Article

Journal of the Brazilian Chemical Society

https://dx.doi.org/10.21577/0103-5053.20240020

*J. Braz. Chem. Soc.* **2024**, *35*, 7, e-20240020, 1-14


# Use of Biochemical Tests and Machine Learning in the Search for Potential Diagnostic Biomarkers of COVID-19, HIV/AIDS, and Pulmonary Tuberculosis

***Alexandre F. Cobre,[a] Amiel A. Morais,[b] Fosfato Selege,[b] Dile P. Stremel,[c] Astrid Wiens,[a] Luana M. Ferreira,[d] Fernanda S. Tonin[e] and Roberto Pontarolo*** *,[a,d]

[a]*Programa de Pós-Graduação em Ciências Farmacêuticas, Universidade Federal do Paraná, 80210-170 Curitiba-PR, Brazil*

[b]*Pharmaceutical Sciences Undergraduate Program, Universidade Lúrio, 3100, Nampula, Mozambique*

[c]*Departamento de Engenharia e Tecnologia, Universidade Federal do Paraná, 80210-170 Curitiba-PR, Brazil*

[d]*Departamento de Famácia, Universidade Federal do Paraná, 80210-170 Curitiba-PR, Brazil*

[e]*Health & Technology Research Centre, ESTeSL, Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, 1990-096, Lisbon, Portugal*

This study aims to develop, validate, and evaluate machine learning algorithms for predicting the diagnosis of coronavirus disease (COVID-19), human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS), pulmonary tuberculosis (TB), and HIV/TB co-infection. We also investigated potential biomarkers associated with the diagnosis. Data from biochemical and hematological tests of infected and controls were collected in a single general hospital, totalizing 6,418 patients. The discriminant analysis by partial least squares (PLS-DA) model had the highest performance in predicting the diagnosis of COVID-19, HIV/AIDS, TB, and HIV/TB co-infection with an accuracy of 94, 97, 95, and 96%, respectively. The biomarkers calcium, lactate dehydrogenase, red blood cells (RBC), white blood cells, neutrophils, basophils, eosinophils, hemoglobin, and hematocrit were associated with COVID-19. HIV infection was associated with mean corpuscular volume, platelets, neutrophils, and mean platelet volume. Red blood cell distribution width and urea were associated with infection by *Mycobacterium tuberculosis*. The following biomarkers were associated with HIV/TB co-infection: lymphocytes, RBC, hematocrit, hemoglobin, aspartate transaminase, alanine transaminase, and glycemia. The PLS-DA model can optimize COVID-19, HIV/AIDS, TB, and HIV/TB co-infection diagnostics. Some biomarkers were potential diagnostic indicators and could be evaluated during the screening of these diseases.

**Keywords:** COVID-19, HIV/AIDS, tuberculosis, co-infection, diagnosis, machine learning

## Introduction

Accessibility to accurate and reliable diagnoses is a major challenge for public health systems. In recent years, powerful tools such as artificial intelligence (AI) and machine learning (ML) have been considered promising solutions to complex problems in many domains. Thus, machine learning-assisted diagnosis revolutionizes healthcare, optimizing early disease detection by providing accurate and personalized diagnoses based on patient data. The data needs to be robust and sufficiently vast for an ML model to be successful and generalizable to new cases. The utility of ML in diagnosis is not only beneficial in areas with limited resources but also presents universal opportunities for health.[1,2]

Access to diagnostic tests for coronavirus disease (COVID-19) and tuberculosis (TB) is a key step in the cascade of care to minimize the transmission of both diseases and allow for the proper management of affected patients. Thus, it is essential to improve access to testing for COVID-19 and tuberculosis by implementing

*e-mail: pontarolo@ufpr.br
Editor handled this article: Ivo M. Raimundo Jr. (Associate)

concurrent testing, particularly in countries with a high burden of disease, reducing the impact of the pandemic on tuberculosis services and identifying people at high risk for both illnesses.

With the COVID-19 pandemic and technological advances, several medical approaches have become available for diagnosis. However, considering that early detection, diagnosis, isolation, and treatment are critical to prevent the further spread of the disease, it is critical to develop accurate and cost-effective detection methods from which all countries can benefit.[3,4]

Thus, none of the tests currently available has a completely satisfactory performance (high diagnostic accuracy), and the search for fast, low-cost tests with adequate sensitivity is essential. In the literature, several methods have been proposed for the diagnosis of COVID-19 using ML and Deep Learning approaches by researchers.[5-8] Some of these studies were previously published by our group.[9,10]

According to the World Health Organization (WHO),[11] the long COVID is the current challenge in combating the COVID-19 pandemic. In a recent study conducted by the Fundação Oswaldo Cruz (FIOCRUZ) group (Brazil) involving 646 patients with COVID-19 followed for 14 months, half of the patients developed long COVID, with mental disorders being one of the most prevalent symptoms.[12] Long COVID is defined as the continuation or development of new symptoms 3 months after recovery from the initial severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) infection, with these symptoms lasting for at least 2 months with no other explanation.[11] Considering the current scenario of the pandemic (long COVID) and the high price of performing the real-time reverse transcription polymerase chain reaction (RT-PCR) test for COVID-19, studies that aim at the development of new, low-cost, accessible, and rapid methods for the diagnosis of COVID-19 are extremely important, because they allow greater coverage in access to diagnosis by the population, rapid treatment of patients, avoiding the development of COVID long.[3,13-16] In addition, the lack of human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) and TB diagnostic tests in low-income countries (e.g., Sub-Saharan Africa), may be compounded by the recent discovery of the most virulent variant of the HIV (HIV-1 subtype 1), which is rapidly spreading across Europe at a rate five times faster than the previous strains of HIV/AIDS. Given the above, in this study, we propose an alternative method for diagnosing COVID-19, pulmonary TB, HIV/AIDS, and HIV/TB co-infection using patient biochemical testing and machine learning. We also investigated potential biomarkers associated with the diagnosis of these diseases.

## Experimental

### Study design and patients

The retrospective datasets of patients with COVID-19, HIV/AIDS, pulmonary tuberculosis, and HIV/TB co-infected patients were obtained from Hospital Geral do Marrere, Mozambique (Universidade Lúrio), a university and reference hospital for the COVID-19, HIV/AIDS, and TB management from Mozambique.

### Obtaining data and ethical considerations

The study observed the guidelines of the Declaration of Helsinki of the World Medical Association. This study was previously approved by the research ethics committee of the Hospital Geral de Marrere (Universidade Lúrio, Mozambique) No. 36.1/Abril/CBISUL/21. As this was research with secondary data, the aforementioned institution's ethics committee waived the signing of the Free and Informed Consent.

### Dataset I: COVID-19 patients

Regardless of the COVID-19 test (negative or positive RT-PCR), patients who presented data from hematological and biochemical tests were included (Table 1). Thus, the dataset was divided according to the RT-PCR result in (*i*) COVID-19 positive patients (n = 816 samples); and COVID-19 negative patients (control, n = 920 samples). The data were collected between April-November 2021.

### Dataset II: HIV/AIDS, TB, and HIV/TB patients

Regardless of the HIV/AIDS or TB test result (negative or positive), patients with hematological and biochemical test results were included in the study (Table 1). Thus, the dataset was organized into four groups: (*i*) patients with HIV/AIDS (n = 49); (*ii*) patients with pulmonary TB (n = 113); (*iii*) patients co-infected with HIV/TB (n = 80); (*iv*) patients who tested negative for HIV/AIDS or TB (control, n = 4,520). The data were also collected between April-November 2021.

### Characterization of the control group

The control group used in this study was a subdivision of the biochemical dataset, which comprised patients treated at

**Table 1.** Biochemical and hematological tests performed on the patients included in the study

| Test name | Measurement unit |
|---|---|
| White blood cells | ($\times 10^9$ L$^{-1}$) |
| Red blood cells | ($\times 10^{12}$ L$^{-1}$) |
| Hemoglobin | (g dL$^{-1}$) |
| Hematocrit | % |
| Mean corpuscular volume | fL |
| Mean corpuscular hemoglobin | (g cell$^{-1}$) |
| Mean corpuscular hemoglobin | (g Hb dL$^{-1}$) |
| Platelets count | ($\times 10^9$ L$^{-1}$) |
| Platelets count | % |
| Platelets count | ($\times 10^9$ L$^{-1}$) |
| Neutrophils count | % |
| Neutrophils count | ($\times 10^9$ L$^{-1}$) |
| Monocytes count | % |
| Monocytes count | ($\times 10^9$ L$^{-1}$) |
| Creatinine Kinase | (U L$^{-1}$) |
| Urea | (mg dL$^{-1}$) |
| Aspartate aminotransferase | (U L$^{-1}$) |
| Alanine aminotransferase | (U L$^{-1}$) |
| Glucose | (mg dL$^{-1}$) |
| Calcium | (mmol L$^{-1}$) |
| Sodium | (mmol L$^{-1}$) |
| Potassium | (mmol L$^{-1}$) |

the hospital who suffered from chronic non-communicable diseases, the most frequent being diabetes, obesity, and high blood pressure. It is essential to highlight that all patients in the control group were negative for HIV/AIDS, *Mycobacterium tuberculosis*, and SARS-CoV-2 infections.

### Combination of datasets I and II: descriptive analysis

For the development of machine learning models, datasets I and II were analyzed separately. However, a general descriptive analysis of the data (data expressed as a median and interquartile range) was performed by combining datasets I and II into a single final dataset. In this combination of datasets, only common biomarkers in both datasets were kept in the final database.

### Dataset III: external validation samples of machine learning models

The Brazilian dataset was used for external validation of the machine learning model. That is, data from Mozambique were used for the training and validation of the model. However, the Brazil dataset was used to test whether the model could predict patient samples from an external source (external model validation).

Data from patients in Brazil used in this study were obtained from the Brazilian public data repository called "FAPESP COVID-19 DataSharing/BR",[17] which is a large repository of data from COVID-19 patients in the state of São Paulo (Brazil). This repository contains demographic data, data from clinical and/or laboratory examinations of COVID-19 patients and controls from the five main hospitals in the State of São Paulo (Brazil), namely: (*i*) Hospital Israelita Albert Einstein; (*ii*) Hospital de Clínicas of the Faculty of Medicine of the University of São Paulo; (*iii*) Hospital Sírio-Libanês and (*iv*) Beneficência Portuguesa de São Paulo. Most of these patients with COVID-19 have several comorbidities such as diabetes, high blood pressure, and hypothyroidism.

The FAPESP COVID-19 DataSharing/BR repository[17] and additional information about the repository are available in the literature.[18] Thus, considering that the data used in this study are open access, the bioethics committee approval was not necessary.

### Univariate analysis

Before developing the machine learning models (multivariate analysis), we performed univariate analysis to compare the levels of biochemical and hematological biomarkers in the four study patient groups (COVID-19, HIV/AIDS, HIV/AIDS/TB, and control group). In this analysis, the normal distribution of each biochemical and hematological biomarker (variable) was initially tested using the Shapiro-Wilk test. After that, comparisons of the biomarkers of the four groups of patients (COVID-19, HIV/AIDS, HIV/TB, and control group) were carried out using the Kruskal Wallis non-parametric test, when the biomarkers did not present a normal distribution (Kruskal Wallis test, $p < 0.05$). On the other hand, when the biomarkers had a normal distribution (Kruskal Wallis test, $p > 0.05$), comparisons between the four groups were performed using the one-way-analysis of variance (ANOVA) test. It is also important to highlight that biomarkers that did not present a normal distribution were reported as median and interquartile range (IQR), and those that presented a normal distribution were reported as mean ± standard deviation. Significance levels $p < 0.05$ were considered statistically significant and all univariate analyzes were performed using SPPS software version 2020 (IBM, USA).[19]

### Data pre-processing in machine learning

Pre-processing plays a vital role in preparing your data for machine learning models. It can lead to improved model performance, better generalization, and more meaningful

insights from data, ultimately increasing the chances of successfully solving a particular problem or task.[10] For our study, the following five different preprocessing methods were tested on datasets (HIV/AIDS, TB, and COVID-19 diagnostic data) to select which preprocessing best fits the data: (*i*) scaling and centering: autoscale, group scale, log decay scaling, mean center, median center, multiway center, multiway scale and square root (sqrt) mean scale; (*ii*) transformations: absolute value, log10; (*iii*) normalization: normalize, standard normal variate (SNV) and multiplicative scatter correction (MSC-mean); (*iv*) imputation: missing data were replaced by the median values; (*v*) imputation: missing data were replaced by the median values; (*vi*) filtering: baseline (specified points), baseline (weighted least square), derivative (Savitzky-Golay), smoothing (Savitzky-Golay), detrend, generalized least squares weighting (GLSW), orthogonal signal correction (OSC) and external parameter orthogonalization (EPO).

### Machine learning models

All machine learning analysis were performed in the SOLO Software (Eigenvector Research, USA).[20] The ML models were developed for each disease class (COVID-19, HIV/AIDS, TB, and HIV/TB) *versus* the control group. To acquire an ML model, exploratory data analysis was initially performed to recognize data standards, select important variables, and detect outliers.[21] In this study, the exploratory analysis of the data was conducted using the principal component analysis (PCA) model, and the detection of outlier samples was performed using the leverage plot *versus* student residuals (samples with high leverage and residual student values were identified, and removed from the dataset).[22]

Afterward, a total of seven ML models (partial least squares discriminant analysis (PLS-DA), artificial neural network (ANN), eXtreme Gradient Boosting (XGBoosted), K-Nearest Neighbors (KNN), logistic regression (LREG), Soft independent modeling by class analogy (SIMCA) and Support vector machine (SVM)) were trained, validated, and evaluated to predict the diagnosis of COVID-19, HIV/AIDS, TB, or HIV/TB co-infection. To ensure meaningful representation of training and testing data, each of the four patient classes (COVID-19, HIV/AIDS, TB, and HIV/TB) were individually analyzed, where 70% of the data was used for calibration and 30% were used for validation, as shown in Table 2. It is important to highlight that this proportion of division of training and test data was carried out based on studies previously published in scientific literature.[10] Samples were randomly selected using the Kennard-Stone algorithm.[23] In Table 2, the unbalanced data observed between the patients, the patient group (COVID-19, HIV/AIDS, TB, and HIV/TB) and the control group can be explained due to the prevalence and incidence rates of the diseases. For example, the prevalence of HIV/AIDS in Mozambique is 13.5, while the incidence of tuberculosis is 551 cases *per* 100 thousand people.[24,25] All ML models were optimized considering lower values of root-mean-square error of cross-validation (RMSECV).[26]

The model validation (performance) was assessed considering specificity (the ability of an ML model to correctly classify true negative samples), sensitivity (the ability of an ML model to correctly classify true positive samples), and accuracy (established to Receiver Operator Characteristic (ROC) curves is the ability of an ML model to correctly classify negative and positive samples and represent the model efficiency).[27-29] Equations 1, 2, 3, 4, 5, and 6 were used to calculate sensitivity (recall), specificity, accuracy, precision, *F*1 score, and Mattew's correlation coefficient, respectively.

The Variable Importance in Projection (VIP) graph was constructed for the model that had high diagnostic accuracy, to investigate potential biomarkers associated with the diagnosis of COVID-19, HIV/AIDS, TB, and HIV/TB co-infection.

$$\text{Sensitivity (recall)} = \frac{TP}{TP + FN} \quad (1)$$

**Table 2.** Subset of calibration and validation data used for the development of machine learning models to predict diagnosis of COVID-19, HIV/AIDS, TB, and HIV/TB patients

| Dataset | Training (70%) | | | | Testing | |
|---|---|---|---|---|---|---|
| | Positive | Negative (control) | Positive | Negative | Positive | Negative |
| COVID-19 | 816 | 920 | 571 | 644 | 245 | 276 |
| HIV/AIDS | 49 | 4520 | 34 | 3164 | 15 | 1356 |
| TB | 113 | 4520 | 79 | 3164 | 34 | 1356 |
| HIV/TB | 80 | 4520 | 56 | 3164 | 24 | 1356 |

COVID-19: coronavirus disease; HIV/AIDS: human immunodeficiency virus; TB: tuberculosis.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$F1\ \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \quad (6)$$

where TP: true positive, TN: true negative, FP: false positive, FN: false negative, and MCC: Mattew's correlation coefficient.

## Results

### Univariate analysis

The results of the univariate analysis comparing the levels of all biochemical and hematological biomarkers between COVID-19, HIV/AIDS, pulmonary TB and HIV/TB co-infection patients are presented in Table 3. The Kruskal Wallis test showed that there were significant differences ($p < 0.05$) in the viability of all these biomarkers in the four groups studied.

### Multivariate analysis: principal component analysis (PCA)

The results of the exploratory analysis (PCA model) of COVID-19 diagnostic data are shown in Figure 1, while exploratory analysis for HIV/AIDS, pulmonary TB, and HIV/TB co-infection patient data are shown in Figure 2. In PCA analysis, a sample is considered a potential outlier

**Table 3.** Biochemical and hematological biomarkers used in patients with COVID-19, HIV/AIDS, tuberculosis, and HIV/TB co-infection

| Variable | COVID-19 | | HIV/AIDS | | Tuberculosis | | HIV/TB | | Control | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Median | IQR[a] | Median | IQR[a] | Median | IQR[a] | Median | IQR[a] | Median | IQR[a] | P[b] |
| Age / years | 62.00 | 51.00-75.00 | 38.00 | 32.00-43.00 | 32.00 | 24.00-40.00 | 33.50 | 28.25-42.75 | 64.00 | 45.00-78.00 | 0.00 |
| White blood cells / ($\times 10^9$ L$^{-1}$) | 6.58 | 5.01-9.58 | 4.50 | 3.70-5.60 | 7.30 | 5.15-9.75 | 7.15 | 4.83-9.98 | 8.80 | 6.75-11.31 | 0.00 |
| Red blood cells / ($\times 10^{12}$ L$^{-1}$) | 4.67 | 4.27-5.07 | 3.93 | 3.46-4.75 | 4.48 | 3.99-4.94 | 3.73 | 3.27-4.33 | 4.50 | 4.03-4.89 | 0.00 |
| Hemoglobin / (g dL$^{-1}$) | 13.70 | 12.31-14.74 | 12.40 | 11.25-13.85 | 11.20 | 9.85-13.15 | 10.00 | 8.53-11.43 | 13.10 | 11.53-14.30 | 0.00 |
| Hematocrit / % | 40.50 | 37.00-43.55 | 38.10 | 34.65-40.90 | 35.00 | 31.35-39.60 | 30.30 | 26.55-34.40 | 39.03 | 35.30-42.20 | 0.00 |
| Mean corpuscular volume / fL | 87.10 | 84.00-90.10 | 98.90 | 76.95-107.85 | 73.00 | -70.30-83.95 | 71.05 | -73.88-87.78 | 87.70 | 84.26-91.24 | 0.00 |
| Mean corpuscular hemoglobin / (g cell$^{-1}$) | 29.40 | 28.21-30.50 | 31.20 | 27.90-34.45 | 25.20 | -19.30-27.75 | 26.50 | -17.75-29.05 | 29.50 | 28.10-30.70 | 0.00 |
| Mean corpuscular hemoglobin / (g Hb dL$^{-1}$) | 33.70 | 32.90-34.35 | 31.90 | 29.70-34.55 | 31.90 | 30.00-33.25 | 32.05 | 28.98-33.80 | 33.50 | 32.56-34.20 | 0.00 |
| Platelets count / ($\times 10^9$ L$^{-1}$) | 209.00 | 162.00-268.75 | 212.00 | 159.50-268.50 | 319.50 | 226.50-396.00 | 305.50 | 226.00-389.75 | 234.67 | 187.00 -284.00 | 0.00 |
| Platelets count / % | 15.65 | 10.16-22.69 | 41.10 | 27.85-51.70 | 27.20 | 21.00-37.90 | 24.20 | 14.80-37.75 | 18.40 | 14.18-22.60 | 0.00 |
| Platelets count / ($\times 10^9$ L$^{-1}$) | 1.00 | 0.71-1.30 | 2.20 | 1.75-2.95 | 2.00 | 1.50-2.65 | 1.60 | 1.10-2.38 | 1.50 | 1.20-1.75 | 0.00 |
| Neutrophils count / % | 75.60 | 66.83-83.50 | 31.80 | 20.65-47.35 | 58.50 | 52.45-66.60 | 57.35 | 47.15-76.75 | 70.40 | 66.40-76.10 | 0.00 |
| Neutrophils count / ($\times 10^9$ L$^{-1}$) | 4.80 | 3.50-7.25 | 1.80 | 1.34-2.35 | 4.40 | 3.06-6.70 | 4.30 | 3.03-6.28 | 5.95 | 5.00-7.10 | 0.00 |
| Monocytes count / % | 7.30 | 5.20-9.50 | 10.70 | 8.00-15.15 | 10.30 | 8.30-12.80 | 9.15 | 7.08-11.88 | 7.78 | 6.80-8.70 | 0.00 |
| Monocytes count / ($\times 10^9$ L$^{-1}$) | 0.50 | 0.30-0.60 | 0.50 | 0.35-0.68 | 0.80 | 0.60-1.10 | 0.80 | 0.50-0.90 | 0.60 | 0.56-0.75 | 0.00 |
| Creatinine kinase / (U L$^{-1}$) | 100.00 | 100.00-100.00 | 83.85 | 76.15-91.35 | 61.83 | 61.83-61.83 | 63.41 | 63.41-63.41 | 78.00 | 78.00-78.00 | 0.00 |
| Urea / (mg dL$^{-1}$) | 34.83 | 30.00-42.00 | 3.55 | 3.10-3.80 | 3.40 | 3.40-3.40 | 4.74 | 4.74-4.74 | 35.00 | 33.00-38.30 | 0.00 |
| Aspartate aminotransferase / (U L$^{-1}$) | 43.00 | 30.00-63.00 | 31.00 | 26.50-34.50 | 29.30 | 29.30-29.30 | 23.40 | 23.40-23.40 | 26.00 | 21.00-33.00 | 0.00 |
| Alanine aminotransferase / (U L$^{-1}$) | 33.85 | 23.00-54.46 | 31.00 | 27.00-38.35 | 21.00 | 21.00-21.00 | 25.20 | 25.20-25.20 | 22.00 | 16.31-31.00 | 0.00 |
| Glucose / (mg dL$^{-1}$) | 107.00 | 95.00-129.00 | 4.41 | 4.41-4.41 | 4.70 | 4.70-4.70 | 4.82 | 4.82-4.82 | 100.75 | 89.00-118.00 | 0.00 |

[a]IQR: interquaril range; [b]P: Kruskal Wallis test; COVID-19: coronavirus disease; TB: tuberculosis; HIV/AIDS: human immunodeficiency virus.

if, and only if, it simultaneously has high Q-residual values and high Hotteling $T^2$ values. For both data sets, although some samples have high Q-residual values, they cannot be considered outliers because they are within Hotteling $T^2$'s 95% confidence interval (Figures S1 and S2, in Supplementary Information (SI) section). It is important to highlight that both PCA models were optimized using the combination of the following process methods: normalize + autoscale. To discriminate patients from dataset 1 (COVID-19 *vs.* control) and dataset 2 (HIV/AIDS *vs.* TB *vs.* HIV/TB *vs.* control) by the PCA model, two principal components (90.42% of the variance explained) and three principal components (96.89% of the variance explained), respectively. From this analysis, no outlier samples were detected from COVID-19 diagnostic data (Figure S1 in SI section) and HIV/AIDS, pulmonary TB, and HIV/TB co-infection diagnostic data (Figure S2 in SI section).

In Figure 1a, the PCA model was able to discriminate the two sample classes (COVID-19 and control). According to the loading plot (Figure 1b), the biomarkers that were most important in discriminating between the COVID-19 group and the controls were: calcium, lactate dehydrogenase, C-reactive protein, white blood cells, red blood cells, lactate dehydrogenase, hemoglobin, hematocrit, lymphocytes, neutrophils, urea and aspartate transaminase. It is important to highlight that we only show the PC1 loading graph because it represents the majority of the explained variance of the dataset, which is 74.39%.

From Figure 1a, it can be seen that the n = 4520 samples from the control group are overlapping each other, making it difficult to visualize. To understand the source of this high overlap of these samples, we constructed a score plot of the PCA model using only the samples from the control group (n = 4520) applying the same combination of the same pre-processing methods (normalize + autoscale) used for building the PCA model shown in Figure 1a, and we observed high overlap remained (see Figure S3 in SI section). Next, we also built a new score plot of the PCA model of the same samples but without any pre-processing method and observed a high dispersion of all 4520 samples belonging to the control group (see Figure S4 in SI section). Based on these analyses, we conclude that the high overlap between the control samples in the control group in Figure 1a (n = 4520) is due to the combination of pre-processing methods used to optimize the PCA model, that is, the combination normalize + autoscale.

In Figure 2a, the PCA model was able to discriminate samples from patients with HIV/AIDS, pulmonary TB, HIV/TB co-infection, and control. According to the PC1 loadings plot (Figure 2b), the most important biomarkers in discriminating the four groups of patients (COVID-19, HIV/AIDS, HIV/TB and control) were red blood cells, hematocrit, mean corpuscular volume concentration, platelets, lymphocytes, neutrophils, monocytes, creatinine, alanine transaminase, red blood cell distribution width and serum glucose.

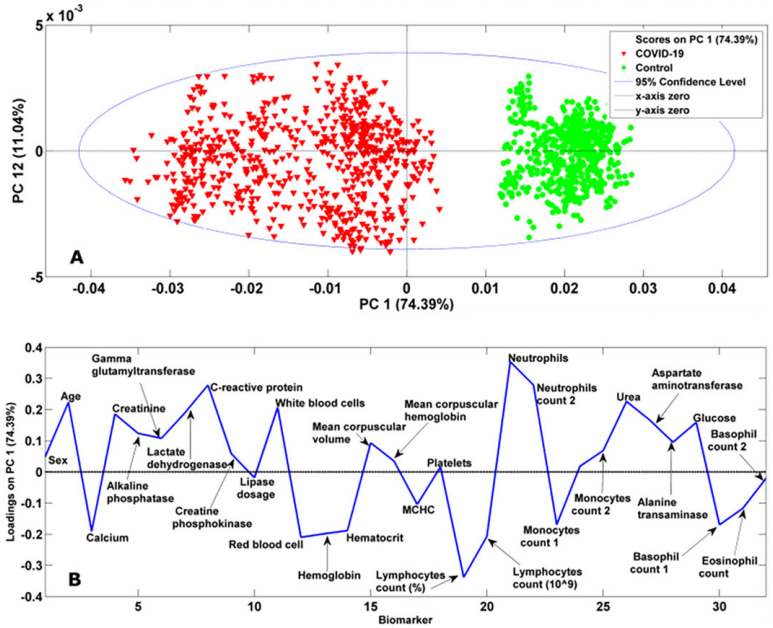


Figure 1. Exploratory analysis of the COVID-19 dataset. (a) PCA model of the blood samples from 816 patients with COVID-19 diagnosed by RT-PCR are represented by the red triangles and blood samples from 920 controls with negative RT-PCR are represented by green circles. (b) The loadings plot represents the most important variables in discriminating samples from the COVID-19 group and controls. Only the loadings plot of the first principal component is shown as it is the one that captured the highest explained variance of the original data, which was 74.39%.

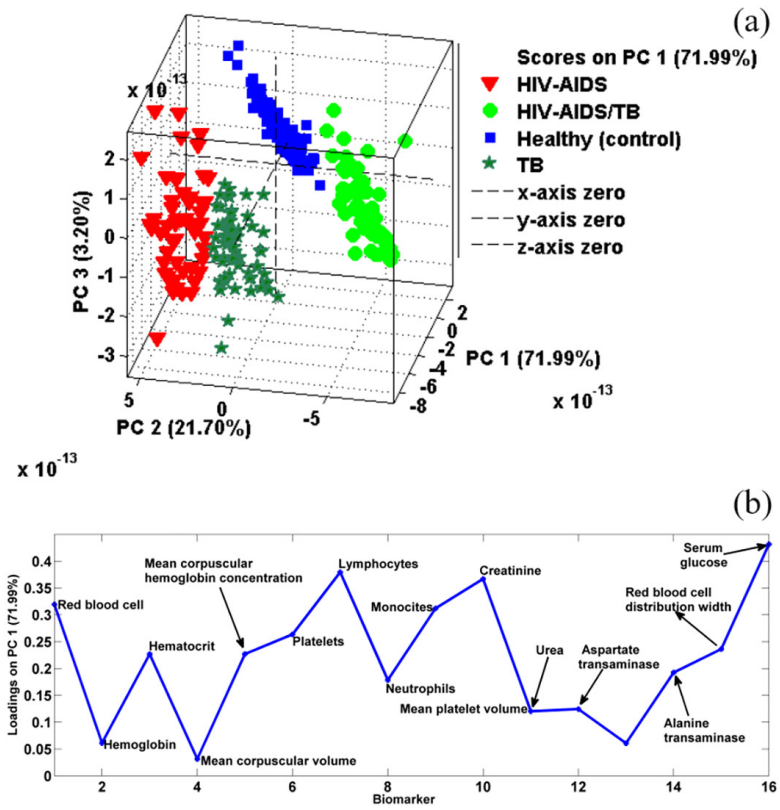


**Figure 2.** Exploratory analysis of the human immunodeficiency (HIV/AIDS), tuberculosis (TB), and HIV/TB co-infection dataset. In (a) it is shown the PCA model of the blood samples from 49 patients with HIV/AIDS are represented by the red triangles; 113 patients with pulmonary TB are represented by the green stars; 80 patients co-infected with HIV/TB are represented by the green circles and 4,520 controls who tested negative for HIV/AIDS or TB are represented by the blue squares (control). In (b) the loadings plot is shown that represents the most important variables in discriminating samples from the HIV/AIDS, TB, HIV/TB, and control. Only the loadings plot of the first principal component is shown as it is the one that captured the highest explained variance of the original data, which was 71.99%.

### Outliers detection

Before developing the classification models, graphs of leverage *versus* student residuals were constructed (Figures S5-S6, SI section). For both data set 1 (COVID-19 *vs*. control) and data set 2 (COVID-19 *vs*. HIV/AIDS *vs*. TB *vs*. HIV/TB *vs*. control), although there are some samples that presented high leverage values, none of them was considered outliers because all samples are within ± 2.5 standard deviations of the student residuals (Figures S5-S6, SI section). Therefore, all samples presented in the two datasets were used to develop machine learning models to predict the diagnosis of COVID-19, HIV/AIDS, TB and the detection of HIV/TB co-infection.

### Multivariate analysis: classification models

All machine learning models classifying different patient groups (COVID-19, HIV/AIDS, TB, HIV/TB and control) were trained and tested using the same processing methods used in the PCA analyzes mentioned above (normalize + autoscale). The performance results of all seven machine learning models (PLS-DA, ANN, XGBoosted, KNN, LREG, SIMCA, and SVM) used to predict the diagnosis of COVID-19, HIV/AIDS, pulmonary TB, and HIV/TB co-infection are presented in Table 4. The PLS-DA model was the one that showed the highest performance in predicting the diagnosis of COVID-19, HIV/AIDS, TB and HIV/TB due to its high sensitivity, specificity, accuracy, precision, *F*1 score and MCC values, varying between 88-96% (Table 4). It is important to highlight that the sensitivity (recall), specificity, accuracy, precision, *F*1 score, and Mattew's correlation coefficient (MCC) values were calculated using equations 1, 2, 3, 4, 5 and 6 shown in the Experimental section.

These models were trained considering lower values of RMSECV (Figures S7-S8, SI section). The PLS-DA model had the best performance (higher values of sensitivity, and specificity as shown in Table 4). The ROC curves of the PLS-DA diagnostic models of HIV/AIDS, pulmonary TB, and HIV/TB co-infection are presented in Figure 3.

Figures 4 and S9 (SI section) demonstrate the VIP graphs with the biomarkers that most contributed to the diagnosis. The *x* axis represents each biomarker used in the

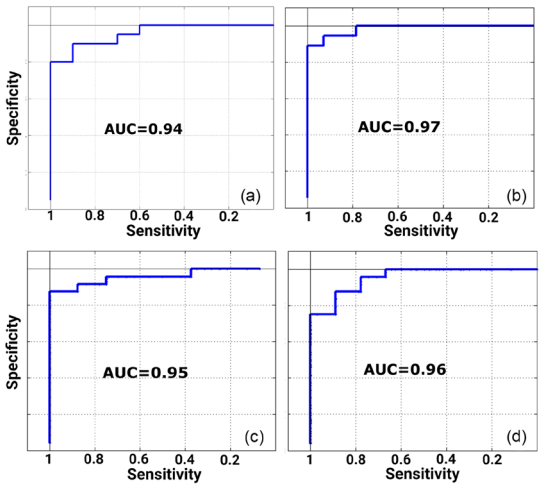


**Figure 3.** The area under the ROC curve of PLS-DA model performance. The area under the curves (AUC) reflects the accuracy of PLS-DA models in predicting patients of different classes of patients ((COVID-19, human immunodeficiency virus (HIV/AIDS), tuberculosis (TB) and (HIV/AIDS/TB) co-infection)) and control. Curves include both sets of samples (training and test samples). The AUC values for predicting COVID-19, HIV/AIDS, and HIV/TB co-infection were 94 (a), 97 (b), 95 (c), and 96% (d), respectively.

study. The *y* axis shows the VIP score that corresponds to the importance of each biomarker in predicting the diagnosis of COVID-19, HIV/AIDS and TB. VIP values greater than 1 (values above the dashed red horizontal line) are considered significantly important in predicting diagnosis. Thus, the following biomarkers were associated with the diagnosis of COVID-19: calcium, lactate dehydrogenase (LDH), white blood cell (WBC), red blood cell (RBC), hemoglobin, hematocrit, neutrophils count, aspartate aminotransferase, basophils count, and eosinophils count (Figure 4). The differences in biomarker levels between the COVID-19 patient group and the controls are shown in Table 3. It is important to highlight that these biomarkers were also identified by the PCA model as important in discriminating the two groups of patients (COVID-19 *vs.* control).

The parameters such as mean corpuscular volume (MCV), platelets, neutrophils, and mean platelet volume (MPV) were associated with HIV/AIDS infection. The platelets, neutrophils, red blood cell distribution width (RBCD), urea, and serum glucose were related

**Table 4.** Performance comparison of the machine learning models for COVID-19, HIV/AIDS, TB, and HIV/TB co-infection. Only the results of the model test are shown

| Class | Model | TP | FN | TN | FP | Sensitivity | Specificity | Accuracy | Precision | *F*1 score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | PLS-DA | 226 | 19 | 265 | 11 | 0.92 | 0.96 | 0.94 | 0.95 | 0.94 | 0.88 |
| | ANN | 224 | 21 | 248 | 28 | 0.91 | 0.90 | 0.91 | 0.89 | 0.90 | 0.81 |
| | KNN | 219 | 26 | 259 | 17 | 0.89 | 0.94 | 0.92 | 0.93 | 0.91 | 0.83 |
| COVID-19 | SVM | 205 | 40 | 259 | 17 | 0.84 | 0.94 | 0.89 | 0.92 | 0.88 | 0.78 |
| | SIMCA | 238 | 7 | 244 | 32 | 0.97 | 0.88 | 0.93 | 0.88 | 0.92 | 0.85 |
| | XGboost | 210 | 35 | 258 | 18 | 0.86 | 0.93 | 0.90 | 0.92 | 0.89 | 0.80 |
| | LREG | 213 | 32 | 264 | 12 | 0.87 | 0.96 | 0.92 | 0.95 | 0.91 | 0.83 |
| | PLS-DA | 15 | 0 | 1310 | 46 | 1.00 | 0.97 | 0.97 | 0.25 | 0.39 | 0.49 |
| | ANN | 12 | 3 | 1273 | 83 | 0.80 | 0.94 | 0.94 | 0.13 | 0.22 | 0.30 |
| | KNN | 12 | 3 | 1250 | 106 | 0.80 | 0.92 | 0.92 | 0.10 | 0.18 | 0.27 |
| HIV/AIDS | SVM | 11 | 4 | 1261 | 95 | 0.73 | 0.93 | 0.93 | 0.10 | 0.18 | 0.26 |
| | SIMCA | 13 | 2 | 1213 | 143 | 0.87 | 0.89 | 0.89 | 0.08 | 0.15 | 0.25 |
| | XGboost | 15 | | 1255 | 101 | 1.00 | 0.93 | 0.93 | 0.13 | 0.23 | 0.35 |
| | LREG | 11 | 4 | 1236 | 120 | 0.73 | 0.91 | 0.91 | 0.08 | 0.15 | 0.23 |
| | PLS-DA | 31 | 3 | 1251 | 105 | 0.91 | 0.92 | 0.92 | 0.23 | 0.36 | 0.43 |
| | ANN | 29 | 5 | 1247 | 109 | 0.85 | 0.92 | 0.92 | 0.21 | 0.34 | 0.40 |
| | KNN | 29 | 5 | 1247 | 109 | 0.85 | 0.92 | 0.92 | 0.21 | 0.34 | 0.40 |
| | SVM | 31 | 3 | 1220 | 136 | 0.91 | 0.90 | 0.90 | 0.19 | 0.31 | 0.39 |
| TB | SIMCA | 25 | 9 | 1207 | 149 | 0.74 | 0.89 | 0.89 | 0.14 | 0.24 | 0.29 |
| | XGboost | 29 | 5 | 1251 | 105 | 0.85 | 0.92 | 0.92 | 0.22 | 0.35 | 0.41 |
| | LREG | 24 | 10 | 1257 | 99 | 0.71 | 0.93 | 0.92 | 0.20 | 0.31 | 0.34 |
| | PLS-DA | 22 | 2 | 1343 | 13 | 0.92 | 0.99 | 0.99 | 0.63 | 0.75 | 0.75 |
| | ANN | 21 | 3 | 1255 | 101 | 0.88 | 0.93 | 0.92 | 0.17 | 0.29 | 0.37 |
| | KNN | 20 | 4 | 1238 | 118 | 0.83 | 0.91 | 0.91 | 0.14 | 0.25 | 0.33 |
| | SVM | 18 | 6 | 1243 | 113 | 0.75 | 0.92 | 0.91 | 0.14 | 0.23 | 0.30 |
| HIV/TB | SIMCA | 11 | 13 | 1214 | 142 | 0.46 | 0.90 | 0.89 | 0.07 | 0.12 | 0.15 |
| | XGboost | 20 | 4 | 1252 | 104 | 0.83 | 0.92 | 0.92 | 0.16 | 0.27 | 0.35 |
| | LREG | 19 | 5 | 1273 | 83 | 0.79 | 0.94 | 0.94 | 0.19 | 0.30 | 0.36 |

PLS-DA: partial least squares discriminant analysis; ANN: artificial neural network; XGBoosted: eXtreme Gradient Boosting; KNN: K-Nearest Neighbors; LREG: logistic regression; SIMCA: soft independent modeling by class analogy; SVM: Support vector machine; TB: tuberculosis; HIV: human immunodeficiency; AIDS: acquired immunodeficiency syndrome; TP: true positive; TN: true negative; FP: false positive; FN: false negative; MCC: Mattew's correlation coefficient.

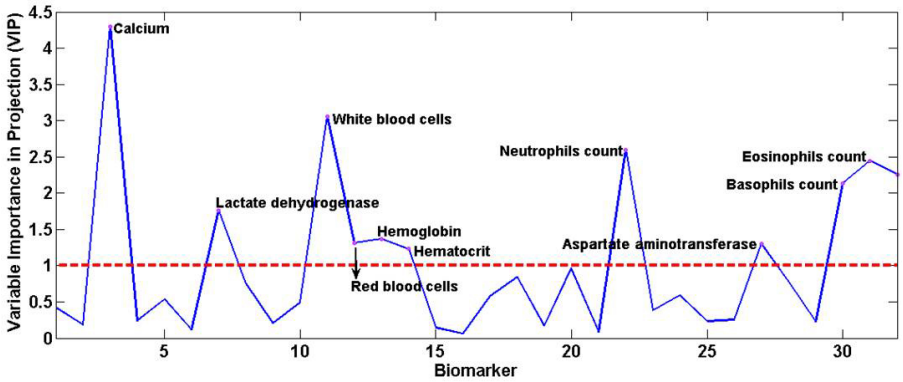


**Figure 4.** Variable importance in projection (VIP) graph of the most important biomarkers for COVID-19 diagnosis. *x* axis represents all analyzed metabolites; *y* axis represents the VIP score that reflects the importance of each metabolite in the prediction of the different classes of the samples (COVID-19 and control). The red dashed line parallel to the x axis represents the VIP score threshold (VIP score threshold = 1). Metabolites that significantly contribute to predicting the diagnosis of COVID-19 are above the threshold (VIP score > 1).

to *Mycobacterium tuberculosis* infection (Figure S3, SI section). The following biomarkers were associated with HIV/TB co-infection: red blood cells, hemoglobin, hematocrit, platelets, lymphocytes, neutrophils, red blood cell distribution width (RBCD), aspartate transaminase (AST), alanine transaminase (ALT), and serum glucose (Figure S3, SI section). The differences in biomarker levels between the HIV/AIDS, TB, HIV/TB co-infection patient group and the controls are shown in Table 3. It is important to highlight that these biomarkers were also identified by the PCA model as important in discriminating the four groups of patients (HIV/AIDS *vs.* TB *vs.* HIV/TB *vs.* control).

### Practical applicability of the machine learning model

The robustness of the PLS-DA model was thoroughly assessed using 1228 external samples sourced from patients in Brazil. These individuals tested negative for COVID-19, HIV/AIDS, and TB, yet presented a spectrum of other conditions including diabetes, hypercholesteremia, hypertriglyceridemia, hypothyroidism, and obesity. The model demonstrated exceptional performance, accurately classifying these cases with a sensitivity of 98% (1205 out of 1228 samples) and an error rate of only 2%. This underscores the model's high predictive efficacy in real-world clinical scenarios (Table 5).

## Discussion

In this study, several machine learning models (PLS-DA, LREG, KNN, XGBoost, SIMCA, ANN, and SVM) were tested to predict the diagnosis of COVID-19, HIV/AIDS, TB, and HIV/TB co-infection, based on data from biochemical. Considering that the data were collected between April-November 2021 and considering the recent new mutations recorded in samples from patients with SARS-CoV-2 (e.g., lambda, gamma, alpha, and beta),[21-23] HIV/AIDS (e.g., mutations 67N, 70R, 184V, 219Q, M184L, and M184T),[26,27] and *M. tuberculosis* (e.g., mutations V91W and delta 438A),[28,29] the ML models of this study are complex. The results of this study showed high diagnostic accuracy in detecting SARS-CoV-2 (area under the curves (AUC) = 0.94), HIV/AIDS (AUC = 0.95), TB (AUC = 0.97), and HIV/TB co-infection (AUC = 0.96). These values are similar to other ML models of COVID-19, HIV/AIDS, and *M. tuberculosis* available in the literature (AUC ROC 70-99%).[30-33]

It is important to highlight that, in real-life data involving COVID-19 (or HIV/AIDS) patients, the data

**Table 5.** Assessment of the predictive performance of the PLS-DA model in predicting external samples from patients without any of the diseases studied (COVID-19, HIV/AIDS and TB) but with other comorbidities

| Group of patients | Total | TP | FN | Non-error rate | Error rate |
|---|---|---|---|---|---|
| Diabetes | 86 | 85 | 1 | 0.988 | 0.012 |
| Deslipidemias[a] | 282 | 275 | 7 | 0.975 | 0.025 |
| Hypothyroidism | 184 | 181 | 3 | 0.984 | 0.016 |
| Obesity | 676 | 664 | 12 | 0.982 | 0.018 |
| Total | 1228 | 1205 | 23 | 0.981 | 0.019 |

[a]Patients with hypercholesterolemia or hypertriglyceridemia. TP: true positive; FN: false negative.

are almost always unbalanced due to the prevalence of the disease, as there are many patients with a negative test result than a positive result, as can be seen in the recent studies by Alves *et al.*[34] Zuin *et al.*[35] Therefore, this imbalance problem was also observed in our study.

In our previous study,[10] using plasma and serum samples from patients with COVID-19 analyzed by liquid chromatography mass spectrometry (LC-MS, metabolomic data), the PLS-DA model was able to predict the diagnosis and severity of COVID-19 with an accuracy of around 93%. In addition, potential new diagnostic and severity biomarkers were identified (ribothymidine, *N*-acetyl-glucosamine-1-phosphate, L-ornithine, and 5,6-dihydro-5-methyluracil).[10] On the other hand, in this present study, we propose a new method for diagnosing COVID-19 using data from routine examinations of patients (biochemical and hematological tests). Although both studies used the PLS-DA model to predict the diagnosis of COVID-19, the potential of our present study (when compared with the previous study)[10] stands out since it uses data from routine examinations of patients, which are more accessible and cheaper, facilitating the application of the PLS-DA model developed in clinical practice. Unlike the previous study, where it is difficult to implement, given the high cost of high performance liquid chromatography (HPLC) and mass spectrophotometer equipment, which are both very sophisticated and very expensive equipment.[36,37] In addition to the difference in the analytical techniques used, another point to be highlighted is that in this study, we also analyzed samples from HIV/AIDS patients, pulmonary TB, and HIV/TB co-infected patients, increasing the complexity of the PLS-DA model developed. Finally, we identified some potential biomarkers associated with the diagnosis of COVID-19 (calcium, LDH, RBC, WBC, neutrophils, basophils, eosinophils, hemoglobin, and hematocrit) and some associated with the diagnosis of HIV/AIDS and TB (lymphocytes, RBC, hematocrit, hemoglobin, AST, ALT, and glycemia); these biomarkers are different from our previous study.

Currently, machine learning and artificial intelligence have often been used to aid in the early diagnosis of diseases (including COVID-19, HIV/AIDS, and TB).[38-40] However, most of these studies compromise their model due to the small sample size.[41-43] In our study, the ML models were trained using a relatively large sample (n = 6,498 patients), obtaining robust and accurate results. Furthermore, to our knowledge, this is the first study where machine learning models were trained for the simultaneous detection of COVID-19, HIV/AIDS, TB, and HIV/TB co-infection.

Alterations in calcium levels in viral diseases have already been demonstrated in previous studies.[44,45] Low calcium is present in COVID-19, according to the systematic review and meta-analysis by Alemzadeh.[46] Recent studies have shown a strong correlation between the positivity of SARS-CoV-2 infection and the low level of calcium in the body, as reported by Cappellini *et al.*[47] and Yang *et al.*[48] Similar results were also addressed in our study, where calcium was considered the main biomarker indicative of the diagnosis of COVID-19, presenting reduced values when compared to controls. *In vitro* and *in vivo* studies involving animals infected with SARS-CoV-2 showed that the SARS-CoV-E gene, a gene located in the apparatus in the Golgi apparatus, is highly synthesized during viral infection and is responsible for encoding the transmembrane protein of calcium ion channel, allowing the permeabilization of calcium. Thus, there is a homeostatic imbalance of intracellular calcium that can activate the inflammatory pathways mediated by TNF, IL-1b, and IL-6, causing cell damage and edema.[45,49] Given the great genomic similarity between SARS-CoV and SARS-CoV-2, the biochemical mechanisms between them may be similar.[50]

In our study, the PLS-DA model identified LDH as a potential diagnostic biomarker of COVID-19, which is in agreement with what has been reported in some systematic reviews[51,52] and our previous study.[9] As LDH is present in lung cells, possibly patients infected with SARS-CoV-2 release a greater amount of LDH in the blood as a result of the injuries suffered by lung cells caused by the virus.[53] Our study also demonstrated elevated levels of AST and ALT in HIV/TB co-infected patients, possibly because these enzymes present high levels during the viral infectious process.[54,55]

Neutropenia in patients infected with SARS-CoV-2, HIV/AIDS, and HIV/TB co-infection agrees with previous studies in the literature.[3,56-58] A potential reason is that patients infected with COVID-19 presented transient agranulocytosis in the initial disease and for the excessive demand for neutrophils in peripheral blood in the face of SARS-CoV-2 infection.[58] In HIV/AIDS infection (or HIV/TB co-infection), neutropenia is due to advanced disease progression, which includes low levels of cluster of differentiation 4 (CD4) cells, and high levels of the HIV/AIDS-1 ribonucleic acid (RNA) virus that causes cytotoxicity.[59] Although there is no evidence demonstrating that HIV/AIDS directly infects and kills mature neutrophils, HIV/AIDS has been shown to have the ability to infect and kill multipotent hematopoietic stem cells by the Fas-Associated death Domain (FAD)-dependent process of apoptosis, and viral proteins are responsible for the suppression of proliferation of granulomonocytic progenitor cells.[60-63]

Anemia is common in serious infectious diseases such as COVID-19, HIV/AIDS infection, and TB, and represents one of the main hematological complications caused by these viruses, contributing to a reduction in the survival rate of patients, low quality of life, and compromised treatment success.[63-66] In SARS-CoV-2 infection, anemia is attributed to the presence of disseminated intravascular coagulation and pulmonary thrombotic microangiopathy, which results in intravascular hemolysis,[64] while in HIV/AIDS, multifactorial causes may be related, such as the presence of opportunistic infections, nutritional deficiencies, changes in the adaptive immune system, pre-existing chronic diseases, and HIV/AIDS infection of stromal cells.[67-72] In the case of *M. tuberculosis* infection, the nutritional factor is the main cause.[73] Despite our results being consistent, the cross-sectional design adopted in this study constituted the main limitation of the work, as this study design does not allow for patient follow-up and blood biomarkers may vary during the disease.

The main limitation of the study is the imbalance between patients (COVID-19, HIV/AIDS, TB, and the control group), which can be justified due to the prevalence and incidence of these diseases. Two other important factors contributed to the imbalance of the data: (*i*) the study data were collected in a hospital in the northern region of Mozambique (Africa) with a lack of financial and human resources, and this is a sad reality in almost all hospitals in Mozambique. We collected this data manually directly from patients' clinical files and there were no electronic forms of patients that could facilitate the data collection process. This entire process was carried out by the study researchers and the study had no funding. (*ii*) The second and last point that contributed to the high imbalance in the data is the high stigmatization and discrimination of patients living with HIV/AIDS and tuberculosis, and this impacted the undercounting of cases and the increase in treatment abandonment rates, drastically reducing the number of data available for collection, and consequently the high imbalance of the groups.

## Conclusions

In conclusion, this study successfully developed and validated several machine learning algorithms (PLS-DA, ANN, XGBoosted, KNN, LREG, SIMCA and SVM) for predicting the diagnosis of COVID-19, HIV/AIDS, TB, and HIV/TB co-infection using data from biochemical and hematological tests. The PLS-DA model demonstrated excellent performance in diagnosing these diseases, achieving accuracy rates ranging from 94 to 97%.

Moreover, the study identified several potential biomarkers associated with each of these diseases, providing valuable insights into diagnostic indicators. For COVID-19, biomarkers such as calcium, lactate dehydrogenase, red blood cells, white blood cells, neutrophils, basophils, eosinophils, hemoglobin, and hematocrit were associated. HIV/AIDS infection was linked to mean corpuscular volume, platelets, neutrophils, and mean platelet volume. In the case of *Mycobacterium tuberculosis* infection, red blood cell distribution width and urea were identified as relevant biomarkers. Finally, HIV/TB co-infection was associated with biomarkers including lymphocytes, red blood cells, hematocrit, hemoglobin, aspartate transaminase, alanine transaminase, and glycemia. The findings underscore the potential of the PLS-DA model for optimizing the diagnosis of these diseases and the significance of specific biomarkers that may aid in screening and early detection. This research contributes to the ongoing efforts to improve diagnostic accuracy and early intervention in infectious diseases, ultimately benefiting patient outcomes and healthcare management.

## Supplementary Information

Supplementary information (Figures S1-S9 and reference for the tutorial of how to use the FAPESP COVID-19 DataSharing/BR repository) is available free of charge at http://jbcs.sbq.org.br as PDF file.

## Acknowledgments

The authors express their gratitude to the Brazilian National Council of Technological and Scientific Development (CNPq) and CAPES (Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil) for research funding - Finance Code 001.

### Author Contributions

AFC, RP, FST, AW was responsible for study concepts; AFC, FST, RP, AAM for study design; AFC, DPS, AAM for data acquisition; AFC, AAM, DPS for statistical analysis; AFC, FST, RP, MMF, AAM for manuscript preparation; AFC, ROV, DPS, LMF, AW for manuscript editing; AFC, RP, FST, LMF, DPS, AW for manuscript review.

## References

1. Richens, J. G.; Lee, C. M.; Johri, S.; *Nat. Commun.* **2020**, *11*, 3923. [Crossref]
2. Hoodbhoy, Z.; Jiwani, U.; Sattar, S.; Salam, R.; Hasan, B.; Das, J. K.; *Front. Artif. Intell.* **2021**, *4*, 708365. [Crossref]

3. Nogueira, M. S.; Leal, L. B.; Marcarini, W. D.; Pimentel, R. L.; Muller, M.; Vassallo, P. F.; Campos, L. C. G.; dos Santos, L.; Luiz, W. B.; Mill, J. G.; Barauna, V. G.; de Carvalho, L. F. C. S.; *Sci. Rep.* **2021**, *11*, 15409. [Crossref]
4. Ruhwald, M.; Hannay, E.; Sarin, S.; Kao, K.; Sen, R.; Chadha, S.; *Lancet Global Health* **2022**, *10*, e465. [Crossref]
5. Banerjee, P.; Eckert, A. O.; Schrey, A. K.; Preissner, R.; *Nucleic Acids Res.* **2018**, *46*, W257. [Link] accessed in January 2024
6. Rehman, A.; Iqbal, M. A.; Xing, H.; Ahmed, I.; *Appl. Sci.* **2021**, *11*, 3414. [Crossref]
7. Hussain, E.; Hasan, M.; Rahman, M. A.; Lee, I.; Tamanna, T.; Parvez, M. Z.; *Chaos, Solitons Fractals* **2021**, *142*, 110495. [Crossref]
8. Haq, A. U.; Li, J. P.; Ahmad, S.; Khan, S.; Alshara, M. A.; Alotaibi, R. M.; *Sensors* **2021**, *21*, 8219. [Crossref]
9. Cobre, A. F.; Stremel, D. P.; Noleto, G. R.; Fachi, M. M.; Surek, M.; Wiens, A.; Tonin, F. S.; Pontarolo, R.; *Comput. Biol. Med.* **2021**, *134*, 104531. [PubMed]
10. de Fátima Cobre, A.; Surek, M.; Stremel, D. P.; Fachi, M. M.; Lobo Borba, H. H.; Tonin, F. S.; Pontarolo, R.; *Comput. Biol. Med.* **2022**, *146*, 105659. [Crossref]
11. WHO Word Health Organization (WHO); *Post COVID-19 Condition (Long COVID)* https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition, accessed in January 2024.
12. de Miranda, D. A. P.; Gomes, S. V. C.; Filgueiras, P. S.; Corsini, C. A.; Almeida, N. B. F.; Silva, R. A.; Medeiros, M. I. V. A. R. C.; Vilela, R. V. R.; Fernandes, G. R.; Grenfell, R. F. Q.; *Trans. R. Soc. Trop. Med. Hyg.* **2022**, *116*, 1007. [Crossref]
13. Panwar, H.; Gupta, P. K.; Siddiqui, M. K.; Morales-Menendez, R.; Singh, V.; *Chaos, Solitons Fractals* **2020**, *138*, 109944. [Crossref]
14. Panahi, A. H.; Rafiei, A.; Rezaee, A.; *Inf. Med. Unlocked* **2021**, *22*, 100506. [Crossref]
15. Minhas, N.; Gurav, Y. K.; Sambhare, S.; Potdar, V.; Choudhary, M. L.; Bhardwaj, S. D.; Abraham, P.; *PLoS One* **2023**, *18*, e0277867. [PubMed]
16. de Oliveira, M. A. L.; Watanabe, A. S. A.; Cesar, D. E.; Candido, J. M. B.; Lima, N. M.; Moreira, O. B. O.; Chellini, P. R.; *Quim. Nova* **2022**, *45*, 760. [Crossref]
17. FAPESP COVID-19 DataSharing/BR, https://repositoriodatasharingfapesp.uspdigital.usp.br/, accessed in January 2024.
18. Mello, L. E.; Suman, A.; Medeiros, C. B.; Prado, C. A.; Rizzatti, E. G.; Nunes, F. L. S.; Barnabé, G. F.; Ferreira, J. E.; Sá, J.; Reis, L. F. L.; Rizzo, L. V.; Sarno, L.; de Lamonica, R.; Maciel, R. M. de B.; Cesar-Jr, R. M.; Carvalho, R.; *Zenodo*, **2020**. [Link] accessed in January 2024
19. *SPPS*, version 20; IBM, New York, USA, 2012.
20. *SOLO*, version 8.8; Eigenvector Research, Inc., Manson, USA, 2020.
21. Voloch, C. M.; Francisco Jr., R. S.; de Almeida, L. G. P.; Cardoso, C. C.; Brustolini, O. J.; Gerber, A. L.; Guimarães, A. P. C.; Mariani, D.; da Costa, R. M.; Ferreira, O. C.; Cavalcanti, A. C.; Frauches, T. S.; de Mello, C. M. B.; Leitão, I. C.; Galliez, R. M.; Faffe, D. S.; Castiñeiras, T. M. P. P.; Tanuri, A.; de Vasconcelos, A. T. R.; *J. Virol.* **2021**, *95*, e00119. [Crossref]
22. Wang, R.; Hozumi, Y.; Yin, C.; Wei, G.-W.; *Genomics* **2020**, *112*, 5204. [Crossref]
23. Nonaka, C. K. V; Franco, M. M.; Gräf, T.; de Lorenzo Barcia, C. A.; de Ávila Mendonça, R. N.; de Sousa, K. A. F.; Neiva, L. M. C.; Fosenca, V.; Mendes, A. V. A.; de Aguiar, R. S.; Giovanetti, M.; de Freitas Souza, B. S.; *Emerging Infect. Dis.* **2021**, *27*, 1522. [Crossref]
24. Ministério da Saúde, Instituto Nacional de Estatística, ICF Macro; *Inquérito de Indicadores de Imunização, Malária e HIV/AIDS/SIDA em Moçambique (IMASIDA) 2015*, https://dhsprogram.com/pubs/pdf/AIS12/AIS12.pdf, accessed in January 2024.
25. Word Health Organization (WHO); *Mozambique-A Comprehensive Community-Based Service Delivery Intervention for TB*, https://www.who.int/publications/m/item/mozambique-a-comprehensive-community-based-service-delivery-intervention-for-tb, accessed in January 2024.
26. Pouga, L.; Santoro, M. M.; Charpentier, C.; Di Carlo, D.; Romeo, I.; Artese, A.; Alcaro, S.; Antinori, A.; Wirden, M.; Perno, C. F.; Ambrosio, F. A.; Calvez, V.; Descamps, D.; Marcelin, A.-G.; Ceccherini-Silberstein, F.; Lambert-Niclot, S.; *Chem. Biol. Drug Des.* **2019**, *93*, 50. [Crossref]
27. Kühnert, D.; Kouyos, R.; Shirreff, G.; Pečerska, J.; Scherrer, A. U.; Böni, J.; Yerly, S.; Klimkait, T.; Aubert, V.; Günthard, H. F.; Stadler, T.; Bonhoeffer, S.; *PLOS Pathog.* **2018**, *14*, e1006895. [Crossref]
28. Machaba, K. E.; Mhlongo, N. N.; Soliman, M. E. S.; *Cell Biochem. Biophys.* **2018**, *76*, 345. [Crossref]
29. Singh, A.; Somvanshi, P.; Grover, A.; *J. Cell. Biochem.* **2019**, *120*, 7386. [Crossref]
30. Meraj, S. S.; Yaakob, R.; Azman, A.; Mohd Rum, S. N.; Ahmad Nazri, A. S.; *Int. J. Adv. Sci. Eng. Inf. Technol.* **2019**, *9*, 81. [Link] accessed in January 2024
31. Wang, L.; Zhang, Y.; Wang, D.; Tong, X.; Liu, T.; Zhang, S.; Huang, J.; Zhang, L.; Chen, L.; Fan, H.; Clarke, M.; *Front. Med.* **2021**, *8*, 704256. [Crossref]
32. Peiffer-Smadja, N.; Rawson, T. M.; Ahmad, R.; Buchard, A.; Georgiou, P.; Lescure, F.-X.; Birgand, G.; Holmes, A. H.; *Clin. Microbiol. Infect.* **2020**, *26*, 584. [Crossref]
33. Pai, N. P.; Karellis, A.; Kim, J.; Peter, T.; *Lancet HIV* **2020**, *7*, e574. [Crossref]
34. Alves, M. A.; Castro, G. Z.; Oliveira, B. A. S.; Ferreira, L. A.; Ramírez, J. A.; Silva, R.; Guimarães, F. G.; *Comput. Biol. Med.* **2021**, *132*, 104335. [Crossref]

35. Zuin, G.; Araujo, D.; Ribeiro, V.; Seiler, M. G.; Prieto, W. H.; Pintão, M. C.; Dos, C.; Lazari, S.; Francisco, C.; Granato, H.; Veloso, A.; *Commun. Med.* **2022**, *2*, 72. [Crossref]
36. Torres, P. B.; Chow, F.; Furlan, C. M.; Mandelli, F.; Mercadante, A.; dos Santos, D. Y. A. C.; *Braz. J. Oceanogr.* **2014**, *62*, 57. [Crossref]
37. Vékey, K.; *J. Chromatogr. A* **2001**, *921*, 227. [Crossref]
38. Caballé-Cervigón, N.; Castillo-Sequera, J. L.; Gómez-Pulido, J. A.; Gómez-Pulido, J. M.; Polo-Luque, M. L.; *Appl. Sci.* **2020**, *10*, 5135. [Crossref]
39. Albahri, A. S.; Hamid, R. A.; Alwan, J. k.; Al-qays, Z. T.; Zaidan, A. A.; Zaidan, B. B.; Albahri, A. O. S.; AlAmoodi, A. H.; Khlaf, J. M.; Almahdi, E. M.; Thabet, E.; Hadi, S. M.; Mohammed, K. I.; Alsalem, M. A.; Al-Obaidi, J. R.; Madhloom, H. T.; *J. Med. Syst.* **2020**, *44*, 122. [Crossref]
40. Marcus, J. L.; Sewell, W. C.; Balzer, L. B.; Krakower, D. S.; *Curr. HIV/AIDS Rep.* **2020**, *17*, 171. [Crossref]
41. Laing, A. G.; Lorenc, A.; del Molino del Barrio, I.; Das, A.; Fish, M.; Monin, L.; Muñoz-Ruiz, M.; McKenzie, D. R.; Hayday, T. S.; Francos-Quijorna, I.; Kamdar, S.; Joseph, M.; Davies, D.; Davis, R.; Jennings, A.; Zlatareva, I.; Vantourout, P.; Wu, Y.; Sofra, V.; Cano, F.; Greco, M.; Theodoridis, E.; Freedman, J. D.; Gee, S.; Chan, J. N. E.; Ryan, S.; Bugallo-Blanco, E.; Peterson, P.; Kisand, K.; Haljasmägi, L.; Chadli, L.; Moingeon, P.; Martinez, L.; Merrick, B.; Bisnauthsing, K.; Brooks, K.; Ibrahim, M. A. A.; Mason, J.; Lopez Gomez, F.; Babalola, K.; Abdul-Jawad, S.; Cason, J.; Mant, C.; Seow, J.; Graham, C.; Doores, K. J.; Di Rosa, F.; Edgeworth, J.; Shankar-Hari, M.; Hayday, A. C.; *Nat. Med.* **2020**, *26*, 1623. [Crossref]
42. Dai, W.; Ke, P.-F.; Li, Z.-Z.; Zhuang, Q.-Z.; Huang, W.; Wang, Y.; Xiong, Y.; Huang, X.-Z.; *J. Med. Internet Res.* **2021**, *23*, e23390. [Crossref]
43. Wu, P.; Ye, H.; Cai, X.; Li, C.; Li, S.; Chen, M.; Wang, M.; Heidari, A. A.; Chen, M.; Li, J.; Chen, H.; Huang, X.; Wang, L.; *IEEE Access* **2021**, *9*, 45486. [Crossref]
44. Deng, B.; Zhang, S.; Geng, Y.; Zhang, Y.; Wang, Y.; Yao, W.; Wen, Y.; Cui, W.; Zhou, Y.; Gu, Q.; Wang, W.; Wang, Y.; Shao, Z.; Wang, Y.; Li, C.; Wang, D.; Zhao, Y.; Liu, P.; *PLoS One* **2012**, *7*, e41365. [Crossref]
45. Nieto-Torres, J. L.; Verdiá-Báguena, C.; Jimenez-Guardeño, J. M.; Regla-Nava, J. A.; Castaño-Rodriguez, C.; Fernandez-Delgado, R.; Torres, J.; Aguilella, V. M.; Enjuanes, L.; *Virology* **2015**, *485*, 330. [Crossref]
46. Alemzadeh, E.; Alemzadeh, E.; Ziaee, M.; Abedi, A.; Salehiniya, H.; *Immunity, Inflamm. Dis.* **2021**, *9*, 1219. [Crossref]
47. Cappellini, F.; Brivio, R.; Casati, M.; Cavallero, A.; Contro, E.; Brambilla, P.; *Clin. Chem. Lab. Med.* **2020**, *58*, e171. [Crossref]
48. Yang, C.; Ma, X.; Wu, J.; Han, J.; Zheng, Z.; Duan, H.; Liu, Q.; Wu, C.; Dong, Y.; Dong, L.; *J. Med. Virol.* **2021**, *93*, 1639. [Crossref]
49. Nieto-Torres, J. L.; DeDiego, M. L.; Verdiá-Báguena, C.; Jimenez-Guardeño, J. M.; Regla-Nava, J. A.; Fernandez-Delgado, R.; Castaño-Rodriguez, C.; Alcaraz, A.; Torres, J.; Aguilella, V. M.; Enjuanes, L.; *PLoS Pathog.* **2014**, *10*, e1004077. [Crossref]
50. Rabaan, A. A.; Al-Ahmed, S. H.; Haque, S.; Sah, R.; Tiwari, R.; Malik, Y. S.; Dhama, K.; Yatoo, M. I.; Bonilla-Aldana, D. K.; Rodriguez-Morales, A. J.; *Infez. Med.* **2020**, *28*, 174. [PubMed]
51. Martha, J. W.; Wibowo, A.; Pranata, R.; *Postgrad. Med. J.* **2022**, *98*, 422, [Crossref]
52. Hariyanto, T. I.; Japar, K. V.; Kwenandar, F.; Damay, V.; Siregar, J. I.; Lugito, N. P. H.; Tjiang, M. M.; Kurniawan, A.; *Am. J. Emerg. Med.* **2021**, *41*, 110. [Crossref]
53. Henry, B. M.; Aggarwal, G.; Wong, J.; Benoit, S.; Vikse, J.; Plebani, M.; Lippi, G.; *Am. J. Emergency Medicine* **2020**, *38*, 1722. [Crossref]
54. Mocroft, A.; Phillips, A. N.; Soriano, V.; Rockstroh, J.; Blaxhult, A.; Katlama, C.; Boron-Kaczmarska, A.; Viksna, L.; Kirk, O.; Lundgren, J. D.; *AIDS Res. Hum. Retroviruses* **2005**, *21*, 743. [Crossref]
55. Enoh, J. E.; Cho, F. N.; Manfo, F. P.; Ako, S. E.; Akum, E. A.; *Biomed Res. Int.* **2020**, *2020*, ID 9631731. [Crossref]
56. Hernandez, J. M.; Quarles, R.; Lakshmi, S.; Casanas, B.; Eatrides, J.; McCoy, E.; Somboonwit, C.; *Open Forum Infect. Dis.* **2021**, *8*, 38. [Crossref]
57. Abay, F.; Yalew, A.; Shibabaw, A.; Enawgaw, B.; *Tuberc. Res. Treat.* **2018**, *2018*, ID 5740951. [Crossref]
58. López-Pereira, P.; Iturrate, I.; de La Cámara, R.; Cardeñoso, L.; Alegre, A.; Aguado, B.; *Clin. Case Rep.* **2020**, *8*, 3348. [Crossref]
59. Hensley-McBain, T.; Klatt, N. R.; *Curr. HIV/AIDS Rep.* **2018**, *15*, 1. [Crossref]
60. Carter, C. C.; Onafuwa-Nuga, A.; McNamara, L. A.; Riddell, J.; Bixby, D.; Savona, M. R.; Collins, K. L.; *Nat. Med.* **2010**, *16*, 446. [Crossref]
61. Banda, N. K.; Tomczak, J. A.; Shpall, E. J.; Sipple, J.; Akkina, R. K.; Steimer, K. S.; Hami, L.; Curiel, T. J.; Singer Harrison, G.; *Apoptosis* **1997**, *2*, 61. [Crossref]
62. Calenda, V.; Graber, P.; Delamarter, J.-F.; Chermann, J.-C.; *Eur. J. Haematol.* **2009**, *52*, 103. [Crossref]
63. Rameshwar, P.; Denny, T. N.; Gascón, P.; *J. Immunol.* **1996**, *157*, 4244. [PubMed]
64. Bergamaschi, G.; Borrelli de Andreis, F.; Aronico, N.; Lenti, M. V.; Barteselli, C.; Merli, S.; Pellegrino, I.; Coppola, L.; Cremonte, E. M.; Croce, G.; Mordà, F.; Lapia, F.; Ferrari, S.; Ballesio, A.; Parodi, A.; Calabretta, F.; Ferrari, M. G.; Fumoso, F.; Gentile, A.; Melazzini, F.; Di Sabatino, A.; Bertolino, G.; Codega, S.; Costanzo, F.; Cresci, R.; Derosa, G.; Di Stefano, M.; Falaschi, F.; Iadarola, C.; Lovati, E.; Lucotti, P. C.; Martignoni, A.; Mengoli, C.; Miceli, E.; Mugellini, A.; Muggia, C.; Noris, P.; Pagani, E.; Palumbo, I.; Pecci, A.; Perrone, T.; Pieresca, C.;

Preti, P. S.; Russo, M. C.; Sgarlata, C.; Siciliani, L.; Staniscia, A.; Vjera, F. T.; Achilli, G.; Agostinelli, A.; Antoci, V.; Banfi, F.; Benedetti, I.; Brattoli, M.; Cambiè, G.; Canta, R.; Cococcia, S.; Conca, F.; Delliponti, M.; Rio, V. Del; Terlizzi, F. Di; Fiengo, A.; Forni, T.; Freddi, G.; Frigerio, C.; Fusco, A.; Gabba, M.; Garolfi, M.; Gori, G.; Grandi, G.; Grimaldi, P.; Lampugnani, A.; Lepore, F.; Lettieri, G.; Mambella, J.; Mercanti, C.; Nardone, A.; Pace, L.; Padovini, L.; Pitotti, L.; Reduzzi, M.; Rigano, G.; Rotola, G.; Sabatini, U.; Salvi, L.; Santacroce, G.; Savioli, J.; Soriano, S.; Spataro, C.; Stefani, D.; *Clin. Exp. Med.* **2021**, *21*, 239. [Crossref]

65. Volberding, P. A.; Levine, A. M.; Dieterich, D.; Mildvan, D.; Mitsuyasu, R.; Saag, M.; *Clin. Infect. Dis.* **2004**, *38*, 1454. [Crossref]
66. Chu, K.-A.; Hsu, C.-H.; Lin, M.-C.; Chu, Y.-H.; Hung, Y.-M.; Wei, J. C.-C.; *PLoS One* **2019**, *14*, e0221908. [Crossref]
67. Teklemariam, Z.; Mitiku, H.; Mesfin, F.; *HIV/AIDS - Res. Palliat. Care* **2015**, *7*, 191. [Crossref]
68. Mihiretie, H.; Taye, B.; Tsegaye, A.; *Anemia* **2015**, *2015*, ID 479329. [Crossref]
69. Kibaru, E. G.; Nduati, R.; Wamalwa, D.; Kariuki, N.; *AIDS Res. Ther.* **2015**, *12*, 26. [Crossref]
70. Bain, B. J.; *Curr. Opin. Hematol.* **1999**, *6*, 89. [Crossref]
71. Volberding, P.; *J. Infect. Dis.* **2002**, *185*, S110. [Crossref]
72. Moore, R. D.; Forney, D.; *JAIDS, J. Acquired Immune Defic. Syndr.* **2002**, *29*, 54. [Crossref]
73. Macallan, D. C.; *Diagn. Microbiol. Infect. Dis.* **1999**, *34*, 153. [Crossref]

*Submitted: April 20, 2023*
*Published online: February 15, 2024*