use of additional clinical indicators, such as computed tomography images, to rule the diagnosis in or out [8]. However, identification of correlation between exam variables and positive PCR results is still a challenge in literature.

The growing availability of data generated in routine care – e.g., from electronic health records, administrative health databases etc – can be exploited to identify additional clinical indicators of SARS-CoV-2 infection. Herein, we propose an approach to filter relevant data from the FAPESP open data repository [3], aiming to answer the following question: is there a set of low-cost laboratory results, or other parameters, that can aid the diagnosis of COVID-19?". More concretely, we describe the process of selecting relevant variables and the use of machine learning techniques to answer our research question.

III. Related works

Research done with open data has some challenges. Although institutions may present the data in a predictable manner, the case of the FAPESP dataset, the data itself is not always standardized. This is the case of the exam's descriptions and the reference value for the exams.

In [9] it is discussed common problems with open data sources of COVID19. The authors of [10] explored the FAPESP dataset using limited data mining techniques and discuss inconsistencies and outliers present in this dataset, given that the data is not standardized between institutions. Another working paper [11] discusses correlations between age, sex and Systemic Inflammation in Individuals with COVID-19. This later article is another example of the problems of dealing with a dataset not standardized with some missing values. In [12], the quality of open access image datasets is discussed and a pre-processing method for these images in order to increase their quality is discussed.

After this first step of joining, standardising and clearing the data is overcome, knowledge discovery is the next challenge. Some initiatives have been published to analyze a variety of issues such as SARS-CoV-2 virus-host interaction mechanisms [13], chest X-Ray Images [14], presence of neuropsychiatric manifestations in COVID-19 adult patients [15], and establish the difference between COVID-19 and community acquired pneumonia and other lung diseases [8].

IV. Methods

We analyzed data from 910 patients attended in three private institutions in São Paulo, undergoing SARS-CoV-2 RT-PCR [3]. We performed an exploratory analysis using principal component analysis, feature selection and predictive algorithms to test for associations between a number of laboratory test abnormalities and the SARS-CoV-2 RT-PCR result.

A. Data sources

The data for this study was obtained from the repository COVID-19 Data Sharing/BR [3]. A full description of the repository and programs used to process the information can be found in [5]. The repository contains anonymized patient data and laboratory test results. The full dataset includes data from a private medical laboratory (Grupo Fleury), as well as patient record information from Albert-Einstein and Sírio-Libanês hospitals. For a subset of patients attended in the hospitals the clinical outcome is also recorded. There was a total 327,844 patients undergoing SARS-CoV-2 RT-PCR (index PCR) in the dataset. The number of tests included in the repository and the site where they were collected can be seen in Table 1.

TABLE I: Location where the exams were performed.

Sites where tests were collected	Number of laboratory tests done
Private medical laboratory	5861700 (61%)
Hospital	3408303 (35%)
Intensive Care Unit	137209 (1.5%)
Inpatient ward	73346 (<1%)
Emergency Room	20373 (<1%)
Other	26115 (<1%)

B. Overview of the analytic approach

Figure 1 presents an overview of our analytic approach. After the selection process, we normalized the value of exams of the selected patients, as detailed in Section IV-B2. Then, we classified patients into two classes: positive and negative according to the RT-PCR result, as presented in Section V-1. For this set of patients, we selected the respective other exams obtained until two weeks before or after this RT-PCR date (Section V-2). The resultant set of exams was submitted to Machine Learning algorithms as detailed in Section V-3.

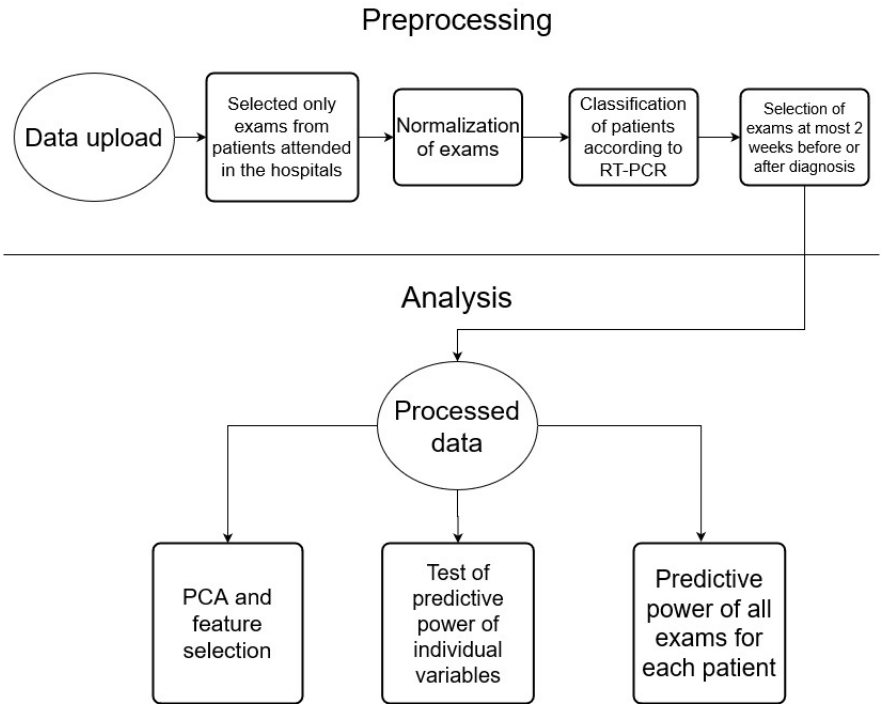


Fig. 1: General view of the approach used to analyze the data.

1) Patient selection and timing considerations: The population of interest was patients undergoing SARS-CoV-2 RT-PCR testing in a hospital, and therefore in whom the diagnosis of COVID-19 was considered. In order to select patients meeting these criteria we first filtered only cases in which the RT-PCR test was performed in an emergency department or inpatient setting. Thus, the studied database included 910 patients. As a single patient could contribute multiple RT-PCR results, we selected only a single instance of testing for each patient. In patients testing positive we selected the first positive result, as this was likely to be closest in time to the onset of symptoms. Where an individual had multiple negative results we selected only the first instance of testing.

Next, in order to include laboratory test that were performed close in time to the index RT-PCR, we defined a window of two weeks either side of the RT-PCR test for inclusion of these variables. This was to ensure that the laboratory results were in relation to the episode of care in which the RT-PCR was performed, and not related to another hospital admission. If an exam was done more than once for the same individual in the time window, we considered the exam closest to the date of the index RT-PCR.

2) Data normalization: The results of laboratory exams vary according to their reference range and units. For feature selection and predictive algorithms, normalized values are required. Normalization was performed as follows:

- all instances of each laboratory test was selected in the database;
- where a reference range (upper and lower bound) was provided (e.g. glucose), we calculated the difference between the observed result (e.g. recorded blood glucose level for a patient) and the mean of the upper and lower bounds for the given laboratory test;
- where only an upper lower or upper bound was provided, we calculated the difference between the result and the upper or lower limit provided;
- where there was no reference value (e.g. RT-PCR), numerically positive results were assigned a value of 1, and all other results were assigned a zero value.

The final scaling between zero (miminum value found) and one (maximum value found), was made with sklearn's MinMaxScaler [16].

V. Analysis Steps

1) Patients classification: We divided the patients into positive and negative RT-PCR classes. Since the dataset was not standardized in relation to RT-PCR result, we analyzed the dataset in order to identify descriptions that could be related to negative or positive RT-PCR. The text of all variables was kept just as it was made available by the institutions (in Portuguese language) in order to guarantee the reproducibility of the analysis.

Thus, we defined a POSITIVE_COVID class, related to a subject tested positive for the virus covid at the time. The original descriptions considered in this class were: "NOVO CORONAVÍRUS 2019 (SARS-CoV-2), DETECÇÃO POR PCR", "PCR em tempo real para detecção de Coron", and "COVID-19-PCR para SARS-COV-2, Vários Materiais (Fleury)".

All the remaining patients with descriptions different from those above mentioned were considered as belonging to negative class. After this step, 629 (69%) individuals were considered within the negative class and 281 (31%) were considered as positive class.

2) Selection of laboratory tests: we first selected the most frequently performed laboratory tests in the dataset. After considering individual analytes, we found that many laboratory tests had a large amount of missing values and were not clinically informative. Table II presents the laboratory test selected for analysis.

TABLE II: Variables (Laboratory tests and personal data) selected for analysis

Analytes included in the analysis	Clinical description
Total bilirubin	Liver function
Aspartate transaminase (AST)	Liver enzymes
Alanine aminotransferase (ALT)	
Creatinine	Renal function
Urea	
Hemoglobin	Full blood count – individual cell counts
Leucocytes	
Neutrophils	
Platelets	
Eosinophils	
Basophils	
Lymphocytes	
Monocytes	
Lactate dehydrogenase (LDH)	Marker of cell turnover
d-dimer	Fibrin breakdown product
Age	Demographic characteristics
Sex	

3) Analysis: to reduce the dimensionality of the dataset and verify the existence of correlations between the variables, we used Principal Component Analysis (PCA). The chosen number of components was three and, as demonstrated by the graph in Figure 2, they are able to explain more than 95% of variance in the data The variance explained by more than three components is rather marginal and it was found to be enough these three components.

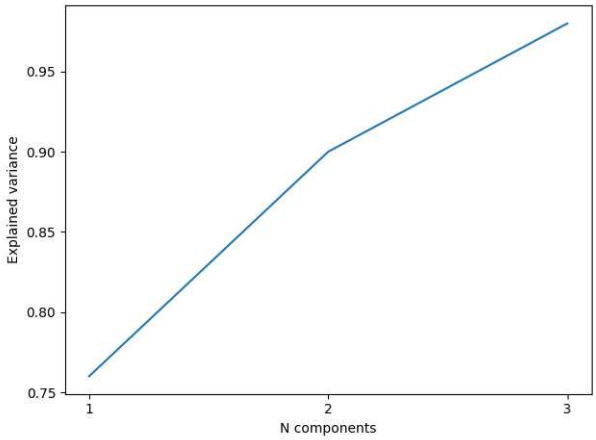


Fig. 2: The variance explained by PCA components was 1 (76%), 2 (90%) and 3 (96%)

.

In order to compare this technique with other means of testing for correlation between the variables (exams) and the dependent variable (diagnosis) we used Decision trees and Logistic regression with lasso to also select the five most important features in the data for patients from each class. The number of five variables was empirically set to explore the main features of the data responsible for predicting the outcome and they represent 27% of the available variables.

Logistic regression with lasso [17] is a method for feature selection that forces the algorithm to choose only some of the variables available to predict the outcome. To predict which of the variables were the most important, the dataset was split into two groups. The first is the training group (90% of the dataset), which was used to fit the model. The second group, the test set, was not used. The other parameters used in this implementation were the standard ones in the scikit-learn implementation of the algorithm. On the other hand, in order to test how accurate or possible it is for each variable to predict the diagnosis, when compared to a gold standard test, their values were trained and validated with a random forest algorithm with a random state set to 0 and the other parameters ere the standard ones in the scikit-learn implementation of the algorithm. This algorithm, after multiple tests, was the best performing algorithm to predict the outcome and it works well with a set of datapoints with a limited number of variables with a normalized range. The implementation used of these algorithms is the one found in python's library scikit-learn [16].

VI. Results

A. Dimensionality reduction

After transforming the data to reduce the dimensionality it is possible to select only the most important variables for each of the three principal components. For each class these are the five variables that have the highest correlation with the result, positive or negative. The descriptions are presented as they were written in the original records in Portuguese Language.

- Class: POSITIVE_COVID;
- Explanatory variables:
  - 'AA_NASCIMENTO'
  - 'Dosagem de D-Dímero__D-Dímero'
  - 'sexo_F'
  - 'Dosagem de Desidrogenase Láctica__Dehidrogenase Láctica'
  - 'Hemograma__Hemoglobina'

The first three variables are the most important and the biggest contributors to the components. In order to visualize the variables and their importance in the components, Figure 3 presents the loading plot showing the data after transformation by PCA and the three main variables as vectors, their size representing their importance in the components (x: Principal Component 1, y: Principal Component 2). The reason why the data is split into two homogeneous groups is because of the sex, which separates the data points in two.[1]

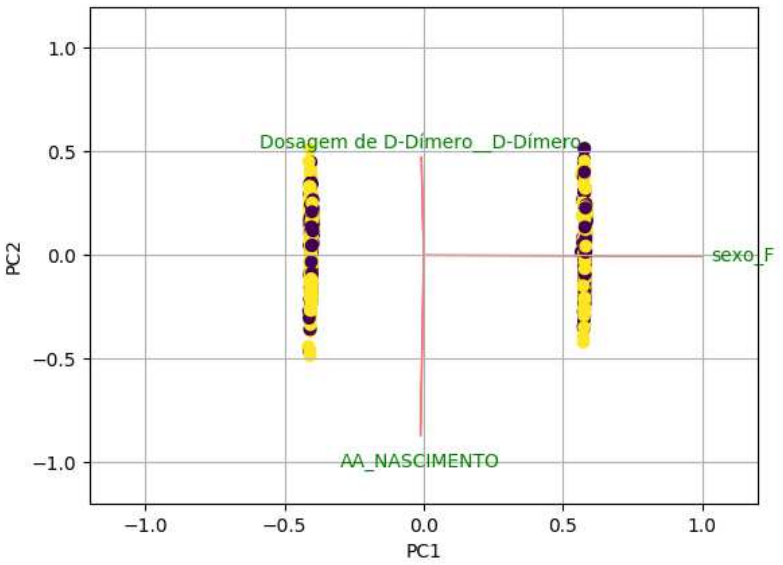


Fig. 3: Biplot of the Principal Component Analysis - main variables

It is important to emphasize that PCA creates a model based on the correlation of the variables and their variance.It can be several variables that, alone, do not have the power to predict the outcome (diagnosis) or be individually correlated with the outcome, but when considered together they may contribute to a component.

[1]The text of all variables was kept just as it was made available by the institutions in order to guarantee the reproducibility of the analysis

B. Feature selection

The feature selection algorithms were used to select the five features that, for each class, are most correlated with the diagnosis. These are the common variables found to be the most important by Logistic Regression with Lasso and Decision Trees. There are overlaps between the variables found by PCA and the other algorithms indicated by the symbol * after the variable name in Table III. The algorithm Logistic Regression with Lasso found only two variables with their beta coefficient bigger than zero, meaning that only two variables were deemed to be relevant for the prediction according to this algorithm.

TABLE III: Variables selected by the algorithms.

Algorithm	Variables selected
Logistic regression with lasso	Age*
	Sex *
	D-dimer *
	LDH
Decision trees	LDH
	Leucocytes
	Basophils - manual differential
	Total bilirubin
	Eosinophils - manual differential

C. Diagnosis prediction

The variables indicated as discriminatory by the algorithms mentioned were confirmed by physicians based on the literature. Lactic dehydrogenase has been shown to be altered in COVID-19 [18]. Furthermore, it has been evaluated as a biomarker for lung infections caused by Pneumocystis [19]. Although not specific, an altered LDH could be an interesting suggestion of a false-negative RT-PCR in patients with clinically suspected infection.

In recent studies D-dimer was described as a factor of prognosis in patients with COVID-19 [20]. In comparison to comunnity acquired pneumonia, patients with COVID-19 presented higher D-dimer levels, which could represent a multifactorial increased prothrombotic state [21]. Lactate dehidrogenase was also found to be increased in patients with COVID-19 and was related to increased deterioration in patients with mild disease [22].

Hypercoagulability is an important aspect of COVID-19. Overall, the frequency of venous thromboembolism (VTE) was found to be approximately 20% of patients in a literature review, and of stroke approximately 3%. Cumulative incidences were reported as high as 49% at varying time periods. There was an unusually high frequency of pulmonary embolism and the frequency of VTE was significantly higher in severely ill patients admitted to the ICU, compared to patients admitted to regular wards [23]. Markers of coagulation are potentially interesting for prognosis. D-dimer is a degradation product of hydrolysis of fibrin. A high D-dimer is nonspecific and often associated with various medical conditions such as infections, trauma, or even hospitalization [23]. The higher the D-dimer in COVID-19, the higher is the risk in-hospital mortality [24]. This is found in other studies [25, 26]. In one study, patients with high D-dimer had 12 times the risk of presenting severe COVID-19 [25]. The initial D-dimer level is higher in patients whose outcome is death than in survivors, and so is the peak level of D-dimer [27]. Thus, in the presence of a suspected case of COVID-19 with a negative RT-PCR, an altered D-dimer could be an interesting suggestion of a false-negative test.

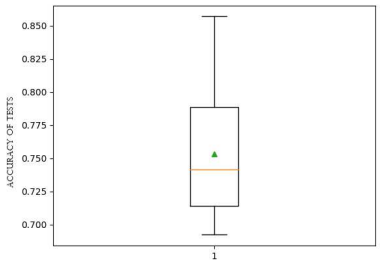
Figures 4a, 4b and 4c show, respectively, accuracy, precision and sensitivity of Random Forest Algorithm to predict COVID-19 diagnosis from the variables previously selected. As it can be noted, the accuracy has a mean value of about 76%, assuming values from 70% to a little more than 85%. The precision has a mean of 77% and varies from 72% to about 83% in most tests. The sensitivity has a mean value of 92% and ranges from 87% to 99%, depending on the random split of the data. In the graphs, the triangle represents the mean and the orange line the median. The range is the set of values obtained in the process of cross-validation, in which the data is randomly split into test and validation groups to test the model.

Furthermore, we estimate that between 13% and 18% of all the patients in our sample had been infected by SARS-CoV-2 even though their RT-PCR was negative, thus they were false-negative according to the algorithm in Figure 5. This was done by splitting the dataset randomly 20 times and comparing the predictions made by the random forest algorithm and the diagnosis made with RT-PCR.

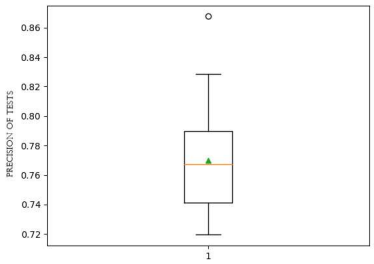
There were no single variables with significant enough predictive power when considered alone. Then, all the variables for the 910 patients were considered together, and the random forest classifier was repeatedly trained and validated using different random parts of the dataset using sklearn's cross-validation function.

VII. Discussion

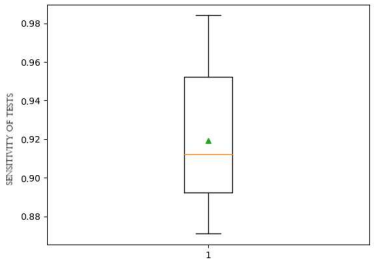
The main contribution of this paper is the use of filters designed by a medical team coupled with mathematical analysis. Our mathematical analysis includes correlation, feature selection with PCA, logistic regression with lasso, and decision tree algorithms using random forests to find laboratory tests and other variables associated with a positive RT-PCR. We found a set of 18 variables that showed some association with a positive PCR result. Among these variables elevated lactic dehydrogenase



(a) Accuracy



(b) Precision



(c) Sensitivity

Fig. 4: Results using Random Forest algorithm.

(LDH) and d-dimer were the most correlated with a positive RT-PCR. We developed a classifier that achieved 76% mean accuracy, 77% mean precision and 92% mean sensitivity to identify individuals with COVID-19.

The diagnosis of acute ongoing infection by SARS-CoV-2 is made by RT-PCR, usually using oropharyngeal and nasopharyngeal swabs. However, in a meta-analysis [28], the proportion of false-negative results varies according to the day on which the specimen was collected. On the first day post exposure, the probability of a false-negative RT-PCR was practically 100% in patients who were infected by the virus. This proportion decreased over time. On day 4, this probability was 38%. This proportion reached a minimum of 20% on day 8 (equivalent to 3-4 days after symptom onset). Because of this high proportion of false negative results, RT-PCR is an unreliable way of excluding the diagnosis of COVID-19 among suspected cases. The implications are: the risk of patients transmitting the virus despite a negative test, especially in the hospital (to other patients and healthcare workers); the false feeling of security that is conveyed by the test leading to non-adherence to preventive measures; and minimizing the importance of symptoms and of clinical follow-up [29]. Finding laboratorial characteristics that can suggest that patients are infected, despite a negative RT-PCR test, may direct clinical management and lead to measures to prevent transmission. Especially if these are tests easily available and used for evaluating the clinical condition of patients when they are attended in the hospital.

Lactic dehydrogenase has been shown to be raised in SARS-CoV-2 infection [18]. Furthermore, it has been evaluated as a biomarker for lung infections caused by Pneumocystis [19]. Although not specific, an altered LDH could indicate a false-negative RT-PCR in patients with clinically suspected infection.

D-dimer has been shown to be a prognostic factor in patients with COVID-19 [20]. In comparison to community acquired pneumonia, patients with COVID-19 presented higher D-dimer levels, which could represent a multifactorial prothrombotic state [21]. Lactate dehydrogenase was also found to be increased in patients with COVID-19 and was related to increased risk of deterioration in patients with mild disease [22].

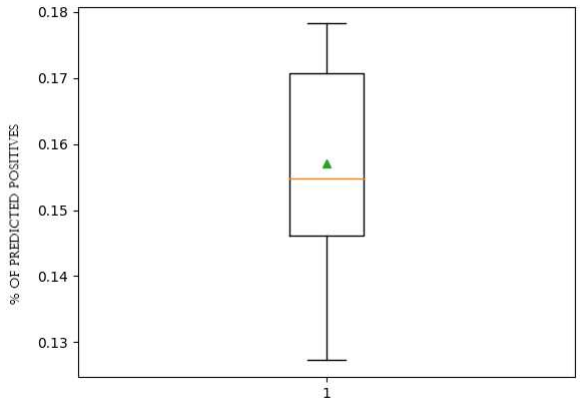


Fig. 5: Percentage of predicted positives according to RF

Hypercoagulability is an important aspect of COVID-19. Overall, the frequency of venous thromboembolism (VTE) was found to be approximately 20% of patients in a literature review, and of stroke approximately 3%. Cumulative incidences were reported as high as 49% at varying time periods. There was an unusually high frequency of pulmonary embolism and the frequency of VTE was significantly higher in severely ill patients admitted to the ICU, compared to patients admitted to regular wards [23].

Markers of coagulation are potentially useful prognostic markers. D-dimer is a fibrin degradation product. A high D-dimer is nonspecific and often associated with various medical conditions such as infections, trauma, or even hospitalization [23]. The higher the D-dimer in COVID-19, the higher is the risk of in-hospital mortality [24]. This is found in other studies [25, 26]. In one study, patients with high D-dimer had 12 times the risk of presenting severe COVID-19 [25]. The initial D-dimer level is higher in patients whose outcome is death than in survivors, and so is the peak level of D-dimer [27]. Thus, in the presence of a suspected case of COVID-19 with a negative RT-PCR, an altered D-dimer could be an interesting suggestion of a false-negative test.

Research using open data has some challenges. The FAPESP dataset exemplifies some of these issues. Data are not always provided in a standardized way, for example the labels and reference range can be different for the same laboratory tests. The authors of [9] discuss common problems with using open data sources in the context of COVID-19 research. Data mining techniques are necessarily limited due to inconsistencies and outliers present in routine clinical datasets [10] and the lack of standardization in the data collection process.

This study has limitations. There is not sufficient information on the bias of subjects that were tested since there is no additional information of symptoms. Since not all patients were tested for the same exams and the name of the exams are not standardized, the accuracy of the model was tested and is presented given the available dataset and according to the methods described above. The name of exams do not need to be standardized, since PCA only considers their variance, but the ideal scenario is for the subjects to be tested for the same set of exams with a standardized measure for their reference value and what is to be considered an abnormal exam.

VIII. Conclusion

This was an exploratory analysis of routinely collected clinical data from patients undergoing RT-PCR testing for SARS-CoV-2 at hospitals in Brazil. We confirm previous reports showing that an elevated d-dimer and LDH are associated with a positive COVID-19 diagnosis. Using a random forest algorithm we were able to classify patients with 76% accuracy. Abnormalities in these laboratory results should alert the clinician to the possibility of a false-negative RT-PCR result when the clinical suspicion is high for COVID-19. The analytic approach that we have presented, integrating expert opinion and machine learning techniques can be applied to other routine clinical data sources.

References

1. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. 2020. Available from: https://covid19.who.int/
2. Ulahannan JP, Narayanan N, Thalhath N, Prabhakaran P, Chaliyeduth S, Suresh SP, Mohammed M, Rajeevan E, Joseph S, Balakrishnan A, Uthaman J, Karingamadathil M, Thomas ST, Sureshkumar U, Balan S, and Vellichirammal NN. A citizen science initiative for open data and visualization of COVID-19 outbreak in Kerala, India. Journal of the American Medical Informatics Association 2020 Aug. DOI: 10.1093/jamia/ocaa203. Available from: https://doi.org/10.1093/jamia/ocaa203

3. São Paulo F de Amparo à Pesquisa do Estado de. FAPESP COVID-19 DataSharing/BR Repository. 2020. Available from: https://repositoriodatasharingfapesp.uspdigital.usp.br/
4. Mello LE, Suman A, Medeiros CB, Prado CA, Rizzatti EG, Nunes FLS, Barnabé GF, Ferreira JE, Sá J, Reis LFL, Rizzo LV, Sarno L, Lamonica R de, Maciel RMdB, Cesar-Jr RM, and Carvalho R. Opening Brazilian COVID-19 patient data to support world research on pandemics. en. 2020. DOI: 10.5281/ZENODO.3966427. Available from: https://zenodo.org/record/3966427
5. Carlotti D, Ferreira JE, and Nunes FLS. Relational data model and programs. 2020 (accessed October 15, 2020). Available from: http://repositorio.uspdigital.usp.br/handle/item/243
6. Kucirka LM, Lauer SA, Laeyendecker O, Boon D, and Lessler J. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure. Annals of Internal Medicine 2020 Aug; 173:262–7. DOI: 10.7326/m20-1495. Available from: https://doi.org/10.7326/m20-1495
7. Kasteren PB van, Veer B van der, Brink S van den, Wijsman L, Jonge J de, Brandt A van den, Molenkamp R, Reusken CB, and Meijer A. Comparison of seven commercial RT-PCR diagnostic kits for COVID-19. Journal of Clinical Virology 2020 Jul; 128:104412. DOI: 10.1016/j.jcv.2020.104412. Available from: https://doi.org/10.1016/j.jcv.2020.104412
8. Li Y, Yao L, Li J, Chen L, Song Y, Cai Z, and Yang C. Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. Journal of Medical Virology 2020 Apr; 92:903–8. DOI: 10.1002/jmv.25786. Available from: https://doi.org/10.1002/jmv.25786
9. Alamo T, Reina D, Mammarella M, and Abella A. Covid-19: Open-Data Resources for Monitoring, Modeling, and Forecasting the Epidemic. Electronics 2020 May; 9:827. DOI: 10.3390/electronics9050827. Available from: https://doi.org/10.3390/electronics9050827
10. Saire JEC. Data Mining Approach to Analyze Covid19 Dataset of Brazilian Patients. 2020 Aug. DOI: 10.1101/2020.08.13.20174508. Available from: https://doi.org/10.1101/2020.08.13.20174508
11. Caten F ten, Gonzalez-Dias P, Castro I, Ogava R, Giddaluru J, Silva JC, Martins F, Goncalves AN, Costa-Martins AG, Araujo JD, Viegas AC, Cunha FQ, Farsky S, Bozza FA, Levin AS, Pannaraj PS, Silva TI de, Minoprio P, Andrade BB, Silva FP da, and Nakaya HI. In-depth Analysis of Laboratory Parameters Reveals the Interplay Between Sex, Age and Systemic Inflammation in Individuals with COVID-19. 2020 Aug. DOI: 10.1101/2020.08.07.20170043. Available from: https://doi.org/10.1101/2020.08.07.20170043
12. Trivizakis E, Tsiknakis N, Vassalou E, Papadakis G, Spandidos D, Sarigiannis D, Tsatsakis A, Papanikolaou N, Karantanas A, and Marias K. Advancing Covid-19 differentiation with a robust preprocessing and integration of multi-institutional open-repository computer tomography datasets for deep learning analysis. Experimental and Therapeutic Medicine 2020 Sep; 20:1–1. DOI: 10.3892/etm.2020.9210. Available from: https://doi.org/10.3892/etm.2020.9210
13. Ostaszewski M, Mazein A, Gillespie ME, Kuperstein I, Niarakis A, Hermjakob H, Pico AR, Willighagen EL, Evelo CT, Hasenauer J, Schreiber F, Dräger A, Demir E, Wolkenhauer O, Furlong LI, Barillot E, Dopazo J, Orta-Resendiz A, Messina F, Valencia A, Funahashi A, Kitano H, Auffray C, Balling R, and Schneider R. COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. Scientific Data 2020 May; 7. DOI: 10.1038/s41597-020-0477-8. Available from: https://doi.org/10.1038/s41597-020-0477-8
14. Wang L and Wong A. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. 2020. arXiv: 2003.09871 [eess.IV]
15. Nalleballe K, Onteddu SR, Sharma R, Dandu V, Brown A, Jasti M, Yadala S, Veerapaneni K, Siddamreddy S, Avula A, Kapoor N, Mudassar K, and Kovvuru S. Spectrum of neuropsychiatric manifestations in COVID-19. Brain, Behavior, and Immunity 2020 Aug; 88:71–4. DOI: 10.1016/j.bbi.2020.06.020. Available from: https://doi.org/10.1016/j.bbi.2020.06.020
16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12:2825–30
17. Fonti V and Belitser E. Paper in Business Analytics Feature Selection using LASSO. 2017
18. Ferrari D, Motta A, Strollo M, Banfi G, and Locatelli M. Routine blood tests as a potential diagnostic tool for COVID-19. Clinical Chemistry and Laboratory Medicine (CCLM) 2020 Jun; 58:1095–9. DOI: 10.1515/cclm-2020-0398. Available from: https://doi.org/10.1515/cclm-2020-0398
19. Esteves F, Calé S, Badura R, Boer M de, Maltez F, Calderón E, Reijden T van der, Márquez-Martín E, Antunes F, and Matos O. Diagnosis of Pneumocystis pneumonia: evaluation of four serologic biomarkers. Clinical Microbiology and Infection 2015 Apr; 21:379.e1–379.e10. DOI: 10.1016/j.cmi.2014.11.025. Available from: https://doi.org/10.1016/j.cmi.2014.11.025
20. Li C, Hu B, Zhang Z, Qin W, Zhu Z, Zhai Z, Davidson BL, and Wang C. D-dimer Triage for COVID-19. Academic Emergency Medicine 2020 Jun; 27. Ed. by Kline JA:612–3. DOI: 10.1111/acem.14037. Available from: https://doi.org/10.1111/acem.14037
21. Yin S, Huang M, Li D, and Tang N. Difference of coagulation features between severe pneumonia induced by SARS-CoV2 and non-SARS-CoV2. Journal of Thrombosis and Thrombolysis 2020 Apr. DOI: 10.1007/s11239-020-02105-8. Available from: https://doi.org/10.1007/s11239-020-02105-8

22. Shi J, Li Y, Zhou X, Zhang Q, Ye X, Wu Z, Jiang X, Yu H, Shao L, Ai JW, Zhang H, Xu B, Sun F, and Zhang W. Lactate dehydrogenase and susceptibility to deterioration of mild COVID-19 patients: a multicenter nested case-control study. BMC Medicine 2020 Jun; 18. DOI: 10.1186/s12916-020-01633-7. Available from: https://doi.org/10.1186/s12916-020-01633-7
23. Al-Ani F, Chehade S, and Lazo-Langner A. Thrombosis risk associated with COVID-19 infection. A scoping review. Thrombosis Research 2020 Aug; 192:152–60. DOI: 10.1016/j.thromres.2020.05.039. Available from: https://doi.org/10.1016/j.thromres.2020.05.039
24. Cummings MJ, Baldwin MR, Abrams D, Jacobson SD, Meyer BJ, Balough EM, Aaron JG, Claassen J, Rabbani LE, Hastie J, Hochman BR, Salazar-Schicchi J, Yip NH, Brodie D, and O'Donnell MR. Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. The Lancet 2020 Jun; 395:1763–70. DOI: 10.1016/s0140-6736(20)31189-2. Available from: https://doi.org/10.1016/s0140-6736(20)31189-2
25. Gao Y, Li T, Han M, Li X, Wu D, Xu Y, Zhu Y, Liu Y, Wang X, and Wang L. Diagnostic utility of clinical laboratory data determinations for patients with the severe COVID-19. Journal of Medical Virology 2020 Apr; 92:791–6. DOI: 10.1002/jmv.25770. Available from: https://doi.org/10.1002/jmv.25770
26. Velavan TP and Meyer CG. Mild versus severe COVID-19: Laboratory markers. International Journal of Infectious Diseases 2020 Jun; 95:304–7. DOI: 10.1016/j.ijid.2020.04.061. Available from: https://doi.org/10.1016/j.ijid.2020.04.061
27. Ye W, Chen G, Li X, Lan X, Ji C, Hou M, Zhang D, Zeng G, Wang Y, Xu C, Lu W, Cui R, Cai Y, Huang H, and Yang L. Dynamic changes of D-dimer and neutrophil-lymphocyte count ratio as prognostic biomarkers in COVID-19. Respiratory Research 2020 Jul; 21. DOI: 10.1186/s12931-020-01428-7. Available from: https://doi.org/10.1186/s12931-020-01428-7
28. Kermali M, Khalsa RK, Pillai K, Ismail Z, and Harky A. The role of biomarkers in diagnosis of COVID-19 – A systematic review. Life Sciences 2020 Aug; 254:117788. DOI: 10.1016/j.lfs.2020.117788. Available from: https://doi.org/10.1016/j.lfs.2020.117788
29. Carpenter CR, Mudd PA, West CP, Wilber E, and Wilber ST. Diagnosing COVID-19 in the Emergency Department: A Scoping Review of Clinical Examinations, Laboratory Tests, Imaging Accuracy, and Biases. Academic Emergency Medicine 2020 Jul; 27. Ed. by Zehtabchi S:653–70. DOI: 10.1111/acem.14048. Available from: https://doi.org/10.1111/acem.14048

Figures

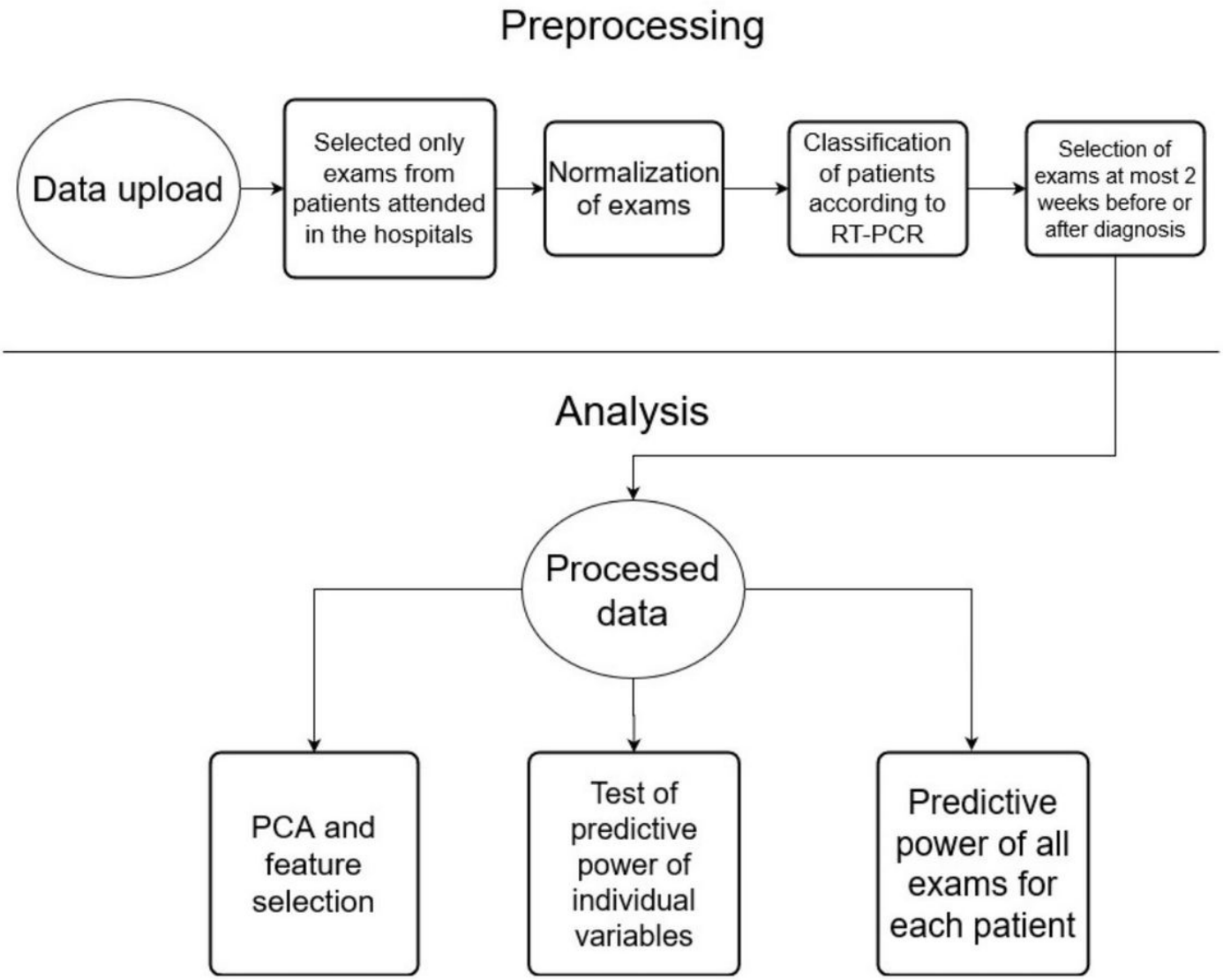


Figure 1

General view of the approach used to analyze the data.

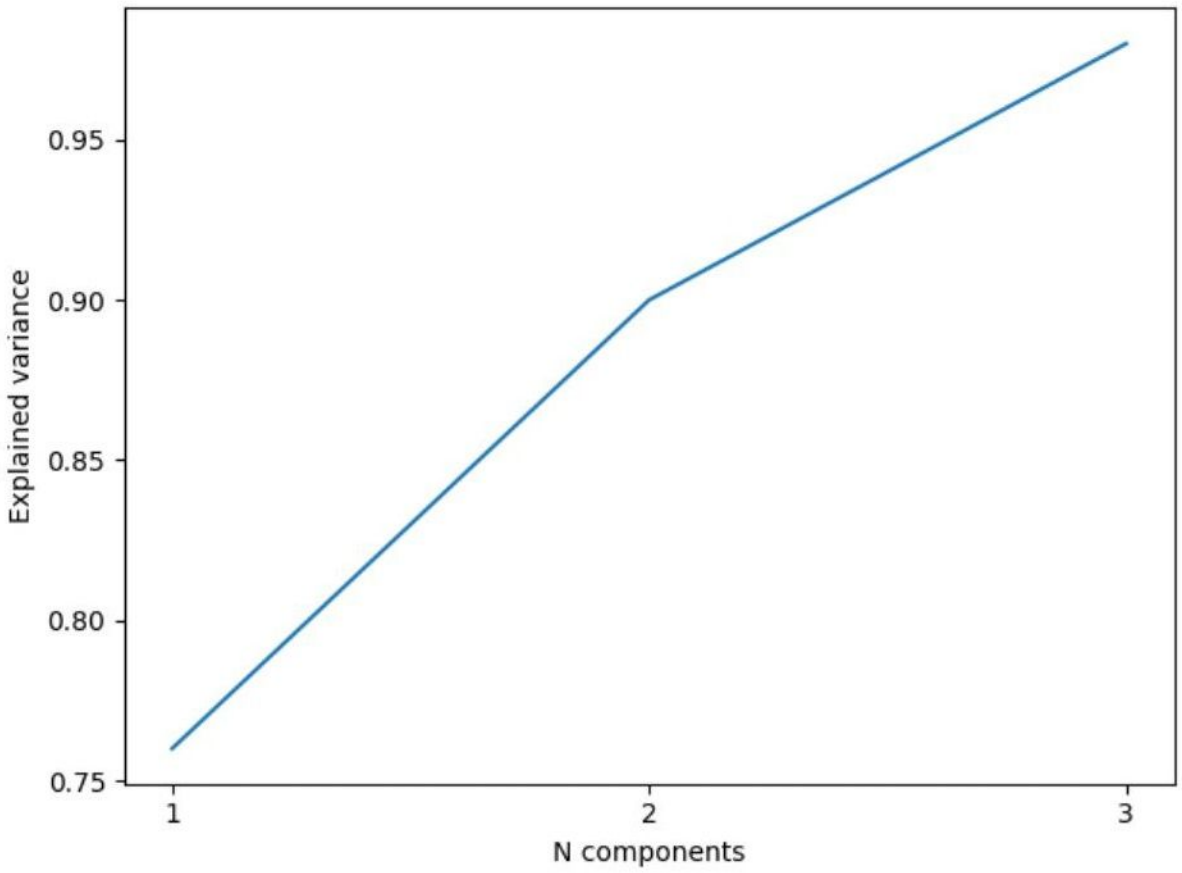


Figure 2

The variance explained by PCA components was 1 (76%), 2 (90%) and 3 (96%)

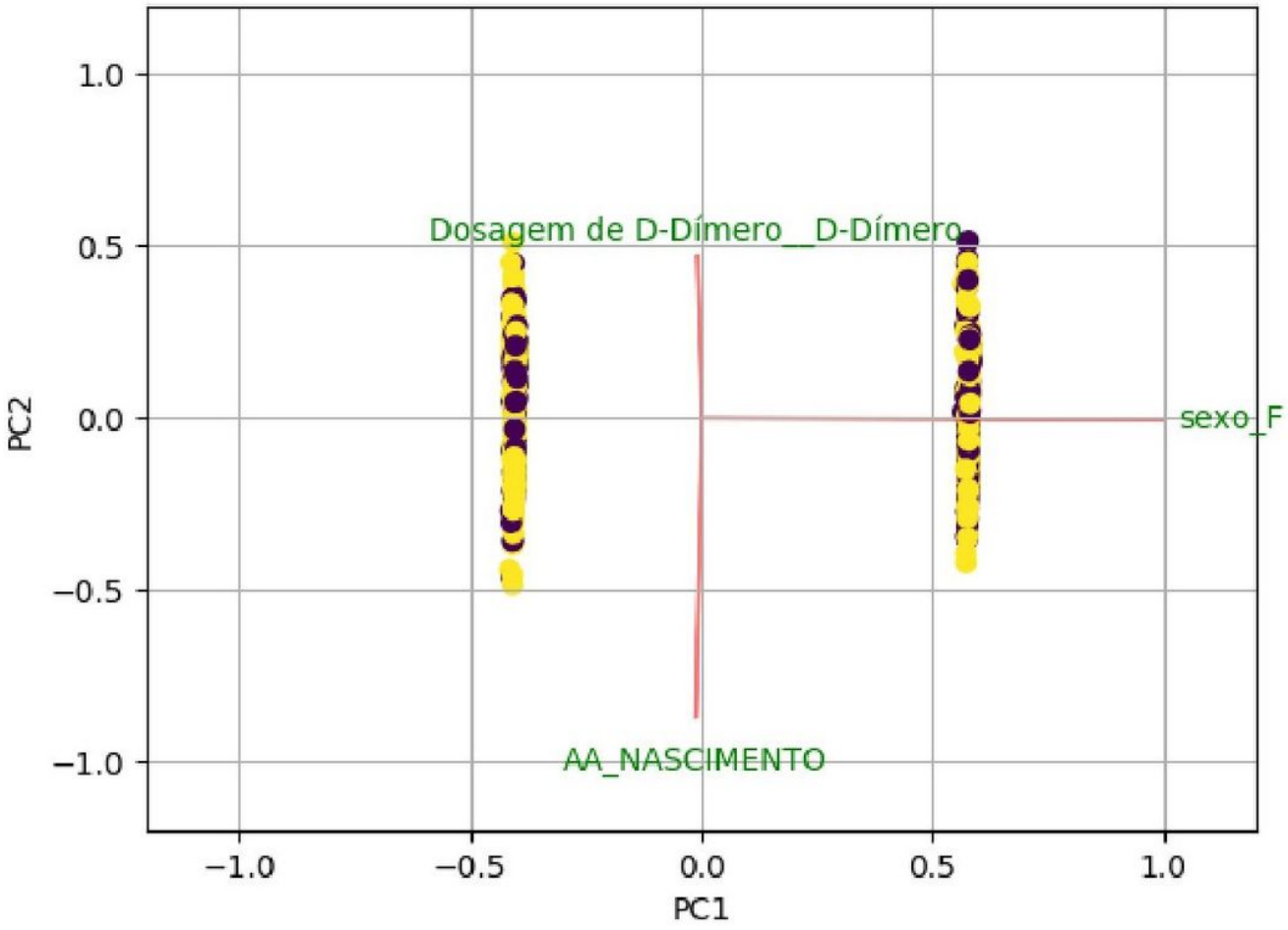
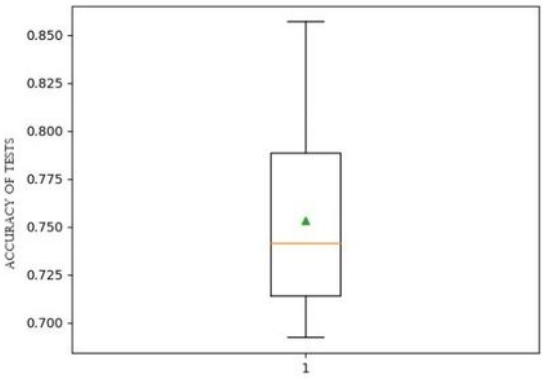
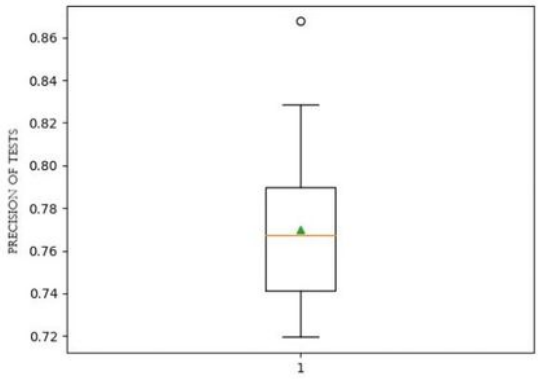


Figure 3

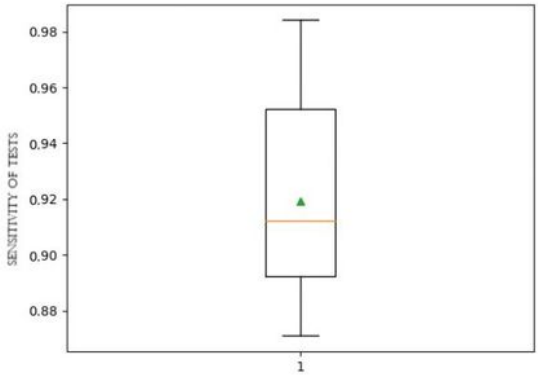
Biplot of the Principal Component Analysis - main variables



(a) Accuracy



(b) Precision



(c) Sensitivity

Figure 4

Results using Random Forest algorithm.

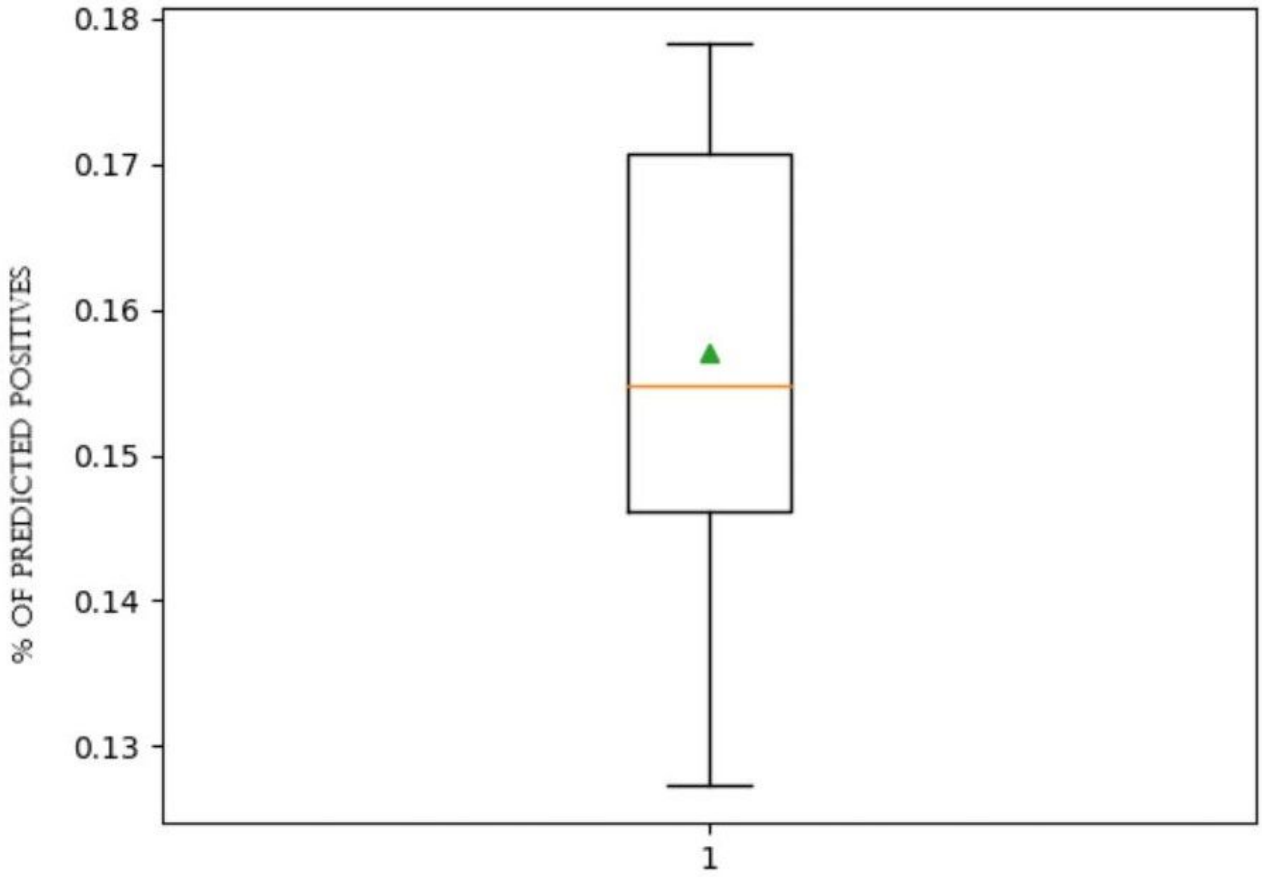


Figure 5

Percentage of predicted positives according to RF