analytical chemistry

pubs.acs.org/ac

Article

Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis

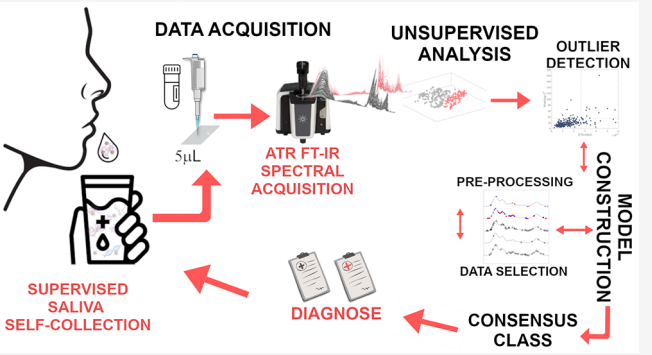
Márcia H. C. Nascimento,[#] Wena D. Marcarini,[#] Gabriely S. Folli, Walter G. da Silva Filho, Leonardo L. Barbosa, Ellisson Henrique de Paulo, Paula F. Vassallo, José G. Mill, Valério G. Barauna, Francis L. Martin, Eustáquio V. R. de Castro, Wanderson Romão, and Paulo R. Filgueiras*

Cite This: *Anal. Chem.* 2022, 94, 2425−2433

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused the worst global health crisis in living memory. The reverse transcription polymerase chain reaction (RT-qPCR) is considered the gold standard diagnostic method, but it exhibits limitations in the face of enormous demands. We evaluated a mid-infrared (MIR) data set of 237 saliva samples obtained from symptomatic patients (138 COVID-19 infections diagnosed via RT-qPCR). MIR spectra were evaluated via unsupervised random forest (URF) and classification models. Linear discriminant analysis (LDA) was applied following the genetic algorithm (GA-LDA), successive projection algorithm (SPA-LDA), partial least squares (PLS-DA), and a combination of dimension reduction and variable selection methods by particle swarm optimization (PSO-PLS-DA). Additionally, a consensus class was used. URF models can identify structures even in highly complex data. Individual models performed well, but the consensus class improved the validation performance to 85% accuracy, 93% sensitivity, 83% specificity, and a Matthew's correlation coefficient value of 0.69, with information at different spectral regions. Therefore, through this unsupervised and supervised framework methodology, it is possible to better highlight the spectral regions associated with positive samples, including lipid (~1700 cm^{-1}), protein (~1400 cm^{-1}), and nucleic acid (~1200−950 cm^{-1}) regions. This methodology presents an important tool for a fast, noninvasive diagnostic technique, reducing costs and allowing for risk reduction strategies.

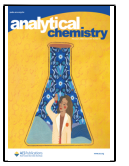


INTRODUCTION

The pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has affected countries around the world since its emergence in Wuhan city, China, in December 2019. Globally, according to the World Health Organization (WHO), as of 1st July 2021, there have been 181 930 736 confirmed cases of coronavirus disease 2019 (COVID-19), including 3 945 832 deaths.[1] One of the critical actions to control the spread of the virus is to quickly isolate infected people. For this, we need virus detection methods that are precise, reliable, and fast, with potential for large-scale implementation.[1−5] The most well-known virus detection assays are the enzyme-linked immunosorbent assay (ELISA) and the reverse transcription quantitative polymerase chain reaction (RT-qPCR). The latter has been used as the gold standard for SARS-CoV-2 infection diagnosis. These methods are touted as repeatable, reproducible, and robust.[6] However, they require laboratory resources and chemical reagents. Besides, the time needed to deliver test results, sample logistics, and other factors require consideration due to the pandemic's enormous demands. Thus, it is urgent to develop reliable and fast methods to accommodate demand for large-scale usage.[1−4,7]

Consequently, vibrational spectroscopy techniques, including infrared (IR) spectroscopy, have been proposed as alternative testing systems since they are reproducible, noninvasive, need little or no sample preparation, and are reagent-free. Moreover, information at the molecular level provides information on functional groups, types of bonds, and molecular conformations, thus potentially identifying important biochemical changes in biological samples in the presence of viruses.[6,8] IR spectroscopy evaluates molecular vibrational modes based on changes in the dipole moment caused by chemical bond vibrations. These vibrational movements allow molecules to absorb radiation energy related to their

Received: September 24, 2021
Accepted: January 13, 2022
Published: January 25, 2022



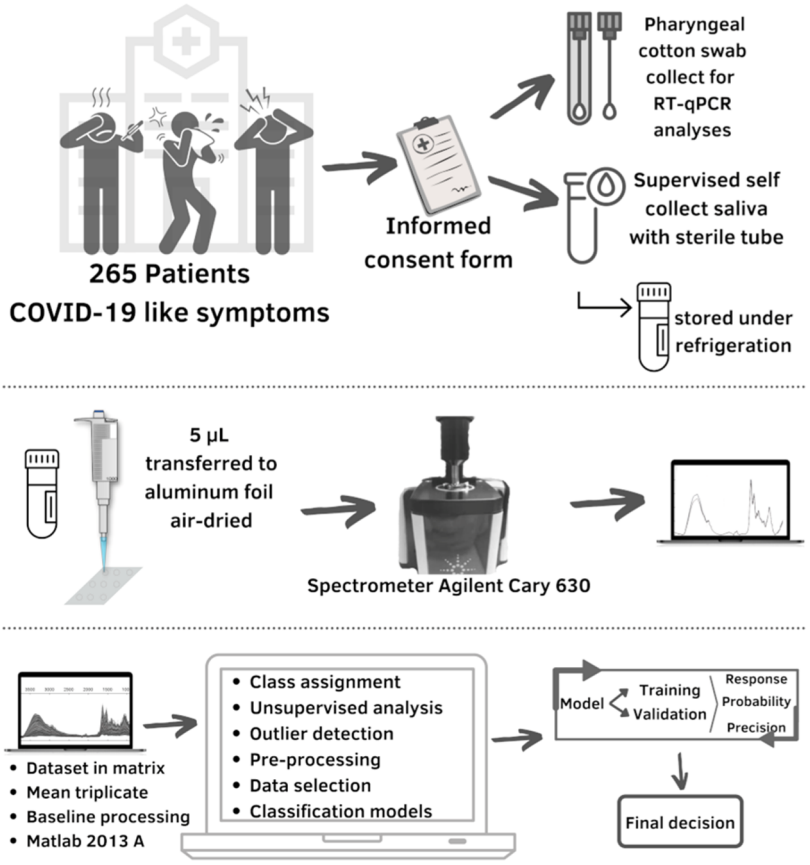


Figure 1. Diagram outlining methodologies for collection of saliva from patients, spectral acquisition, and multivariate statistical data analyses.

vibrational energy levels. Within the IR regions, mid-IR (MIR), whose wavenumbers range from 4000 to 200 cm^{-1},[9] seems to be the most promising in biological analyses since this spectral range includes important biomolecules. This is mainly in the 1800−900 cm^{-1} spectral region, known as the biofingerprint region. In this spectral region, absorptions of different biomolecular constituents occur; they may be biomarkers such as lipids (C═O symmetric stretching at ∼1750 cm^{-1}), carbohydrates (CO−O−C symmetric stretching at ∼1155 cm^{-1}), and nucleic acid (asymmetric phosphate stretching at ∼1225 cm^{-1}, and symmetric phosphate stretching at ∼1080 cm^{-1}), in addition to glycogen and protein phosphorylation (between ∼1030 and 900 cm^{-1}). Proteins exhibit higher signal contributions around amide I at ∼1650 cm^{-1} (80% C═O stretching, 10% C−N stretching, and 10% C−N bending) and amide II at ∼1550 cm^{-1} (60% N−H bending, and 40% C−N stretching); and a lower contribution of amide III at ∼1260 cm^{-1} (C−N stretching).[6,8,10−12]

Studies have reported the use of MIR spectroscopy to detect dengue virus in blood samples,[13] identification of *Staphylococcus aureus* bacteria in blood samples,[14] and stability studies of blood composition from healthy people.[15] However, one of the difficulties in the use of MIR spectroscopy in virology studies is related to the variability of viruses affecting human organisms. Greater virus variability produces more overlapping spectral information within a heterogeneous, complex biological sample derived from human hosts. To address these challenges, biospectroscopy studies have been associated with statistical learning methods. The main methods are principal component analysis (PCA), hierarchical cluster analysis (HCA), and linear discriminant analysis (LDA) with dimension reduction or variable selection methods. This is because spectroscopy data usually present collinear variables. Thus, alongside LDA are applied: genetic algorithm (GA-LDA), successive projection algorithm (SPA-LDA), PCA−DA, and partial least squares (PLS-DA).[6,8,10−14,16−18]

Since 2020, many research groups have directed their attention to spectroscopy methods to detect the SARS-CoV-2 virus, such as the development of plasmonic biosensors[19] and virus detection in blood samples,[20] oral or pharyngeal cell smears,[21] and saliva.[4] Barauna et al.[21] analyzed oral and pharyngeal cell smears in swabs collected from patients with COVID-19 infection-like symptoms. Samples were separated in training (50 positives and 50 negatives) and validation (20 positives and 61 negatives) and classified by GA-LDA with a sensitivity of 95% and a specificity of 89%. However, they associated only one of the five selected variables with the virus' RNA, and other selected variables with the organism's inflammatory responses. Wood et al.[4] studied characteristic spectroscopic signals of SARS-CoV-2 biomarkers with synchrotron-Fourier transform infrared (FTIR) and Raman spectra of purified virus. For COVID-19 infection diagnosis, they modeled 57 mean spectra, of which 29 are positive for SARS-CoV-2 infection by RT-qPCR and 28 are negative. With truncated spectra at 1300−900 cm^{-1}, Monte Carlo double cross-validation, and PLS-DA with an optimized threshold of 0.6, they obtained a sensitivity of 93% and a specificity of 82%. However, they concluded that they needed a larger patient cohort to improve the technique's sensitivity and specificity.

To the best of our knowledge, there are no studies with a diagnosis via spectrochemical analysis of saliva from a large patient cohort. Herein, we aim to evaluate the use of MIR

https://doi.org/10.1021/acs.analchem.1c04162
Anal. Chem. 2022, 94, 2425−2433

spectroscopy associated with pattern recognition methods to classify a higher number of patients via saliva tests into positive or negative SARS-CoV-2 infections. Our aim is to develop a rapid and less invasive diagnostic technique as an alternative to screening patients with COVID-19-like symptoms.

METHODS

Participants. In this study, we evaluated MIR spectra from a total of 265 healthcare services patients from the state of Espírito Santo in Brazil. These patients were assisted according to State Health Secretaries directives and according to the World Health Organization (WHO). This study was carried out in agreement with the Helsinki declaration and authorized by the Hospitals Directive due to the emergency situation. Ethical approval for the investigation was granted by the Ethics Committee at the Universidade Federal do Espírito Santo (#0993920.1.0000.5071 and #31411420.9.0000.8207). Full ethical approval was given to undertake the studies described herein. All patients provided the Informed Consent Form. Next, a nasopharyngeal swab was collected by a healthcare provider for RT-qPCR analysis. Then, the patient received a sterile tube for supervised saliva self-collection. All steps from the point of patient admission for classification models are described in Figure 1.

RT-qPCR analyses were carried out in the central laboratory from the State Health Secretary of Espírito Santo (LACEN−SESA, Brazil). These RT-qPCR results were used for a class assignment for samples, giving a vector of class response.

MIR Spectroscopy. For spectral analysis, 5 μL of saliva were transferred to an aluminum foil and air-dried at room temperature overnight. Spectra were obtained from the aluminum foil containing the dried sample using a transportable benchtop Cary 630 FTIR spectrometer (Agilent Technologies, Inc.), equipped with a diamond attenuated total reflectance (ATR) sampling accessory. The spectral range was from 4000 to 650 cm^{-1}, in the absorbance mode with a 4 cm^{-1} resolution, with 32 scans for the background and the sample.[22,23] For each analysis, the diamond sampling window and the sample press tip were cleaned with 70% ethanol v/v. MIR spectra were acquired in triplicate, with an average time of 90 s per sample, giving us a data set with 795 rows (samples in triplicate) and 1798 columns (variables).

Multivariate Analyses. Average triplicate spectra were obtained and processed for baseline correction using the iteratively reweighted penalized least squares algorithm (airPLS).[24] This procedure can reduce the impact of scattering artifacts, undesirable slopes, and offsets in MIR data sets. This is important for biological studies because the MIR wavelength (2.5−25 μm) includes dimensions of biological cells, providing potential conditions for scattering.[11]

For multivariate analyses, we truncated the spectral data set in the biofingerprint region (1800−900 cm^{-1}) since this region contained relevant biological information.[6,10,12,17] With mean and truncated MIR spectra, we obtained a data set with 265 rows (samples) and 484 columns (variables). These spectra were preprocessed for testing with one or a combination of methods: mean centering, first and second derivatives,[25] standard normal variate (SNV),[26] vector normalization, and multiplicative scatter correction (MSC).[27] All processing was carried out with MATLAB 13A version (The MathWorks, Inc.), with a few toolboxes[28] for modeling and our scripts. This processing was divided into unsupervised analysis to identify trends in the data set and supervised analysis to classify samples as positive or nonpositive for SARS-CoV-2 infections.

Unsupervised Analysis. Due to the complexity of biological spectroscopic information, we chose the random forest algorithm for unsupervised pattern recognition. Random forest (RF) is a machine learning algorithm, developed by Breiman,[29] from the fusion of classification and regression trees (CART) and bootstrapping aggregation (BAGGING).[30] A comprehensive description is in the Supporting Information, but more detailed information can be found elsewhere.[29,31,32]

Herein, we used the unsupervised random forest (URF) model according to the methodology proposed by Afanador et al.[32] to visualize the similarities and differences in the samples. For this, the concatenated matrix with the generated synthetic outliers and the original data set was modeled via the RF model with 1000 trees.[31] The model was evaluated and used to calculate the proximity and dissimilarity matrices. Finally, data trends were evaluated through the dissimilarity matrix in reduced spaces by principal coordinates analysis (PCoA) with the Euclidean distance of samples.

Supervised Analysis. For supervised analysis, we used the RT-qPCR results for the sample class assignment. We calculated the cycle threshold (C_T) average of target genes in the RT-qPCR analysis (Gene N and Gene ORF1ab). Samples with mean C_T < 37 were assigned class one "positive", and samples with mean C_T > 37 were assigned class two "negative". Next, the original data set and the vector of the classes were divided into a training set (70%) and a test set (30%) by the Kennard Stone method,[33] keeping the original proportion between the two classes (positive and negative). Then, the training and test data sets were preprocessed, and outliers were identified by the control chart of Q residual and Hotelling's T^2. Variable selection methods (GA,[34,35] SPA,[14,36−38] and particle swarm optimization (PSO)[39]) associated with classification models (PLS-DA[28,40,41] and LDA[41−43]) were carried out. These methods are described in the Supporting Information.

The selection of variables by GA was performed with 100 generations, each one containing 200 chromosomes; crossover and one-point mutation probabilities were set at 60 and 10%, respectively. A solution was chosen after three cycles were performed. For variable selection through PSO, we used the C_T mean as a response vector, and PSO was tested five times using autoscaling, in which the number of particles (popsize) equaled 10 during 10 iterations. Finally, several models were acquired by a selected variable set (SPA-LDA; GA-LDA; PLS-DA; PSO-PLS-DA). The PLS-DA and PSO-PLS-DA were k-fold cross-validated (k = 10) to optimize latent variable (LV) via the error rate in the cross-validation. Settings to train the models are shown in Table S-1.

Consensus Class. Like high-level data fusion,[44,45] which is operated at the decision level, separate models were built, and their predictions were integrated into a single final response.[46] Then, each sample was classified considering the predicted categories and their calculated probability among all of the models (eq 1). For this, we evaluated a combination of the individual decision of three classification models (GA-LDA, PLS-DA, and PSO-PLS-DA) to a final class decision for the samples. The class defined for the sample of each built model is shown in Table S-2, in which each sample was arranged for two results [positive (1) and negative (2)].

$$\mathrm{Prob}_{\mathrm{mult}} = \sum (\mathrm{Prob}_{\mathrm{class1}}) / \sum (\mathrm{Prob}_{\mathrm{class2}}) \quad (1)$$

https://doi.org/10.1021/acs.analchem.1c04162
Anal. Chem. 2022, 94, 2425−2433

where $Prob_{class1}$ is the class 1 calculated probability in the model, $Prob_{class2}$ is the class 2 calculated probability, and $Prob_{mult}$ is the deciding factor obtained between the two classes. The higher ratio value formulates the final decision class for the sample.

Classification models were evaluated for accuracy, sensitivity, specificity, and other metrics as described in Table S-3. To present the statistical significance, we evaluated these models by the *y*-permutation test. For this, class labels of the training data set were permuted, the permuted model was built, and the predicted class was provided via the permuted model. Performance parameters of the original model are expected out of the distribution of the permuted models. This *y*-permutation test was evaluated by the *F*1 score metric for class 1 (positive) (eq 2). This metric was used to statistically represent the conjunct of performance parameters via the only scalar value. The *F*1 score is a harmonic mean of the precision and sensitivity, where the *F*1 score reaches its best value at 1 and worst value at 0.

$$F1 = 2 \cdot \frac{(\text{precision} \cdot \text{sensitivity})}{(\text{precision} + \text{sensitivity})} \quad (2)$$

Finally, the evaluated models were used to estimate diagnosis in a new data set. For this, models were applied to a data set from 59 randomly selected and newly collected samples (177 spectral triplicates). These new spectra were processed like the training and validation data sets and were classified by individual models and the consensus class. The RT-qPCR of the new data set was obtained, and we calculated metrics for revision.

RESULTS AND DISCUSSION

In this study, we successfully identified a structure in the data set via the URF model. Then, we built linear classification models and tested them to diagnose saliva samples as either positive or negative for COVID-19 infections via MIR spectra. The procedures to identify and remove outliers (before and after preprocessing data) resulted in a data set with 237 samples. Table 1 describes sample profiles grouped by gender and comorbidities, with numbers and percentages in their respective subset. There are 138 samples of patients with positive and 99 with negative COVID-19 RT-qPCR diagnoses. Additionally, out of the positive samples, 67% are women and 33% are men. Similarly, out of the negative samples, 71% are women and 29% are men. However, the data showed insufficient statistical evidence (α = 0.05) to reject the null hypothesis: diagnostic distribution is independent of patients' genders (biological difference) through the χ^2 test (*p*-value of 0.5095).

Table 1. Health Data for Participants

	number (%)		
	total (n = 237)	positive	negative
gender			
female	162 (68%)	92 (57%)	70 (43%)
male	75 (32%)	46 (61%)	29 (39%)
hypertension	58 (24%)	33 (57%)	25 (43%)
diabetes	20 (8%)	13 (65%)	7 (35%)
chronic obstructive pulmonary disease (COPD)	15 (6%)	10(67%)	5 (33%)
obesity	13 (5%)	8 (62%)	5 (38%)

Participants in this study were patients assisted through the health services, between the ages of 20 and 97 years old. However, most of the cohort comprises people aged between 30 and 60 years old (Figure S-1), with only one patient >90 years. Through the χ^2 test, there is no sufficient statistical evidence (α = 0.05) to reject the null hypothesis, i.e., the diagnostic distribution is independent of patients' ages (*p*-value of 0.5541). C_T values of target genes were distributed between 12.33 to 40.38 of gene N and 10.99 to 41.56 to gene ORF1ab; those nondetected were assigned values of 42 (Figure S-2).

Spectra profiles in the biofingerprint (Figure 2) and full spectral regions (Figure S-3) exhibit high intraclass variability, with few observable differences between positive (Figure 2A) and negative samples (Figure 2B).

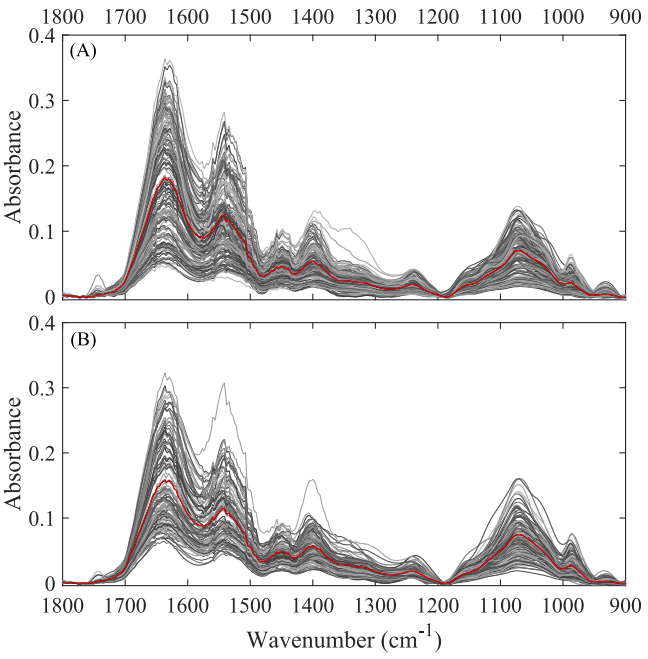


Figure 2. Mid-infrared (MIR) spectral data set from saliva samples of n = 237 patients with RT-qPCR diagnoses for COVID-19 infection with an average spectrum (red line). (A) Positive (n = 138 samples) and (B) negative (n = 99 samples).

This chosen biofingerprint region is important for biological studies due to the information on molecular vibrations, including lipids (~1750 cm^{-1}), carbohydrates (~1155 cm^{-1}), proteins (amide I, ~1650 cm^{-1}; amide II, ~1550 cm^{-1}; amide III, ~1260 cm^{-1}), in addition to DNA/RNA (~1225 and ~1080 cm^{-1}).[6,8] Table 2 and Figure S-4 show the principal MIR band assignment[6,8] for this data set in the biofingerprint region, while Figure S-5 shows raw spectra and baseline corrected aspects.

Unsupervised Random Forest. A URF model was applied to identify a possible structure of the spectral data set. Since data were modeled with bootstrapping of samples and variables in the presence of synthetic outliers, this structure allows a distinction between the original and synthetic data. The RF model distinguished the original and synthetic data sets with an accuracy of 98.2%, a sensitivity of

https://doi.org/10.1021/acs.analchem.1c04162
Anal. Chem. 2022, 94, 2425−2433

Table 2. Principal Mid-Infrared (MIR) Bands of the Data Set and Chemical Assignments[a,6,8]

band	tentative assignment
~3275 cm^{-1}	stretching O−H symmetric
~3200−3550 cm^{-1}	symmetric and asymmetric vibrations attributed to water
~2930 cm^{-1}	stretching C−H
~2800−3000 cm^{-1}	C−H lipid region
~2100 cm^{-1}	combination of hindered rotation and O−H bending (water)
~1750 cm^{-1}	lipids: ν(C═C)
~1650 cm^{-1}	amide I: ν(C═O)
~1550 cm^{-1}	amide II: δ(N−H) coupled to ν(C−N)
~1450 cm^{-1}	methyl groups of proteins: $\delta[(CH_3)]$ asymmetric
~1400 cm^{-1}	methyl groups of proteins: $\delta[(CH_3)]$ symmetric
~1250−1260 cm^{-1}	amide III: ν(C−N)
~1155 cm^{-1}	carbohydrates: ν(C−O)
~1225 cm^{-1}	DNA and RNA: $\nu_{as}(PO_2^-)$
~1080 cm^{-1}	DNA and RNA: $\nu_s(PO_2^-)$
~1030 cm^{-1}	glycogen vibration: ν_s(C−O)
~971 cm^{-1}	nucleic acids and proteins: $n(PO_4)$
~960−966 cm^{-1}	C−O, C−C, deoxyribose

[a] ν_s = symmetric stretching; ν_{as} = asymmetric stretching; and δ = bending.

98.9%, and a specificity of 97.5%, indicating that the data were structured.

From this URF model, the proximity matrix allowed PCoA (80% variance), and samples were projected in three dimensions by a PCoA scores graph (Figure 3). It can be seen that there is a structure allowing visualization of different groups. However, classes were unsatisfactorily distanced in PCo1, which we expected to classify these samples.

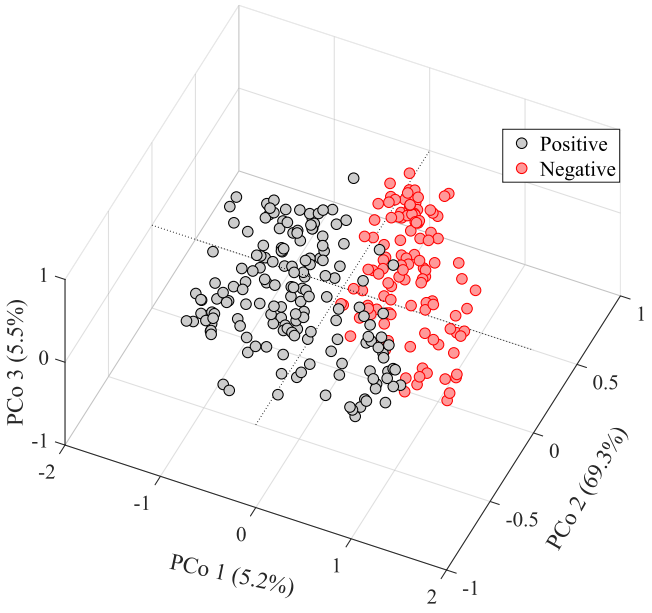


Figure 3. Principal coordinates analysis (PCoA) scores plot from the unsupervised random forest (URF) model from the mid-infrared (MIR) saliva data set (n = 246).

In this URF model, we identified 82 variables with higher frequencies (Figure S-6). These variables show the band characteristic of lipid regions (1785−1729 cm^{-1}; stretching C═C and C═O of ester groups) and proteins (1680 and 1718 to 1705 cm^{-1}: stretching C═O and C−N; 1600−1250 cm^{-1}: amides I, II, and III). Moreover, they also show the characteristic acid nucleic bands (1612−1606 cm^{-1}: adenine vibration in DNA; 1244−1100 cm^{-1}: stretching PO_4 of phosphodiester groups; 1025−1021 cm^{-1}: C−O stretching (carbohydrates); 961 cm^{-1}: deoxyribose; and 930−909 cm^{-1}: phosphodiester stretching bands).[6,8]

Supervised Analyses. We applied linear models to classify samples with variable selection methods (SPA-LDA and GA-LDA), dimension reduction (PLS-DA), and a combination of variable selection and dimension reduction (PSO-PLS-DA). SPA-LDA, GA-LDA, and PLS-DA are the most applied classification methods in biological studies.[6] The same training and test sets were used for each model. Several preprocessing methods were tested, but the 2nd derivative (21 points of the window, and second-degree polynomial; Figure S-7) produced better results in the classification models. From the response of individual models, the consensus class was assigned to samples via the probability of models (Table S-2). Out of the training set, 35 samples (21%) were misclassified, characterizing false positives and false negatives, and in the test set, this number decreased by 12 (17%) according to the consensus class confusion matrix (Table 3 and Figure S-8). The confusion matrix of individual models is shown in Table S-4.

Table 3. Confusion Matrix of the Consensus Class of Training and Test Data Sets[a]

actual class	TP	TN	FP	FN
training data set				
positive	82	48	20	15
negative	48	82	15	20
test data set				
positive	38	22	9	3
negative	22	38	3	9

[a] TP = true positive; TN = true negative; FP = false positive; and FN = false negative.

The principal performance metrics of individual models and parameters from the consensus class are shown in Table 4 for the training and test sets. Out of the individual models, GA-LDA and PSO-PLS-DA highlight better parameters. Matthew's correlation coefficient (MCC) is used mainly for the unbalanced number of samples between the classes.[47−49] This parameter uses the confusion matrix to calculate a correlation between actual and estimated classes. An MCC value near zero suggests that the prediction was not better than a random prediction.[47−49] GA-LDA, PSO-PLS-DA, and SPA-LDA models had an MCC value >0.5, despite the parameter obtained from the consensus class for the final decision.

Distinct bands were selected by the models. In SPA, five variables were selected; in GA, 34 variables were selected; and in PSO, 45 variables were selected. In PLS-DA models, more important variables were identified by coefficient values. This identification was carried out separately for classes 1 and 2. From the PLS-DA model without variable selection, 63 wavenumbers were highlighted for class 1 and 37 for class 2, whereas from PLS-DA after the PSO method, 6 variables were highlighted for class 1 and 5 for class 2. It can be seen (Figure 4) that the lipid regions are highlighted in these selections (1707−1792 cm^{-1}), mainly GA (Figure 4A), PSO (Figure 4C), and PLS-DA (Figure 4D,E). From PLS-DA (Figure 4D),

https://doi.org/10.1021/acs.analchem.1c04162
Anal. Chem. 2022, 94, 2425−2433

Table 4. Quality Parameters of Classification Models[a]

model	preprocessing	set	samples/class POS	NEG	outlier	quality parameters SENS (%)	SPEC (%)	PREC CL.1 (%)	PREC CL.2 (%)	ACC (%)	MCC
SPA-LDA	second derivative	train	97	68	7	69.1	63.2	72.8	58.9	66.7	0.32
		test	41	31	2	87.8	67.7	78.3	80.8	79.8	0.57
GA-LDA	second derivative	train	97	68	7	86.6	67.6	79.2	77.9	78.8	0.56
		test	41	31	2	95.1	70.9	81.2	91.7	84.7	0.69
PLS-DA (7 LV)	second derivative	train	97	68	7	70.0	76.5	80.9	64.2	72.7	0.46
		test	41	31	2	75.6	74.2	79.5	69.7	75.0	0.49
PSO-PLS-DA (9 LV)	second derivative /mean-centered	train	97	68	7	79.4	76.5	82.8	72.2	78.2	0.55
		test	41	31	2	82.9	74.2	80.9	76.7	79.2	0.57
consensus class		train	97	68	7	82.5	75.0	82.0	75.0	79.0	0.57
		test	41	31	2	93.0	74.0	83.0	88.0	85.0	0.69

[a]SENS = sensitivity; SPEC = specificity; PREC CL.1 = precision of class 1, or prevalence positive value (PPV); PREC CL.2 = precision of class 2, or prevalence negative value (PNV); ACC = accuracy; and MCC = Mathew's correlation coefficient.

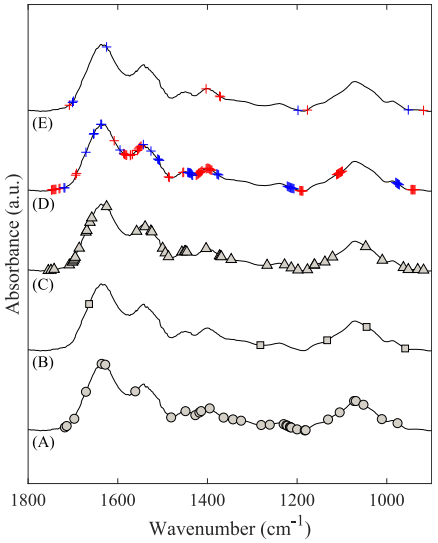


Figure 4. Most important and selected variables through selection methods and classification models. (A) Genetic algorithm linear discriminant analysis (GA-LDA) selected variables; (B) successive projection algorithm LDA (SPA-LDA) selected variables; (C) particle swarm optimization (PSO) selected variables; (D) most important variables for class 1 (red +) and class 2 (blue +) through partial least squares discriminant analysis (PLS-DA) coefficient values; and (E) most important variables for class 1 (red +) and class 2 (blue +) through PSO-PLS-DA coefficient values.

this region is more important to distinguish class 1, i.e., positive class. There are few studies with evidence of a relationship between triglyceride levels and COVID-19 infections in biofluids and other resources.[50,51] The amide I region (~1650 cm^{-1}) was selected for the GA (Figure 4A) and PSO (Figure 4C) methods. Also, this region is highlighted in PLS-DA and PSO-PLS-DA (Figure 4E) for class 2, i.e., negative. Regions showing higher PLS-DA (Figure 4D) coefficient values for class 1 are bands closer to 1400 cm^{-1}, that is the protein region, and closer to 1200 cm^{-1}, 1155 cm^{-1}, and 950 cm^{-1} that comprise carbohydrates, DNA/RNA, and nucleic acid regions, respectively. Moreover, PSO-PLS-DA (Figure 4E) reduced 90% of the variables most important for class 1 and 86.5% for class 2 when compared to PLS-DA without variable selection (Figure 4D).

A few of the selected variables match those described in Barauna et al.[21] (~1429, ~1220, ~1069 cm^{-1}), despite the higher number of selected regions in this study. However, Barauna et al.[21] used spectra from swabs with saliva collected and dried, containing a few better-defined bands at 1100−900 cm^{-1} regions. Because the cell smear can present a higher component concentration, this may explain the spectral difference and increased variance in this region. Moreover, Wyllie et al.[7] reported higher SARS-CoV-2 RNA copies in saliva (5.58 mean log copies mL^{-1}) compared to nasopharyngeal swabs (4.93 mean log copies mL^{-1}). This virus has a preferential tropism to human airway epithelial cells, and salivary glands could be a potential target for SARS-CoV-2.[2,3,5,7,52]

In another paper with a classification of biological samples for the diagnosis of COVID-19 through MIR spectroscopy, Zhang et al.[20] achieved a distinction between the MIR spectra of blood serum samples through the PLS-DA model (a sensitivity of 83.1% and a specificity of 98%) with data processed by the second derivative among control group patients (healthy people) and patients with the confirmed diagnosis of COVID and respiratory infection diseases. The most important regions for the models were 1450−1650 and 1050−1100 cm^{-1}. However, besides the invasive samples and increased time for analyses, the authors emphasized, in that study, that either the spectra of asymptomatic patients or those diagnosed, but with a few days that showed symptoms, the model may not correctly identify. This challenge is corroborated by our results since even with acceptable accuracy, models can show a high false-positive rate (FPR).

Herein, considering the participants presented with respiratory infection symptoms (Table S-5), the potential to distinguish the relevant biochemical changes related to SARS-CoV-2 presence in their biological system is expected with the proposed method. In addition, the prevalence negative value (PNV) or precision of the consensus class for the negative class (class 2) was 88%. This suggests that although the symptoms are similar, the model distinguished the negative samples for SARS-CoV-2 with good precision. In this case, the false-negative rate (FNR) may be a problem with more preoccupation levels. One infected person classified as healthy can potentially contribute to spreading the virus. For this reason, the sensitivity (93%) and the prevalence negative value (88%) are potential indicators that the modeled biomarkers in MIR spectra are related to SARS-CoV-2 virus presence in saliva samples. The participant cohorts present variability of symptoms from mild to moderate and days of

https://doi.org/10.1021/acs.analchem.1c04162
Anal. Chem. 2022, 94, 2425−2433

symptoms range from 1 to 10. However, it is more concentrated between days 3 and 6, with a few outliers >10 days (Figure S-9). In samples with 3 days of symptoms, there is a higher false-negative number. Between 4 and 5 days, there is a higher false-positive level, and from 6 days of symptoms onward, the trend is toward an increase in false-negative levels (Figure S-10). However, the χ^2 test (α = 0.05) shows no sufficient statistical evidence to reject the null hypothesis that the distribution of misclassified samples through the consensus class is independent of the days a patient showed symptoms (p-value of 0.4224).

To evaluate their clinical application, the models were tested on a new data set (n = 59, from symptomatic patients at the same region and health services) to classify with individual classification models and final decision by the consensus class. A few outliers in this new data set were identified and excluded from this application (n = 8, ∼13%). After clinical diagnoses of these samples, we calculated the quality parameters of this new prediction (Table S-6). The accuracy was decreased by 59% from the final decision and 63% from the PSO-PLS-DA model. This suggests that models need to improve robustness. GA-LDA gave a higher FPR in this new application (68%), while other models gave FPRs of ∼50%. PSO-PLS-DA gave better quality parameters in this new prediction when compared to other models. Recently, Wood et al.[4] modeled 29 positive saliva samples for SARS-CoV-2 infection and 28 negatives. Moreover, they developed a modified reflection accessory for transflection IR to optimize the point-of-care diagnosis and to maximize signal absorbance. They obtained a sensitivity of 93% and a specificity of 82% using the spectral region at 1300−900 cm^{-1}. However, given their small data set, they concluded that they need a larger patient cohort to improve sensitivity and specificity. In this study, we evaluated a high number of samples (n = 237), and we tested a new data set (n = 59). In addition, we identified important spectral regions through variable selection methods and the consensus class that may clarify a relationship between spectral information and the biological COVID-19 infection response. Furthermore, from the y-permutation test (Figure S-11), we see the consensus class contributes to turn the classification response statistically significant compared to an individual model.

CONCLUSIONS

The variable selection methods and linear classification models can identify positive saliva samples with 83% accuracy, and precision values of 80 and 88% for positive and negative for COVID-19 infections, respectively. Although the individual GA-LDA model performs well in the validation set with 95% sensitivity and 85% accuracy, the consensus class adds robustness to the prediction of new samples since GA-LDA incurs a higher false-positive rate (68%). The models' estimated classes for a new random set of samples (n = 59) were not equivalent to those obtained in the validation set. However, PSO-PLS-DA estimated classes better (77% sensitivity, 48% specificity, and 63% accuracy). This suggests that PSO-PLS-DA may be an alternative classification method for screening suspected samples. Their performance at the validation set also suggests this (83% sensitivity, 74% specificity, 79% accuracy, and an MCC value of 0.57).

MIR spectroscopy sensitivity for this analysis has been confirmed in recent studies with biological fluids. The unsupervised analysis of the URF method shows a specific structure in the MIR spectroscopic data. Besides, supervised analyses highlight relevant spectral regions related to virus biomarkers and infection responses.

The wider implementation of this methodology will require the identification of confounding factors, like COVID-19 biological response, other types of infections, or other viruses, besides asymptomatic people. Our results show that this methodology is a potential tool to isolate possible spreaders of the disease, due to the possibility of rapid diagnosis (minutes) and reduced demand for supplies. In addition, collection of saliva samples by patients themselves avoids the direct interaction between healthcare providers and patients and may be an alternative for screening infected people.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c04162.

Complementary theory about techniques applied in the experimental section; tables include classification model settings; class predicted; performance characteristic equations; confusion matrix of classification models; and cross-table of χ^2 test symptoms; figures include the histogram of patient ages; histogram of target genes from RT-qPCR analysis; spectra with band assignments, preprocessing aspect, and higher frequency variables in the URF model; actual and consensus class prediction per sample; predicted class by the consensus class versus days of first symptoms; boxplot of days of first symptoms; and histogram of $F1$ score distribution from the y-permutation test (PDF)

AUTHOR INFORMATION

Corresponding Author

Paulo R. Filgueiras − *Chemometrics Laboratory of the Center of Competence in Petroleum Chemistry − NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29075-910, Brazil;* orcid.org/0000-0003-2617-1601; Email: paulo.filgueiras@ufes.br

Authors

Márcia H. C. Nascimento − *Chemometrics Laboratory of the Center of Competence in Petroleum Chemistry − NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29075-910, Brazil*

Wena D. Marcarini − *Department of Physiological Sciences, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29040-090, Brazil*

Gabriely S. Folli − *Chemometrics Laboratory of the Center of Competence in Petroleum Chemistry − NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29075-910, Brazil*

Walter G. da Silva Filho − *Department of Physiological Sciences, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29040-090, Brazil*

Leonardo L. Barbosa − *Department of Physiological Sciences, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29040-090, Brazil*

Ellisson Henrique de Paulo − *Chemometrics Laboratory of the Center of Competence in Petroleum Chemistry − NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29075-910, Brazil;* orcid.org/0000-0001-9960-846X

https://doi.org/10.1021/acs.analchem.1c04162
Anal. Chem. 2022, 94, 2425−2433

Paula F. Vassallo − *Clinical Hospital, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais 31270-901, Brazil*
José G. Mill − *Department of Physiological Sciences, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29040-090, Brazil*
Valério G. Barauna − *Department of Physiological Sciences, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29040-090, Brazil*
Francis L. Martin − *Biocel UK Ltd, Hull HU10 6TS, U.K.;* orcid.org/0000-0001-8562-4944
Eustáquio V. R. de Castro − *Chemometrics Laboratory of the Center of Competence in Petroleum Chemistry − NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29075-910, Brazil*
Wanderson Romão − *Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo, Vila Velha 29106-010, Brazil*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c04162

Author Contributions

#M.H.C.N. and W.D.M. contributed equally to the study. M.H.C.N. and W.D.M.: Methodology, Validation, Investigation, Data curation, Writing - original draft, editing; G.S.F. and E.H.P.: Data curation, Writing - original draft, editing; W.G.S.F. and L.L.B.: Data acquisition, Writing - original draft, visualization; P.F.V., J.G.M., V.G.B., F.L.M., E.V.R.C., W.R., and P.R.F.: Conceptualization, Writing - review & editing, supervision, project administration, and funding acquisition.

Notes

The authors declare the following competing financial interest(s): F.L.M. holds a position and shareholdings in Biocel UK Ltd. and its subsidiary companies; these companies are developing a spectrochemical test for SARS-CoV-2 commercial prescreening.

ACKNOWLEDGMENTS

The authors would like to thank the patients who, even in such a delicate time, voluntarily agreed to participate in this study; the health services and LACEN for the analysis; and LabPetro (UFES, Brazil) for performing the FTIR measurements (Technical Cooperation Agreements No. 0050.0022844.06.4). The authors also would like to thank all of the hospitals (Vila Velha Hospital, Hospital Universitário Cassiano Antonio Moraes, Pronto Atendimento da Glória, and Hospital Roberto Arnizaut Silvares), employees, nurses, doctors, and patients that accepted to participate in this study. The authors would like to thank the municipality of Vila Velha and the Secretaria Municipal de Saúde. The authors would like to thank the Instituto Capixaba de Ensino, Pesquisa e Inovação em Saúde (ICEPi) and the Secretaria Estadual de Saúde do Espírito Santo's (SESA) clinical patient data and RT-PCR results. This study was supported by FAPES (#151/2020, 165/2021, and 442/2021) and CNPq (#310057/2020-5, 401870/2020-0, and 313500/2021-5). This study was financed, in part, by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

REFERENCES

(1) Geneva: World Health Organization WHO COVID-19 Explorer. https://worldhealthorg.shinyapps.io/covid/ (accessed 30th Jun, 2021).
(2) Khiabani, K.; Amirzade-Iranaq, M. H. *Am. J. Infect. Control* **2021**, *49*, 1165−1176.
(3) Cañete, M. G.; Valenzuela, I. M.; Garcés, P. C.; Massó, I. C.; González, M. J.; Providell, S. G. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2021**, *131*, 540−548.
(4) Wood, B. R.; Kochan, K.; Bedolla, D. E.; Salazar-Quiroz, N.; Grimley, S. L.; Perez-Guaita, D.; Baker, M. J.; Vongsvivut, J.; Tobin, M. J.; Bambery, K. R.; Christensen, D.; Pasricha, S.; Eden, A. K.; Mclean, A.; Roy, S.; Roberts, J. A.; Druce, J.; Williamson, D. A.; McAuley, J.; Catton, M.; Purcell, D. F. J.; Godfrey, D. I.; Heruad, P. *Angew. Chem., Int. Ed.* **2021**, *60*, 17102−17107.
(5) Vogels, C. B. F.; Watkins, A. E.; Harden, C. A.; Brackney, D. E.; Shafer, J.; Wang, J.; Caraballo, C.; Kalinich, C. C.; Ott, I. M.; Fauver, J. R.; Kudo, E.; Lu, P.; Venkataraman, A.; Tokuyama, M.; Moore, A. J.; Muenker, M. C.; Casanovas-Massana, A.; Fournier, J.; Bermejo, S.; Campbell, M.; Datta, R.; Nelson, A.; dela Cruz, C. S.; Ko, A. I.; Iwasaki, A.; Krumholz, H. M.; Matheus, J. D.; Hui, P.; Liu, C.; Farhadian, S. F.; Sikka, R.; Wyllie, A. L.; Grubaugh, N. D.; et al. *Med* **2021**, *2*, 263−280.e6.
(6) Santos, M. C. D.; Morais, C. L. M.; Nascimento, Y. M.; et al. *TrAC, Trends Anal. Chem.* **2017**, *97*, 244−256.
(7) Wyllie, A. L.; Fournier, J.; Casanovas-Massana, A.; Melissa, C.; Maria, T.; Pavithra, V.; Joshua, L. W.; Bertie, G.; M Catherine, M.; Adam J, M.; Chantal B F, V.; Mary E, P.; Isabel, M. O.; Peiwen, L.; Arvind, V.; Alice, L.-C.; Jonathan, K.; Rebecca, E.; et al. *N. Engl. J. Med.* **2020**, *383*, 1283−1286.
(8) Movasaghi, Z.; Rehman, S.; Rehman, I. U. *Appl. Spectrosc. Rev.* **2008**, *43*, 134−179.
(9) Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, D. L. *Spectrometric Identification of Organic Compounds*; 8th ed.; John Wiley & Sons:New York, 2014.
(10) Morais, C. L. M.; Lima, K. M. G.; Singh, M.; Martin, F. L. *Nat. Protoc.* **2020**, *15*, 2143−2162.
(11) Butler, H. J.; Smith, B. R.; Fritzsch, R.; Radhakrishnan, P.; Palmer, D. S.; Baker, M. J. *Analyst* **2018**, *143*, 6121−6134.
(12) Mitchell, A. L.; Gajjar, K. B.; Theophilou, G.; Martin, F. L.; Martin-Hirsch, P. L. *J. Biophotonics* **2014**, *7*, 153−165.
(13) Santos, M. C. D.; Nascimento, Y. M.; Araújo, J. M. G.; Lima, K. M. G. *RSC Adv.* **2017**, *7*, 25640−25649.
(14) de Sousa Marques, A.; de Melo, M. C. N.; de Cidral, T. A.; de Lima, K. M. G. *J. Microbiol. Methods* **2014**, *98*, 26−30.
(15) Huber, M.; Kepesidis, Kv.; Voronina, L.; Božić, M.; Trubetskov, M.; Harbeck, N.; Krausz, F.; Žigman, M. *Nat. Commun.* **2021**, *12*, No. 1511.
(16) Sakudo, A.; Suganuma, Y.; Kobayashi, T.; Onodera, T.; Ikuta, K. *Biochem. Biophys. Res. Commun.* **2006**, *341*, 279−284.
(17) Pupeza, I.; Huber, M.; Trubetskov, M.; Schweinberger, W.; Hussain, S. A.; Hofer, C.; Fritsch, K.; Poetzlberger, M.; Vamos, L.; Fill, E.; Amotchkina, T.; Kepesidis, Kv.; Apolonski, A.; Karpowicz, N.; Pervak, V.; Pronin, O.; Fleischmann, F.; Azzeer, A.; Žigman, M.; Krausz, F. *Nature* **2020**, *577*, 52−59.
(18) Bel'skaya, Lv. *J. Appl. Spectrosc.* **2019**, *86*, 187−205.
(19) Peng, X.; Zhou, Y.; Nie, K.; Zhou, F.; Yuan, Y.; Song, J.; Qu, J. *New J. Phys.* **2020**, *22*, No. 103046.
(20) Zhang, L.; Xiao, M.; Wang, Y.; Peng, S.; Chen, Y.; Zhang, D.; Zhang, D.; Guo, Y.; Wang, X.; Luo, H.; Zhou, Q.; Xu, Y. *Anal. Chem.* **2021**, *93*, 2191−2199.
(21) Barauna, V. G.; Singh, M. N.; Barbosa, L. L.; Marcarini, W. D.; Vassallo, P. F.; Mill, J. G.; Ribeiro-Rodrigues, R.; Campos, L. C. G.; Warnke, P. H.; Martin, F. L. *Anal. Chem.* **2021**, *93*, 2950−2958.
(22) Baker, M. J.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, H. J.; Dorling, K. M.; Fielden, P. R.; Fogarty, S. W.; Fullwood, N. J.; Heys, K. A.; Hughes, C.; Lasch, P.; Martin-Hirsch, P. L.; Obinaju, B.; Sockalingum, G. D.; Sulé-Suso, J.; Strong, R. J.; Walsh, M. J.; Wood, B. R.; Gardner, P.; Martin, F. L. *Nat. Protoc.* **2014**, *9*, 1771−1791.

(23) Martin, F. L.; Kelly, J. G.; Llabjani, V.; Martin-Hirsch, P. L.; Patel, I. I.; Trevisan, J.; Fullwood, N. J.; Walsh, M. J. *Nat. Protoc.* **2010**, *5*, 1748−1760.
(24) Zhang, Z. M.; Chen, S.; Liang, Y. Z. *Analyst* **2010**, *135*, 1138−1146.
(25) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627−1639.
(26) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. *Appl. Spectrosc.* **1989**, *43*, 772−777.
(27) Isaksson, T.; Naes, T. *Appl. Spectrosc.* **1988**, *42*, 1273−1284.
(28) Ballabio, D.; Consonni, V. *Anal. Methods* **2013**, *5*, 3790−3798.
(29) Breiman, L. *Mach. Learn.* **2001**, *45*, 5−32.
(30) Breiman, L. *Mach. Learn.* **1996**, *24*, 123−140.
(31) Lovatti, B. P. O.; Nascimento, M. H. C.; Neto, ÁC.; Castro, E. V. R.; Filgueiras, P. R. *Microchem. J.* **2019**, *145*, 1129−1134.
(32) Afanador, N. L.; Smolinska, A.; Tran, T. N.; Blanchet, L. *J. Chemom.* **2016**, *30*, 232−241.
(33) Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137−148.
(34) Siqueira, L. F. S.; Araújo Júnior, R. F.; de Araújo, A. A.; Morais, C. L. M.; Lima, K. M. G. *Chemom. Intell. Lab. Syst.* **2017**, *162*, 123−129.
(35) Shaffer, R. E.; Small, G. W. *Chemom. Intell. Lab. Syst.* **1996**, *35*, 87−104.
(36) Pontes, M. J. C.; Galvão, R. K. H.; Araújo, M. C. U.; et al. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 11−18.
(37) Marques, A. S.; Castro, J. N. F.; Costa, F. J. M. D.; Neto, R. M.; Lima, K. M. G. *Microchem. J.* **2016**, *124*, 306−310.
(38) Araújo, M. C. U.; Saldanha, teresa C. B.; Galvão, R. K. H.; Yoneyama, T.; Chame, H. C.; Visani, V. *Chemom. Intell. Lab. Syst.* **2001**, *57*, 65−73.
(39) Marini, F.; Walczak, B. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 153−165.
(40) Brereton, R. G.; Lloyd, G. R. *J. Chemometrics* **2014**, *28*, 213−225.
(41) Barker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166−173.
(42) Chen, L. F.; Liao, H. Y. M.; Ko, M. T.; Lin, J. C.; Yu, G. J. *Pattern Recognit.* **2000**, *33*, 1713−1726.
(43) Martinez, A. M.; Kak, A. C. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 228−233.
(44) Biancolillo, A.; Bucci, R.; Magrì, A. L.; Magrì, A. D.; Marini, F. *Anal. Chim. Acta* **2014**, *820*, 23−31.
(45) Castanedo, F. *Sci. World J.* **2013**, *2013*, 1−19.
(46) Deng, B.-C.; Yun, Y.-H.; Liang, Y.-Z. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 166−176.
(47) Jurman, G.; Riccadonna, S.; Furlanello, C. *PLoS One* **2012**, *7*, No. e41882.
(48) Boughorbel, S.; Jarray, F.; El-Anbari, M. *PLoS One* **2017**, *12*, No. e0177678.
(49) Brereton, R. G. *J. Chemom.* **2021**, *35*, No. e3331.
(50) Song, J. W.; Lam, S. M.; Fan, X.; Cao, W. J.; Wang, S. Y.; Tian, H.; Chua, G. H.; Zhang, C.; Meng, F. P.; Xu, Z.; Fu, J. L.; Huang, L.; Xia, P.; Yang, T.; Zhang, S.; Li, B.; Jiang, T. J.; Wang, R.; Wang, Z.; Shi, M.; Zhang, J. Y.; Wang, F. S.; Shui, G. *Cell Metab.* **2020**, *32*, 188−202.e5.
(51) Spick, M.; Longman, K.; Frampas, C.; Lewis, H.; Costa, C.; Walters, D. D.; Stewart, A.; Wilde, M.; Greener, D.; Evetts, G.; Trivedi, D.; Barran, P.; Pitt, A.; Bailey, M. *EClinicalMedicine* **2021**, *33*, No. 100786.
(52) Liu, L.; Wei, Q.; Alvarez, X.; Wang, H.; Du, Y.; Zhu, H.; Jiang, H.; Zhou, J.; Lam, P.; Zhang, L.; Lackner, A.; Qin, C.; Chen, Z. *J. Virol.* **2011**, *85*, 4025−4030.

https://doi.org/10.1021/acs.analchem.1c04162
Anal. Chem. 2022, 94, 2425−2433