

# scientific reports



OPEN

## ATR-FTIR spectroscopy in blood plasma combined with multivariate analysis to detect HIV infection in pregnant women

Lidiane G. Silva<sup>1</sup>, Ana F. S. Péres<sup>2,3</sup>, Daniel L. D. Freitas<sup>1</sup>, Camilo L. M. Morais<sup>4</sup>, Francis L. Martin<sup>5</sup>, Janaina C. O. Crispim<sup>2,3</sup> & Kassio M. G. Lima<sup>1</sup>✉

The primary concern for HIV-infected pregnant women is the vertical transmission that can occur during pregnancy, in the intrauterine period, during labour or even breastfeeding. The risk of vertical transmission can be reduced by early diagnosis. Therefore, it is necessary to develop new methods to detect this virus in a quick and low-cost fashion, as colorimetric assays for HIV detection tend to be laborious and costly. Herein, attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy combined with multivariate analysis was employed to distinguish HIV-infected patients from healthy uninfected controls in a total of 120 blood plasma samples. The best sensitivity (83%) and specificity (92%) values were obtained using the genetic algorithm with linear discriminant analysis (GA-LDA). These good classification results in addition to the potential for high analytical frequency, the low cost and reagent-free nature of this method demonstrate its potential as an alternative tool for HIV screening during pregnancy.

The human immunodeficiency virus (HIV) is the cause of a sexually transmitted infection that compromises the immunological system. This virus, belonging to the *Lentiviridae* family, attacks the cells of the immunological system, especially the T lymphocyte CD4+ cells, making them vulnerable to other infections<sup>1</sup>. The decrease of T lymphocytes CD4+ cells by HIV is one of the main causes of AIDS (acquired immune deficiency syndrome)<sup>2</sup>.

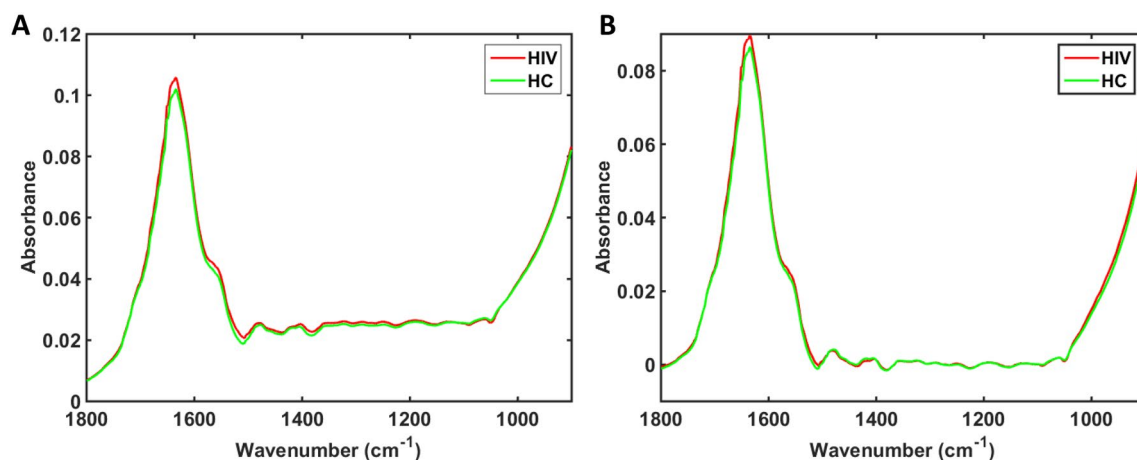
HIV was first identified in the 1980's and since then it has affected millions of people, becoming a major public health problem<sup>3</sup>. Worldwide, around 38 million people are infected with HIV, from which 1.8 million are children and 35 million are adults<sup>4</sup>. According to the epidemiological bulletin from the Brazilian Ministry of Health<sup>5</sup>, 116,292 cases of gestational HIV were diagnosed in Brazil between 2000 and 2018, from which 7,888 cases were recorded in 2017.

One of the main concerns for HIV-infected pregnant women is the viral transmission from the mother to the child<sup>3</sup>. This is termed vertical transmission and can occur during pregnancy (intrauterine), during labour or during breastfeeding. The risk of vertical transmission can be reduced by early diagnosis whereupon antiretroviral treatment can be implemented to reduce viral load. Early HIV diagnosis has increased 21% in the last 10 years, enabling adequate treatment for women during pregnancy, thus improving their quality of life and reducing the risk of vertical transmission. Between 2007 and 2017, prenatal diagnosis resulted in an increase from 2.3 cases per thousand live births to 2.8 cases per thousand live births<sup>5</sup>.

The diagnosis of gestational HIV can be made in blood using immunoassays such as immunochromatography (a rapid test) or the ELISA (enzyme-linked immunosorbent assay) test, which enables one to score both antibodies and antigens. The viral load in HIV-infected patients are estimated > 1,000–1,500 copies/mL<sup>6,7</sup>. Colorimetric assays to detect HIV have a high-cost and require labour-intensive procedures, since each metabolite is detected one at a time<sup>8</sup>.

Spectrometric techniques have been used as metabolomic tools to discriminate biofluids derived from infected vs. healthy patients. Mass spectrometry (MS) and nuclear magnetic resonance (NMR) are the most widely used techniques. Although having high sensitivity and specificity, these techniques have a high-cost and detailed

<sup>1</sup>Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande Do Norte, Natal, RN 5072-970, Brazil. <sup>2</sup>Department of Clinical and Toxicological Analysis, Federal University of Rio Grande Do Norte, Natal, RN 59072-970, Brazil. <sup>3</sup>Maternity School Januário Cicco, Federal University of Rio Grande Do Norte, Natal, RN 59072-970, Brazil. <sup>4</sup>School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, UK. <sup>5</sup>Biocel Ltd, Hull HU10 7TS, UK. ✉email: [kassiolima@gmail.com](mailto:kassiolima@gmail.com)



**Figure 1.** Infrared (IR) spectra. (A) Mean raw IR spectra in the bio-fingerprint region (1800–900  $\text{cm}^{-1}$ ) for HIV-infected (HIV) and healthy uninfected controls (HC) samples. (B) Mean pre-processed IR spectra (AWLS baseline correction) in the bio-fingerprint region (1800–900  $\text{cm}^{-1}$ ) for HIV-infected (HIV) and healthy uninfected controls (HC) samples.

	Group	
	Case (HIV-infection)	Control uninfected
Age (mean $\pm$ SD)	28 $\pm$ 6	31 $\pm$ 6
Gestational age (weeks $\pm$ SD)	21 $\pm$ 7	24 $\pm$ 7
Smoking, <i>n</i> (%)		
Yes	4 (10.3%)	0
No	35 (89.7%)	35 (100%)
Drugs use, <i>n</i> (%)		
Yes	7 (18%)	3 (8.57%)
No	32 (82%)	33 (94.3%)

**Table 1.** Study participant information and demographics.

sample preparation is often required<sup>2</sup>. Attenuated total reflection Fourier-transform infrared (ATR-FITR) spectroscopy is a fast, low-cost, non-destructive and reagent-free alternative to these spectrometric techniques, being able to generate a fingerprint spectrum containing both structural and concentration information for the constituent metabolites in the sample<sup>9</sup>.

There remain a limited number of applications using ATR-FTIR spectroscopy to detect HIV-infected samples. Sitole et al.<sup>8</sup> showed for the first time metabolic differences between healthy uninfected and infected groups using ATR-FTIR spectroscopy where chemometric tools were employed to distinguish healthy uninfected vs. HIV-infected patients, in addition to distinguishing infected patients with or without antiretroviral treatment. Herein, ATR-FTIR spectroscopy combined with multivariate analysis is used to discriminate healthy uninfected and HIV-infected pregnant women based on their blood plasma spectra. Different chemometric techniques were tested, including approaches based on principal component analysis (PCA), successive projections algorithm (SPA) and genetic algorithm (GA).

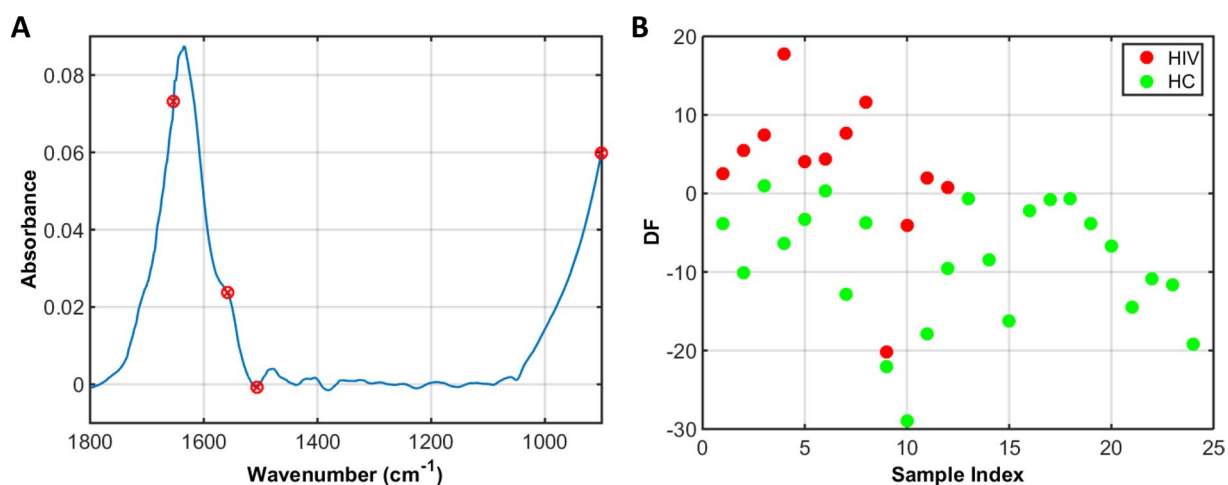
## Results

The infrared (IR) spectra of blood plasma for control (healthy uninfected pregnant women) and HIV-infected pregnant women are shown in Fig. 1A. A total of 120 samples were analysed, generating 360 spectra (3 spectra per sample). The 3 spectra per sample were averaged before model construction. The sample set was composed of 80 healthy uninfected controls and 40 HIV-infected samples. Study participant information and demographics is shown in Table 1.

In the region between 1800 to 900  $\text{cm}^{-1}$  (bio-fingerprint region), the spectra show characteristic absorptions: Amide I band at 1635  $\text{cm}^{-1}$ <sup>10</sup>, an amide II band at 1560  $\text{cm}^{-1}$  (C=O, Amide II)<sup>11</sup> and three small intensity absorptions at 1480  $\text{cm}^{-1}$  (C-H asymmetric deformation of methyl groups), 1404  $\text{cm}^{-1}$  (COO – symmetric stretching of proteins and lipids) and 1060  $\text{cm}^{-1}$  (C-O nucleic acids)<sup>10</sup>. To correct for baseline distortions, the raw spectral data were pre-processed by automatic weighted least squares (AWLS) baseline correction (Fig. 1B). Due to the high similarity between the spectral features for both groups (control uninfected and HIV infected), chemometric approaches were employed to identify spectral features responsible for class differentiation.

Model	AC	SENS	SPEC	F-Score	G-Score
PCA-LDA	70	67	71	69	69
PCA-QDA	58	83	46	59	62
PCA-SVM	81	83	79	81	81
SPA-LDA	75	50	88	64	66
SPA-QDA	64	8	92	15	28
SPA-SVM	67	8	96	15	87
GA-LDA	89	83	92	87	87
GA-QDA	64	8	100	15	29
GA-SVM	66.7	0	100	0	0

**Table 2.** Quality parameters calculated in the test set to classify healthy uninfected controls vs. HIV-infected samples. AC = accuracy, SENS = sensitivity, SPEC = specificity.



**Figure 2.** GA-LDA results. (A) Selected wavenumbers (901, 1506, 1558, 1653 cm<sup>-1</sup>) responsible for class separation. (B) Discriminant function (DF) for the samples in the test set, where HIV stands for HIV-infected samples and HC for healthy uninfected controls.

PCA with linear discriminant analysis (PCA-LDA), quadratic discriminant analysis (PCA-QDA) or support vector machines (PCA-SVM); successive projections algorithm with linear discriminant analysis (SPA-LDA), quadratic discriminant analysis (SPA-QDA) or support vector machines (SPA-SVM); and, genetic algorithm with linear discriminant analysis (GA-LDA), quadratic discriminant analysis (GA-QDA) or support vector machines (GA-SVM) were tested as feature extraction/selection and classification tools.

The performance for each algorithm tested was evaluated by calculating the accuracy, sensitivity, specificity, F-score and G-score for the test set. Both classes were separated into training (70%) and test (30%) sets using the Kennard-Stone algorithm<sup>12</sup>, whereby the training set was used for model construction and the test set for model validation. The classification metrics in the test set for each algorithm is shown in Table 2.

From all models tested, the best classification performances were obtained by PCA-SVM, SPA-LDA and GA-LDA. PCA-SVM showed a good accuracy (81%) and sensitivity (83%), but the specificity was < 80% (79%) in the test. SPA-LDA selected 22 spectral wavenumbers for class differentiation, generating a good specificity (88%), but poor accuracy (75%) and sensitivity (50%). GA-LDA was the best classification algorithm, with an accuracy of 89%, and sensitivity and specificity of 83% and 92%, respectively. The F-score (87%) and G-score (87%) show consistency, indicating that the classification performance was not affected by imbalanced class sizes.

GA-LDA selected 4 spectral wavenumbers, among the 467 in the entire bio-fingerprint spectral range, for class differentiation. These wavenumbers were: 1653, 1558, 1506 and 901 cm<sup>-1</sup> (Fig. 2A). The GA-LDA discriminant function (DF) in the test set is depicted in Fig. 2B, while the tentative biomarkers assignment<sup>9</sup> for the 4 selected wavenumbers are shown in Table 3.

## Discussion

HIV infection remains a major cause of morbidity and mortality worldwide<sup>13</sup>, being mainly transmitted during unprotected sexual intercourse, as well as by the sharing of infected needles or syringes, blood transfusion, and by mother-to-child transmission during pregnancy, labour or breastfeeding<sup>14</sup>. This latter phenomenon is termed vertical transmission and is the focus of this study.

The vertical transmission rate ranges from 15 to 45% worldwide, although this figure can be reduced to < 5% if effective intervention measures are applied. Early diagnosis and correct treatment with anti-retrovirals are among

Selected wavenumber (cm <sup>-1</sup> )	Tentative assignment
901	Phosphodiester stretching
1506	Ring base
1558	Amide II (N–H bending vibration)
1653	Amide I (C=O stretching)

**Table 3.** Wavenumbers selected by GA-LDA to discriminate healthy uninfected controls vs. HIV-infected samples. Tentative assignment based on Movasaghi et al.<sup>10</sup>.

these measures<sup>15</sup>. In Brazil, a considerable proportion of HIV diagnosis for the female population occurs during pregnancy, where diagnosis via an anti-HIV serological screening is performed<sup>15</sup>. This reflects the adequacy of the health policy in prenatal care, although conscientisation campaigns on measures to prevent sexually transmitted infections and unintended pregnancy must still be disseminated widely to improve population health<sup>16</sup>.

Early diagnosis is of fundamental importance towards maintaining both the mother's and new-born's health with the intention of reducing vertical transmission risks. Diagnoses are commonly performed by screening using immunoassays; that is, tests that seek for the presence of anti-HIV antibodies and/or antigen p24 (protein in the virus nucleocapsid). These tests normally have high sensitivity and specificity, and are available in rapid-test formats (e.g., immunochromatography), where the result is obtained within 30 min, or in the ELISA format<sup>17</sup>. Despite these advantages, current biochemical tests require skilled technicians, the use of unstable reagents highly dependent on storage conditions, and production of biochemical waste. Using these approaches, there is a possibility of HIV detection after 2 to 3 weeks from initial viral exposure.

There are other more complex tests generally used as complementary and/or confirmatory tests when screening tests diverge, such as the Western Blot, Immunoblot, Fast Immunoblot and molecular biology testing<sup>18</sup>, although these tests have a substantially higher cost and require highly-skilled labour and specialist facilities. Molecular biology testing is of fundamental importance for early diagnosis, since it is capable of identifying the virus' presence before the period of production of immunoglobulins through the amplification of viral nucleic acids.

ATR-FTIR spectroscopy is capable of detecting metabolic changes after viral infection<sup>19</sup>, and has been successfully used to identify a range of viruses, such as Zika, Dengue and Chikungunya<sup>20,21</sup>. Herein, ATR-FTIR spectroscopy harnessed to chemometric techniques was employed to detect HIV infection in pregnant women. The spectral information responsible for class differentiation is probably associated with metabolic changes caused by the viral infection.

Since the samples were measured in liquid state, water is a possible interference that may compress and mask some of the absorption bands, altering their profiles in comparison with desiccated samples. Water mainly affects the high-region of the IR spectrum, with strong bands at 3,300–3,400 cm<sup>-1</sup> (hydrogen-bonded O–H stretching) and 3,600–3,650 cm<sup>-1</sup> (free O–H stretching)<sup>22</sup>. Since, we worked with the bio-fingerprint region (900–1,800 cm<sup>-1</sup>) only, these main water bands were removed. However, water has an absorbance band within the bio-fingerprint region at 1646 cm<sup>-1</sup>, which broadens the Amide I band and creates a shoulder<sup>23</sup>. In addition, absorbances < 1,000 cm<sup>-1</sup> tend to greatly increase in a liquid environment, hence, compressing the bands between 1,200–1,500 cm<sup>-1</sup><sup>23</sup>.

The virus identification was successfully performed on 40 HIV-infected samples vs. 80 healthy controls using GA-LDA algorithm, with a classification accuracy of 89%. For details, see Supporting information. Four spectral wavenumbers were responsible for class separation based on GA-LDA (Table 2): these included phosphodiester stretching bands at 901 cm<sup>-1</sup>, 1506 cm<sup>-1</sup> (Ring base), protein bands at 1653 cm<sup>-1</sup> (Amide I) and 1558 cm<sup>-1</sup> (Amide II). The sample discrimination caused by changes in these protein features can be justified by the increase of immunoglobulin levels to fight the infection, since when viral infection happens, the immunological system is hyper-activated, hence, increasing production of immunoglobulin<sup>13</sup>.

Sitole et al.<sup>8</sup> using ATR-FTIR spectroscopy also assigned the observed spectral changes between uninfected controls and HIV-infected patients to spectral regions associated with lipids, carbohydrates and proteins. Some protein wavenumbers responsible for class discrimination were: 1652 and 1650 cm<sup>-1</sup> (Amide I) and 1500 and 1504 cm<sup>-1</sup> (Amide II).

Herein, ATR-FTIR spectroscopy with multivariate analysis was able to accurately predict HIV-infected pregnant women based on blood plasma, showing the potential of this method for early detection of HIV in a fast and reagent-free approach. Successful development of this method in a clinical environment could aid early diagnosis of gestational HIV and help treatment.

## Methods

**Samples.** This study was approved by the ethics committee at the Maternidade Escola Januário Cicco (MEJC) from the Federal University of Rio Grande do Norte with the protocol number 1.808.891 and informed consent was obtained from all subjects. Also, all the methods carried out in this study were by the approved guidelines. The patients were selected according to the automatic request that they underwent prenatal appointments, where the patients were asked whether they were willing to participate in this study. If accepted, an Informed Consent Form was read and signed by the participants. The participants answered a questionnaire and some clinical details were obtained from their medical records after consent. The inclusion criteria were: pregnant women between 12 to 30 weeks of pregnancy (Table 1), with or without HIV diagnosed and undergoing prenatal care.

Forty blood plasma samples ( $n = 40$ ) were collected from HIV-positive patients (HIV-infected group) and eighty blood plasma samples ( $n = 80$ ) from healthy uninfected controls. The samples were collected from May 2017 to March 2019 from patients attending the Maternidade Escola Januário Cicco (MEJC) and the Centro Reprodutivo Dra. Leide Moraes in Natal/Brazil. Blood samples were collected into EDTA tubes and, thereafter, centrifuged for 10 min at 1500 rpm at room temperature to separate the blood plasma.

**Spectral acquisition.** An FTIR IR-Affinity-1S spectrometer (Shimadzu Corporation, Japan) was used for spectral acquisition attached to an ATR accessory containing a diamond crystal as reflector element. The instrument was setup to run 32 scans per sample with a spectral resolution of  $4 \text{ cm}^{-1}$ . Measurements were carried out in triplicate.

The IR spectra were recorded between  $4000$  to  $600 \text{ cm}^{-1}$  using  $10 \mu\text{L}$  of sample directly applied onto the ATR crystal. Before starting the measurements, the ATR crystal was cleaned with ethanol 70% v/v and acetone; and before each new sample, the ATR crystal was cleaned with ethanol 70% v/v. A background spectrum was taken before measuring every new sample to account for environmental changes.

**Data analysis.** The spectral data analysis was performed within MATLAB R2014b environment (Math-Works, Inc., USA) using the PLS-Toolbox version 7.9.3 and lab-made routines. The data were pre-processed by automatic weighted least squares (AWLS) baseline correction and organised so that the samples were split into training (70%) and test (30%) sets using the Kennard-Stone uniform sampling algorithm<sup>12</sup>. The following feature extraction/selection and classification algorithms were tested to discriminate healthy controls vs. HIV-infected samples: principal component analysis with linear discriminant analysis (PCA-LDA), quadratic discriminant analysis (PCA-QDA) and support vector machines (PCA-SVM); successive projections algorithm with linear discriminant analysis (SPA-LDA), quadratic discriminant analysis (SPA-QDA) and support vector machines (SPA-SVM); and genetic algorithm with linear discriminant analysis (GA-LDA), quadratic discriminant analysis (GA-QDA) and support vector machines (GA-SVM).

In PCA-LDA/QDA/SVM, the principal component analysis (PCA) scores<sup>24</sup> are used as input variables for linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) or support vector machines (SVM)<sup>25,26</sup>. The successive projections algorithm (SPA)<sup>27</sup> and the genetic algorithm (GA)<sup>28</sup> are feature selection algorithms, where the first works by selecting spectral wavenumbers that minimises the data collinearity, and the latter selecting spectral wavenumbers according to an evolution-based process, where sets of wavenumbers undergo combinations, crossover and mutations until a fitting criterion is reached, which is the minimisation of the following cost-function  $G$ <sup>29</sup>:

$$G = \frac{1}{N_V} \sum_{n=1}^{N_V} g_n \quad (1)$$

where  $N_V$  is the number of validation samples, and  $g_n$  is defined as:

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\text{dmin}_{I(m) \neq I(n)} r^2(x_n, m_{I(m)})} \quad (2)$$

where the numerator is the squared Mahalanobis distance between object  $x_n$  of class index  $I(n)$  and the sample mean  $m_{I(n)}$  of its true class; and the denominator is the squared Mahalanobis distance between object  $x_n$  and the center of the closest wrong class. GA was performed using 100 generations with 200 chromosomes each, with a mutation and crossover probability of 10% and 60%, respectively. LDA, QDA and SVM were used as classification techniques coupled SPA and GA selected variables.

LDA and QDA are discriminant analysis algorithms based on a Mahalanobis distance calculation between the samples. The main difference between them is that LDA assumes classes having similar variance structures, by using a pooled covariance matrix, while QDA assumes classes having different variance structures by using the variance-covariance matrix of each class individually for calculation<sup>25</sup>. The LDA classification score for sample  $i$  of class  $k$  ( $L_{ik}$ ) is calculated in a non-Bayesian as follow<sup>25,30</sup>:

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \mathbf{C}_{\text{pooled}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \quad (3)$$

where  $\mathbf{x}_i$  is a vector with the input variables for sample  $i$ ;  $\bar{\mathbf{x}}_k$  is the mean of class  $k$ ; and  $\mathbf{C}_{\text{pooled}}$  is the pooled covariance matrix between the classes. The QDA classification score for sample  $i$  of class  $k$  ( $Q_{ik}$ ) is estimated using the variance-covariance for each class  $k$  ( $\mathbf{C}_k$ ) in a non-Bayesian form as follows<sup>25,30</sup>:

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \mathbf{C}_k^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \quad (4)$$

SVM is a binary linear classifier using a nonlinear step that transforms the input sample space into a feature space using a kernel function that maximizes the margins of separation between the sample groups<sup>31</sup>. In this study, a radial basis function (RBF) kernel was used since it is able to adapt to different data distributions<sup>32</sup>. The SVM classification rule is obtained by the following Eq. 26:

$$f(x) = \text{sign} \left( \sum_{i=1}^{N_{SV}} \alpha_i y_i k(\mathbf{x}_i, \mathbf{z}_j) + b \right) \quad (6)$$

Parameter	Equation
Accuracy (AC) (%)	$\left(\frac{TP+TN}{TP+FP+TN+FN}\right) \times 100$
Sensitivity (SENS) (%)	$\left(\frac{TP}{TP+FN}\right) \times 100$
Specificity (SPEC) (%)	$\left(\frac{TN}{TN+FP}\right) \times 100$
F-Score (%)	$\left(\frac{2 \times \text{SENS} \times \text{SPEC}}{\text{SENS} + \text{SPEC}}\right)$
G-Score (%)	$\sqrt{\text{SENS} \times \text{SPEC}}$

**Table 4.** Equations to calculate the classification quality parameters. TP stands for true positive, TN for true negative, FP for false positive and FN for false negative.

where  $N_{SV}$  is the number of support vectors;  $\alpha_i$  is the Lagrange multiplier;  $y_i$  is the class membership ( $\pm 1$ );  $k(x_i, z_j)$  is the kernel function; and  $b$  is the bias parameter. These SVM parameters were obtained and optimized via cross-validation.

**Quality performance.** The accuracy, sensitivity, specificity, F-Score and G-Score were calculated in the test set to evaluate the models classification performances. The accuracy represents the total number of samples correctly classified considering true and false negatives, the sensitivity represents the proportion of positive samples (*i.e.*, HIV-infected) correctly classified, the specificity represents the proportion of negative samples (*i.e.*, healthy uninfected controls) correctly classified, the F-score measures the overall classification performance considering imbalanced data, and the G-score measures the overall classification performance not accounting for class sizes<sup>33</sup>. The equations to calculate these parameters are shown in Table 4.

Received: 10 April 2020; Accepted: 4 November 2020

Published online: 19 November 2020

## References

- BRASIL. *Aids/HIV: o que é, causas, sintomas, diagnóstico, tratamento e prevenção*. Brasília: Ministério da Saúde, 2017. Accessed 08 April 2020. <https://www.saude.gov.br/saude-de-a-z/aids-hiv>.
- Bahmani, M. K. *et al.* Spectroscopic characterization of human immunodeficiency virus type-1-infected plasma by principal component analysis and soft independent modeling of class analogy of visible and near-infrared spectra. *Mol. Med. Rep.* **2**, 805–809. [https://doi.org/10.3892/mmr\\_00000176](https://doi.org/10.3892/mmr_00000176) (2009).
- de Lima, S. S. *et al.* HIV na gestação: pré-natal, parto e puerpério. *Ciência Saúde* **10**, 56–61. <https://doi.org/10.15448/1983-652X.2017.1.22695> (2017).
- UNAIDS Brasil. *Programa Conjunto das Nações Unidas sobre HIV/AIDS*. Accessed 08 April 2020. <https://unaids.org.br/prevencao-combinada>.
- BRASIL. *Boletim epidemiológico HIV/Aids 2018*. Brasília: Ministério da Saúde, 2018. Accessed 08 April 2020. <https://www.aids.gov.br/pt-br/pub/2018/boletim-epidemiologico-hivaids-2018>.
- Estill, J. *et al.* Viral load monitoring of antiretroviral therapy, cohort viral load and HIV transmission in Southern Africa: a mathematical modelling analysis. *AIDS* **26**, 1403–1413. <https://doi.org/10.1097/QAD.0b013e3283536988> (2012).
- Marks, G. *et al.* Time above 1500 copies: a viral load measure for assessing transmission risk of HIV-positive patients in care. *AIDS* **29**, 947–954. <https://doi.org/10.1097/qad.0000000000000640> (2015).
- Sitole, L., Steffens, F., Krüger, T. P. & Meyer, D. Mid-ATR-FTIR spectroscopic profiling of HIV/AIDS sera for novel systems diagnostics in global health. *OMICS* **18**, 513–523. <https://doi.org/10.1089/omi.2013.0157> (2014).
- Ellis, D. I., Harrigan, G. G. & Goodacre, R. Metabolic Fingerprinting with Fourier Transform Infrared Spectroscopy. In *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis* (eds Harrigan, G. G. & Goodacre, R.) (Springer, Boston, 2003).
- Movasaghi, Z., Rehman, S. & ur Rehman, I. Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Appl. Spectrosc. Rev.* **43**, 134–179. <https://doi.org/10.1080/05704920701829043> (2008).
- Barth, A. Infrared spectroscopy of proteins. *Biochem. Biophys. Acta Bioenerg.* **1767**, 1073–1101. <https://doi.org/10.1016/j.bbabo.2007.06.004> (2007).
- Kennard, R. W. & Stone, L. A. Computer aided design of experiments. *Technometrics* **11**, 137–148. <https://doi.org/10.1080/00401706.1969.10490666> (1969).
- Deeks, S. G., Overbaugh, J., Phillips, A. & Buchbinder, S. HIV infection. *Nat. Rev. Dis. Primers* **1**, 15035. <https://doi.org/10.1038/nrdp.2015.35> (2015).
- Santos, N. S. O., Romanos, M. T. V. & Wigg, M. D. *Virologia Humana* 3rd edn. (Guanabara Koogan Ltda, Rio de Janeiro, 2015).
- World Health Organization. *Mother-to-child transmission of HIV*. Accessed 08 April 2020. <https://www.who.int/hiv/topics/mtct/about/en/>.
- BRASIL. *Recomendações para profilaxia da transmissão vertical do HIV e terapia antirretroviral em gestantes*. Brasília: Ministério da Saúde, 2010. Accessed 08 April 2020. <https://www.aids.gov.br/pt-br/pub/2010/recomendacoes-para-profilaxia-da-transmissao-vertical-do-hiv-e-terapia-antirretroviral-em>.
- Alexander, T. S. Human immunodeficiency virus diagnostic testing: 30 years of evolution. *Clin. Vaccine Immunol.* **23**, 249–253. <https://doi.org/10.1128/CVI.00053-16> (2016).
- BRASIL. *Manual técnico para o diagnóstico da infecção pelo HIV*. Brasília: Ministério da Saúde, 2013. Accessed 08 April 2020. [https://bvsm.saude.gov.br/bvs/publicacoes/manual\\_tecnico\\_diagnostico\\_infeccao\\_hiv.pdf](https://bvsm.saude.gov.br/bvs/publicacoes/manual_tecnico_diagnostico_infeccao_hiv.pdf).
- Santos, M. C. D., Morais, C. L. M., Nascimento, Y. M., Araujo, J. M. G. & Lima, K. M. G. Spectroscopy with computational analysis in virological studies: a decade (2006–2016). *Trends Analyt. Chem.* **97**, 244–256. <https://doi.org/10.1016/j.trac.2017.09.015> (2017).

20. Santos, M. C. D., Nascimento, Y. M., Araújo, J. M. G. & Lima, K. M. G. ATR-FTIR spectroscopy coupled with multivariate analysis techniques for the identification of DENV-3 in different concentrations in blood and serum: a new approach. *RSC Adv.* **7**, 25640–25649. <https://doi.org/10.1039/C7RA03361C> (2017).
21. Santos, M. C. D. *et al.* ATR-FTIR spectroscopy with chemometric algorithms of multivariate classification in the discrimination between healthy vs. dengue vs. chikungunya vs. zika clinical samples. *Anal. Methods* **10**, 1280–1285. <https://doi.org/10.1039/C7AY02784B> (2018).
22. Morais, C. L. M. *et al.* Standardization of complex biologically derived spectrochemical datasets. *Nat. Protoc.* **14**, 1546–1577. <https://doi.org/10.1038/s41596-019-0150-x> (2019).
23. Passos, J. O. S. *et al.* Spectrochemical analysis in blood plasma combined with subsequent chemometrics for fibromyalgia detection. *Sci. Rep.* **10**, 11769. <https://doi.org/10.1038/s41598-020-68781-x> (2020).
24. Bro, R. & Smilde, A. K. Principal component analysis. *Anal. Methods* **6**, 2812–2831. <https://doi.org/10.1039/C3AY41907J> (2014).
25. Morais, C. L. M. & Lima, K. M. G. Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. *J. Braz. Chem. Soc.* **29**, 472–481. <https://doi.org/10.21577/0103-5053.20170159> (2018).
26. Morais, C. L. M., Costa, F. S. L. & Lima, K. M. G. Variable selection with a support vector machine for discriminating *Cryptococcus* fungal species based on ATR-FTIR spectroscopy. *Anal. Methods* **9**, 2964–2970. <https://doi.org/10.1039/C7AY00428A> (2017).
27. Soares, S. F. C., Gomes, A. A., Araujo, M. C. U., Galvão Filho, A. R. & Galvão, R. K. H. The successive projections algorithm. *Trends Analyt. Chem.* **42**, 84–98. <https://doi.org/10.1016/j.trac.2012.09.006> (2013).
28. McCall, J. Genetic algorithms for modelling and optimisation. *J. Comput. Appl. Math.* **184**, 205–222. <https://doi.org/10.1016/j.cam.2004.07.034> (2005).
29. Siqueira, L. F. S., Araújo Júnior, R. F., de Araújo, A. A., Morais, C. L. M. & Lima, K. M. G. LDA vs. QDA for FT-MIR prostate cancer tissue classification. *Chemometr. Intell. Lab. Syst.* **162**, 123–129. <https://doi.org/10.1016/j.chemolab.2017.01.021> (2017).
30. Dixon, S. J. & Brereton, R. G. Comparison of performance of five common classifiers represented as boundary methods: euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure. *Chemometr. Intell. Lab. Syst.* **95**, 1–17. <https://doi.org/10.1016/j.chemolab.2008.07.010> (2009).
31. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. <https://doi.org/10.1023/A:1022627411411> (1995).
32. Morais, C. L. M., Lima, K. M. G., Singh, M. & Martin, F. L. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat. Protoc.* **15**, 2143–2162. <https://doi.org/10.1038/s41596-020-0322-8> (2020).
33. Morais, C. L. M. & Lima, K. M. G. Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data. *Chemometr. Intell. Lab. Syst.* **170**, 1–12. <https://doi.org/10.1016/j.chemolab.2017.09.001> (2017).

## Acknowledgements

D.L.D. Freitas would like to thank CAPES/PPGQ/UFRN for financial support. L.G. Silva would like to thank the Post-Graduate Program in Chemistry (PPGQ) of UFRN and the CNPq for financial support. C.L.M. Morais would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil (grant 88881.128982/2016-01), for his research grant. The authors greatly appreciate the contribution of all women involved in this study, and the collaboration of the Universidade Federal do Rio Grande do Norte (UFRN), Maternidade Escola Januário Cicco (MEJC), Centro Reprodutivo Dra. Leide Morais, and the Laboratório de Pesquisa em Imunologia Celular e Molecular—Departamento de Análises Clínicas e Toxicológicas da Universidade Federal do Rio Grande do Norte (UFRN) for the support during sample acquisition.

## Author contributions

L.G.S. was responsible for acquiring the spectral data. L.G.S. and A.F.S.P. were responsible for writing the first draft of the manuscript. D.L.D.F. was responsible for the construction of the chemometric models and multivariate analysis. C.L.M.M. and F.L.M. provided chemometric support and finalised the manuscript. J.C.O.C. and K.M.G.L. designed the experiments, supervised the project and revised the manuscript.

## Competing interests

FLM is a shareholder and holds a research position in Biocel UK Ltd, a company that seeks to develop data analytic tools as a service for commercial gain.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-77378-3>.

**Correspondence** and requests for materials should be addressed to K.M.G.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020