Supporting Information

NON-INVASIVE DIAGNOSTIC FOR COVID 19 FROM SALIVA BIOFLUID VIA FTIR SPECTROSCOPY AND MULTIVARIATE ANALYSIS

Márция H. C. Nascimento1+, Wena D. Marcarini2+, Gabriely S. Folli1, Walter G. da Silva Filho2, Leonardo L. Barbosa2, Ellisson H. de Paulo1, Paula F. Vassallo3, José G. Mill2, Valério G. Barauna2, Francis L. Martin4, Eustáquio V. R. de Castro4, Wanderson Romão5, Paulo R. Filgueiras1*

1 Chemometrics Laboratory of the Center of Competence in Petroleum Chemistry – NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo, 29075-910, Brazil.

2 Department of Physiological Sciences, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo 29040-090, Brazil.

3 Clinical Hospital, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais 31270-901, Brazil.

4 Biocel UK Ltd, Hull HU10 6TS, UK.

5 Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo, Vila Velha, 29106-010, Brazil.

* Both authors contributed equally to the study

* Corresponding author: Paulo R. Filgueiras

E-mail address: paulo.filgueiras@ufes.br
**Supporting Information**

**Summary**

1. Supporting information about Experimental section
2. **Table S-1.** Model settings.
3. **Table S-2.** Class prediction of test dataset by four classification models.
4. **Table S-3.** Performance characteristic equations of the classification models
5. **Figure S-1.** Patient age (years) histogram. (A) Positive RT-qPCR diagnostic; (B) Negative RT-qPCR diagnostic.
6. **Figure S-2.** Target genes from RT-PCR analysis histograms. (A) Gene N; (B) Gene ORF1ab.
7. **Figure S-3.** MIR spectral dataset from saliva samples of \( n=237 \) patients with RT-qPCR diagnoses for COVID-19 infection. A) positive \( (n=138 \) samples), and B) negative \( (n=99 \) samples).
8. **Figure S-4.** Mean spectrum of positive samples (red) and negative samples (blue) with main MIR band assignments.
9. **Figure S-5.** Original MIR spectra dataset from saliva samples \( (n=265) \) with the mean spectrum (red line) and with different evaluated baseline correction approaches: A) raw spectra B) Spectra after baseline corrected via the airPLS algorithm; C) Spectra after first derivative pre-processing applied to baseline correction; D) Spectra after SNV pre-processing and Savitzky-Golay smoothing applied to baseline correction.
10. **Figure S-6.** Mean derived spectrum of positive samples (blue) and negative samples (red) with identifying the higher frequency in URF model (dotted line).
11. **Figure S-7.** MIR truncated spectra (after baseline correction) of 237 samples (A) with a mean spectrum (red line), and MIR processed spectra via second derivative with a mean spectrum (red line).
12. **Figure S-8.** Actual and consensus predicted classes for \( n=165 \) saliva samples in training dataset (A), and for 72 saliva samples in test dataset (B).
13. **Table S-4.** Confusion matrix of classification models from training and test dataset.
14. **Table S-5.** Cross-table of \( \chi^2 \) test symptoms (according to Centers for Disease Control and Prevention, CDC) of participant cohorts dataset.
15. **Figure S-9.** Boxplot of days of first symptoms from dataset \( (n=237) \).
16. **Figure S-10.** Graph of actual and predicted classes by the consensus class and days of first symptoms from dataset \((n=237)\). (A) Positive class markers; (B) Negative class markers.

17. **Table S-6.** Quality parameters of prediction new data set \((n=59)\) from applied models.

18. **Figure S-11.** Histograms of distribution F1 score class 1 Train and Test datasets from permuted classification models (blue bars), with the F1 score class 1 train and Test datasets from original classification models (red vertical line).

19. Supplementary References
Supporting information about Experimental section

Unsupervised analysis

Random forest (RF) is a machine learning algorithm developed by Breiman\textsuperscript{1} from the fusion of classification and regression trees (CART), and bootstrapping aggregation (BAGGING)\textsuperscript{2}. In CART, only one tree is grown by recursive partitions resulting in hierarchical connected nodes. Each node corresponds to a cut-off value that is equivalent to a variable split in two ways, in which one is a final node (leaf), and another, a branch for another node. The tree grows from the root node, which is a variable split with a higher contribution to distinguish samples between classes. Then, recursive binary partitions are carried out until an answer is reached as a final node\textsuperscript{1,3}. Then, the RF model is an ensemble of N trees as independent classifiers.

At each tree grown, a sample is drawn by the bootstrapping algorithm\textsuperscript{1,3} from the original dataset, resulting in a random subset with a size of the original data set. One tree is grown on the new bootstrapped data, and about 2/3 of the samples comprise the original dataset, that is, “a bag”. Consequently, 1/3 of the samples in the original dataset consist of the test set, called “out-of-bag”\textsuperscript{1}. Bootstrap sampling provides different possibilities for the training set in the classification trees. Moreover, a fraction of the variables is also selected randomly for this set to construct the first tree. When building a classifier tree, cross-validation by the k-fold method (k=10) is used to identify predictor variables among the fraction initially inserted in the training set. These predictor variables are selected by the Gini impurity index (Equations S-1 and S-2) splitting the ascending node into descending nodes by recursive partitions\textsuperscript{1,3}.

\[
Gini(node) = 1 - \sum_{i=1}^{C} p \left( \frac{i}{\text{node}} \right)^2
\]  
(S-1)
where $p$ is the probability of any sample belonging to class $C$; $\left( \frac{i_{\text{node}}}{\text{node}} \right)$ is the fraction of classes with class $i$ reaching the node. Next, the chosen split variable is based on information gain value ($Gain_{\text{Gini}}$), which is the difference from the Gini index between the ascending and descending nodes (Equation S-2) $^{1,3}$.

$$Gain_{\text{Gini}} = Gini_{(\text{ascending})} - \sum_{nd=1}^{n} \left( \frac{nd}{na} \right) * Gini_{(nd)}$$ (S-2)

where $n$ is the number of descending nodes, $nd$ is the total of descending node events, and $na$ is the total of ascending node events. Next, the test samples (out-of-bag, OOB) were classified by the grown classifier tree, accuracy was calculated, and the OOB error rate was reported with the RF algorithm. This prediction of the OOB set has internal validation, usually reported as an estimation of the true prediction error $^{1,3-5}$.

An unsupervised RF (URF) is based on an assumption about the structure of the original data: if the data holds any structure, it should be distinguishable from a synthetic dataset version of itself $^5$. For this, artificial outliers of the same size as the original dataset are generated. Then, the matrix with the synthetic and original datasets are concatenated, and the vector of the two classes are labeled “synthetic” and “original” for the respective samples. From this complete dataset, the RF model is built as a supervised RF binary classification problem. For this model, the number of trees is optimized, but generally 500 trees are enough $^5$. After this, proximity and dissimilarity matrices are calculated. A final procedure is the principal coordinates analysis (PCoA) to produce ordinations (Euclidean distance) of the samples by the dissimilarity matrix in reduced spaces. Multidimensional scaling (MDS) and PCoA are similar to PCA, but instead of converting correlations into a two-dimensional graph, they convert distances among the samples.
**Variable selection methods**

Genetic algorithm (GA) and successive projection algorithm (SPA) variable selection methods were associated with linear discriminant analysis (LDA) to constitute the classification GA-LDA and SPA-LDA models. In the GA model, variables are randomly binary-coded with 0 and 1. The identification of the most relevant variables is based on Darwinian natural selection, in which the selected variables are those that have the best genetic code (variables coded with 1) via the evolution of individuals over many generations. A new set, smaller and more correlated with the parameter of interest will be used with the LDA 6,7.

The SPA algorithm aims to eliminate variables whose information is redundant, that is, correlated information, reducing data collinearity. It starts with a variable that incorporates another variable, until the desired number of variables is reached. Chains of variables are generated and evaluated using a risk equation ($G$) (Equations S-3 and S-4). The new set with the lowest $G$ value will be used to build the SPA-LDA model 8–11.

\[
G = \frac{1}{k_v} \sum_{k=1}^{K_v} gK
\]  

(S-3)

where $gK$ is defined as

\[
gK = \frac{r^2(x_k, \mu_{Ik})}{\min_{l \neq k} r^2(X_k, \mu_{lj})}
\]  

(S-4)

Particle swarm optimization (PSO) is another technique that uses binary encoding to select variables. PSO simulates the behavior of animals in search of food or shelter to find the best suitable response surface for the optimization problem. Individuals “fly” in the search space, communicate with each other, and update their $X_i$ positions, and $v_i$ velocities, until the optimal solution is found, or the stopping
condition is achieved (Equations S-5 and S-6). Updates occur through factors of inertia $\mathbf{w}$, the factor of individual memory $\mathbf{c}_1$ (cognitive factor), and social knowledge $\mathbf{c}_2$ (social factor). These last two factors are controlled by the cognitive $\mathbf{r}_1$ and social $\mathbf{r}_2$ weights, respectively, and they are randomly created with values between 0 and 1. Additionally, the each particle’s best positions are given by the $\mathbf{p}_{\text{best}}$ vector, which serves to update the particle's best position (fitness) individually. Finally, by the fitness value, the swarm updates its position ($\mathbf{G}_{\text{best}}^{\text{it}}$) 

\[
X_i^{\text{it+1}} = X_i^{\text{it}} + v_i^{\text{it+1}} 
\]  

(S-5)

\[
v_i^{\text{it+1}} = w \ast v_i^{\text{it}} + (c_1^{\text{it}} \ast r_1^{\text{it}} \ast (p_{\text{best}}^{\text{it}} - X_i^{\text{it}})) + (c_2^{\text{it}} \ast r_2^{\text{it}} \ast (G_{\text{best}}^{\text{it}} - X_i^{\text{it}})) 
\]  

(S-6)

For variable selection, a response vector should be used, and the selection can be run one or many times, reducing the selected variables set. Number of particles (popsize) and iterations are also optimized. Finally, the combination of PSO and the partial least squares discriminant analysis (PSO-PLS-DA) model are acquired by a selected variables set.

**Linear classification models**

PLS-DA is a supervised linear classification method based on PLS regression, where the information of each class, i.e., the $\mathbf{y}$ vector, is used as criteria for sample classification \(^{13-15}\). The PLS-DA uses latent variables (LV) to capture the important information (variance) of the data matrix ($\mathbf{X}$) combined with the answer vector ($\mathbf{y}$). In PLS-DA, the probability is calculated for each class, and the classification of samples is performed by the class with the highest probability. The class choice limit (belongs (1) or does not belong (0)) can be calculated by Bayes' theorem, assuming a normal distribution of the data \(^{13,14}\). LDA is a methodology that can map sets of high-
dimensional discriminant projection vectors (projection axes) and reallocate them in low-dimensional space. The projection axes, through the projected ones, form the maximum dispersion between the different classes and the minimum dispersion between them\textsuperscript{15–17}. 
| Model    | Preprocessing type | Class_prob or assignment method | Method | Number of components or latent variables (LV) | Cross-validation method | Cross-validation groups |
|----------|--------------------|----------------------------------|--------|-----------------------------------------------|-------------------------|------------------------|
| GA-LDA  | None               | Proportional probability         | Linear | ----                                           | Venetian blinds         | 10                     |
| PLS-DA  | Mean centering     | Thresholds based on Bayes Theorem|        | 7                                              | Venetian blinds         | 10                     |
| SPA-DA  | None               | Proportional probability         | Linear | ----                                           | Venetian blinds         | 10                     |
| PSO-PLSDA | Mean centering   | Thresholds based on Bayes Theorem|        | 9                                              | Venetian blinds         | 10                     |
**Table S-2.** Class prediction of test dataset by four classification models.

| Samples | Actual class | Predicted class from models | GA-LDA* | SPA-LDA | PLS-DA* | PSO-PLS-DA* | *Consensus class |
|---------|--------------|-------------------------------|---------|---------|---------|-------------|-----------------|
| 1       | 1            | 1                             | 1       | 2       | 2       | 1           |                 |
| 2       | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 3       | 1            | 2                             | 2       | 2       | 2       | 2           |                 |
| 4       | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 5       | 1            | 1                             | 1       | 2       | 1       | 1           |                 |
| 6       | 1            | 1                             | 1       | 2       | 2       | 2           |                 |
| 7       | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 8       | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 9       | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 10      | 1            | 1                             | 1       | 2       | 1       | 1           |                 |
| 11      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 12      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 13      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 14      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 15      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 16      | 1            | 1                             | 1       | 2       | 1       | 1           |                 |
| 17      | 1            | 1                             | 1       | 2       | 1       | 1           |                 |
| 18      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 19      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 20      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 21      | 1            | 1                             | 1       | 2       | 1       | 1           |                 |
| 22      | 1            | 1                             | 1       | 2       | 1       | 1           |                 |
| 23      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 24      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 25      | 1            | 1                             | 1       | 1       | 2       | 1           |                 |
| 26      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 27      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 28      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 29      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 30      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 31      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 32      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 33      | 1            | 2                             | 1       | 2       | 2       | 2           |                 |
| 34      | 1            | 1                             | 1       | 1       | 2       | 1           |                 |
| 35      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 36      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 37      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 38      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |
| 39      | 1            | 1                             | 1       | 1       | 1       | 1           |                 |

Continue...
| Samples | Actual class | GA-LDA* | SPA-LDA | PLS-DA* | PSO-PLS-DA* | Consensus class |
|---------|--------------|---------|---------|---------|-------------|----------------|
| 40      | 1            | 1       | 1       | 1       | 1           | 1              |
| 41      | 1            | 1       | 1       | 1       | 1           | 1              |
| 42      | 2            | 2       | 1       | 2       | 2           | 2              |
| 43      | 2            | 2       | 2       | 2       | 2           | 2              |
| 44      | 2            | 2       | 1       | 2       | 2           | 2              |
| 45      | 2            | 2       | 2       | 2       | 2           | 2              |
| 46      | 2            | 2       | 2       | 2       | 2           | 2              |
| 47      | 2            | 2       | 2       | 2       | 2           | 2              |
| 48      | 2            | 1       | 2       | 1       | 1           | 1              |
| 49      | 2            | 1       | 1       | 1       | 1           | 1              |
| 50      | 2            | 2       | 2       | 2       | 2           | 2              |
| 51      | 2            | 2       | 2       | 2       | 2           | 2              |
| 52      | 2            | 2       | 1       | 2       | 2           | 2              |
| 53      | 2            | 1       | 1       | 1       | 1           | 1              |
| 54      | 2            | 1       | 2       | 2       | 2           | 2              |
| 55      | 2            | 1       | 1       | 2       | 1           | 1              |
| 56      | 2            | 2       | 1       | 1       | 2           | 2              |
| 57      | 2            | 2       | 1       | 2       | 2           | 2              |
| 58      | 2            | 2       | 1       | 2       | 2           | 2              |
| 59      | 2            | 2       | 1       | 2       | 2           | 2              |
| 60      | 2            | 2       | 2       | 2       | 2           | 2              |
| 61      | 2            | 2       | 2       | 1       | 1           | 1              |
| 62      | 2            | 2       | 2       | 2       | 2           | 2              |
| 63      | 2            | 1       | 1       | 1       | 1           | 1              |
| 64      | 2            | 2       | 1       | 2       | 2           | 2              |
| 65      | 2            | 2       | 1       | 2       | 2           | 2              |
| 66      | 2            | 2       | 2       | 2       | 2           | 2              |
| 67      | 2            | 1       | 1       | 1       | 1           | 1              |
| 68      | 2            | 2       | 2       | 2       | 2           | 2              |
| 69      | 2            | 2       | 2       | 2       | 2           | 2              |
| 70      | 2            | 1       | 1       | 1       | 1           | 1              |
| 71      | 2            | 2       | 2       | 2       | 2           | 2              |
| 72      | 2            | 1       | 2       | 2       | 2           | 2              |
Table S-3. Performance characteristics of the classification models.

| Metric                        | Equation                                                                 | Reference |
|-------------------------------|--------------------------------------------------------------------------|-----------|
| Sensitivity                   | $Sens = \frac{TP}{TP + FN}$                                             | (S-7)     |
| Specificity                   | $Spec = \frac{TN}{FP + TN}$                                             | (S-8)     |
| Accuracy                      | $Acc = \frac{TP + TN}{TP + FN + FP + TN}$                               | (S-9)     |
| Non-error rate                | $NER = \sum_{c=1}^{C} Sens_C$                                           | (S-10)    |
| Error rate                    | $ER = 1 - NER$                                                           | (S-11)    |
| False-positive rate           | $FPR = \frac{FP}{FP + TN}$                                              | (S-12)    |
| False-negative rate           | $FNR = \frac{FN}{FN + TP}$                                              | (S-13)    |
| Precision                     | $Prec = \frac{TP}{TP + FP}$                                             | (S-14)    |
| Efficiency rate               | $EFR = 1 - (FPR + FNR)$                                                 | (S-15)    |
| Mathews correlation coefficient (MCC) | $MCC = \frac{TP/N - S \cdot P}{\sqrt{PS(1 - S)(1 - P)}}$ | (S-16) |

Where:

$N = TN + TP + FN + FP$

$S = \frac{TP + FN}{N}$

$P = \frac{TP + FP}{N}$
Figure S-1. Patient age (years) histogram. (A) Positive RT-qPCR diagnostic; (B) Negative RT-qPCR diagnostic.
Figure S-2. Target genes from RT-PCR analysis histograms. (A) Gene N; (B) Gene ORF1ab.
Figure S-3. MIR spectral dataset from saliva samples of $n=237$ patients with RT-qPCR diagnoses for COVID-19 infection. A) positive ($n=138$ samples), and B) negative ($n=99$ samples).
Figure S-4. Mean spectrum of Positive samples (red) and Negative samples (blue) with main MIR band assignments.
Figure S-5. Original MIR spectra dataset from saliva samples (n=265) with the mean spectrum (red line) and with different evaluated baseline correction approaches: A) raw spectra B) Spectra after baseline corrected via the airPLS algorithm; C) Spectra after first derivative pre-processing applied to baseline correction; D) Spectra after SNV pre-processing and Savitzky-Golay smoothing applied to baseline correction.
Figure S-6. Mean derived spectrum of Positive samples (blue) and Negative samples (red) with identifying the higher frequency in URF model (dotted line).
Figure S-7. MIR truncated spectra (after baseline correction) of 237 samples (A) with a mean spectrum (red line), and MIR processed spectra via second derivative with a mean spectrum (red line).
Figure S-8. Actual and consensus predicted classes for $n=165$ saliva samples in training dataset (A), and for $n=72$ saliva samples in test dataset (B).
Table S-4. Confusion matrix of classification models from training and test dataset.

|       | GA-LDA | SPA-LDA | PLS-DA | PSO-PLSDA |
|-------|--------|---------|--------|-----------|
|       | Cl.1   | Cl.2    | Cl.1   | Cl.2      | Cl.1   | Cl.2    | Cl.1   | Cl.2    |
| TP    |        |         |        |           |        |         |        |         |
| Train | 84     | 46      | 67     | 43        | 68     | 52      | 77     | 52      |
| Test  | 39     | 22      | 36     | 21        | 31     | 23      | 34     | 23      |
| TN    |        |         |        |           |        |         |        |         |
| Train | 46     | 84      | 43     | 67        | 52     | 68      | 52     | 77      |
| Test  | 22     | 39      | 21     | 36        | 23     | 31      | 23     | 34      |
| FP    |        |         |        |           |        |         |        |         |
| Train | 22     | 13      | 25     | 30        | 16     | 29      | 16     | 20      |
| Test  | 9      | 2       | 10     | 5         | 8      | 10      | 8      | 7       |
| FN    |        |         |        |           |        |         |        |         |
| Train | 13     | 22      | 30     | 25        | 29     | 16      | 20     | 16      |
| Test  | 2      | 9       | 5      | 10        | 10     | 8       | 7      | 8       |

TP = True positive; TN = True negative; FP = False positive; FN = False negative; Cl.1 = Class 1; Cl.2 = Class 2.
Table S-5. Cross-table of $\chi^2$ test symptoms (according to Centers for Disease Control and Prevention, CDC) of participant cohorts dataset.

|   | Symptom                  | Positive | Negative | $\chi^2_{calc}$ | $\chi^2_{Tab(\alpha=0.05)}$ | p-value   |
|---|--------------------------|----------|----------|------------------|-------------------------------|-----------|
| 1 | **Fever**                | **No**   | 60       | 62               | 8,461                         | 3,841     | **0.0036** |
|   | **Yes**                  | 78       | 37       |                  |                               |           |           |
| 2 | **Dyspnea** (shortness of breath) | **No**   | 104      | 73               | 0.0805                        | 3,841     | 0.7766    |
|   | **Yes**                  | 34       | 26       |                  |                               |           |           |
| 3 | **Saturation O₂ < 95%**  | **No**   | 130      | 97               | 2.0347                        | 3,841     | 0.1537    |
|   | **Yes**                  | 8        | 2        |                  |                               |           |           |
| 4 | **Cough**                | **No**   | 53       | 34               | 0.4095                        | 3,841     | 0.5222    |
|   | **Yes**                  | 85       | 65       |                  |                               |           |           |
| 5 | **Runny nose**           | **No**   | 83       | 44               | 5,7134                        | 3,841     | **0.0168** |
|   | **Yes**                  | 55       | 55       |                  |                               |           |           |
| 6 | **Sore throat**          | **No**   | 75       | 54               | $9,077.10^{-4}$               | 3,841     | 0.9760    |
|   | **Yes**                  | 63       | 45       |                  |                               |           |           |
| 7 | **Diarrhea**             | **No**   | 117      | 78               | 1.420                         | 3,841     | 0.2333    |
|   | **Yes**                  | 21       | 21       |                  |                               |           |           |
| 8 | **Nausea or vomiting**   | **No**   | 121      | 85               | 0.1684                        | 3,841     | 0.6815    |
|   | **Yes**                  | 17       | 14       |                  |                               |           |           |
| 9 | **Headache**             | **No**   | 63       | 38               | 1.245                         | 3,841     | 0.2645    |
|   | **Yes**                  | 75       | 61       |                  |                               |           |           |
| 10| **Fatigue** (tiredness or weakness) | **No**   | 110      | 77               | 0.1293                        | 3,841     | 0.9760    |
|   | **Yes**                  | 28       | 22       |                  |                               |           |           |
| 11| **Loss of smell**        | **No**   | 104      | 88               | 6,856                         | 3,841     | **0.0088** |
|   | **Yes**                  | 34       | 11       |                  |                               |           |           |
**Figure S-9.** Boxplot of days of first symptoms from dataset ($n=237$).

**Figure S-10.** Graph of actual and predicted classes by the consensus class and days of first symptoms from dataset ($n=237$). (A) Positive class markers; (B) Negative class markers.
Table S-6. Quality parameters of prediction new dataset (n=59) from applied models.

| Model       | SET | TRAIN | OUTLIER | VAL | OUTLIER | VL/P/C | VARIABLES | SENS | SPEC | NER | ER | NOT ASSIGNED | VPP | VPN | TFP | TFN | EXAT | MCC |
|-------------|-----|-------|---------|-----|---------|--------|-----------|-------|------|-----|----|--------------|-----|-----|-----|-----|------|-----|
| GA-LDA      | 246 | 172   | 7       | 72  | 2       | 0      | 34        | 76%  | 32%  | 54% | 46%| 0%           | 56% | 54% | 68% | 24% | 55%  | 0.08|
| PLSDA       | 246 | 172   | 7       | 74  | 2       | 7      | 464       | 70%  | 48%  | 59% | 41%| 0%           | 57% | 61% | 52% | 30% | 59%  | 0.18|
| PSO-PLS-DA  | 246 | 172   | 7       | 74  | 2       | 9      | 45        | 77%  | 48%  | 62% | 38%| 0%           | 63% | 65% | 52% | 23% | 63%  | 0.26|
| Consensus class | ----| ---- | ----     | ----| ----    | ----   | ----      | 70%  | 41%  | 56% | 44%| 0%           | 58% | 56% | 58% | 29% | 57%  | 0.13|
Fig. S-11. Histograms of distribution F1 score class 1 Train and Test datasets from permuted classification models (blue bars), with the F1 score class 1 train and Test datasets from original classification models (red vertical line).
Supplementary References

(1) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:101093340.

(2) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24* (2), 123–140. https://doi.org/10.1007/bf00058655.

(3) Lovatti, B. P. O.; Nascimento, M. H. C.; Neto, Á. C.; Castro, E. V. R.; Filgueiras, P. R. Use of Random Forest in the Identification of Important Variables. *Microchem. J.* **2019**, *145* (December 2018), 1129–1134. https://doi.org/10.1016/j.microc.2018.12.028.

(4) Oliveira, B. P. L.; Nascimento, M. H. C.; Rainha, K. P.; Alvaro, E. C. S. O.; Filgueiras, P. R. Different Strategies for the Use of Random Forest in NMR Spectra. *J. Chemometrics* **2020**, No. Special issue, 1–10. https://doi.org/10.1002/cem.3231.

(5) Afanador, N. L.; Smolinska, A.; Tran, T. N.; Blanchet, L. Unsupervised Random Forest: A Tutorial with Case Studies. *J. Chemometrics* **2016**, *30* (5), 232–241. https://doi.org/10.1002/cem.2790.

(6) 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. *Chemometrics Intell. Lab. Syst.* **2017**, *162*, 123–129. https://doi.org/10.1016/j.chemolab.2017.01.021.

(7) Shaffer, R. E.; Small, G. W. Genetic Algorithms for the Optimization of Piecewise Linear Discriminants. *Chemometrics Intell. Lab. Syst.* **1996**, *35*, 87–104.
(8) Pontes, M. J. C.; Galvão, R. K. H.; Araújo, M. C. U. The Sucesive Projections Algorithm for Spectral Variable Selection in Classification Problems. *Chemometrics Intell. Lab. Syst.* **2005**, *78* (1–2), 11–18.
https://doi.org/10.1016/j.chemolab.2004.12.001.

(9) Marques, A. S.; Castro, J. N. F.; Costa, F. J. M. D.; Neto, R. M.; Lima, K. M. G. Near-Infrared Spectroscopy and Variable Selection Techniques to Discriminate Pseudomonas Aeruginosa Strains in Clinical Samples. *Microchem. J.* **2016**, *124*, 306–310. https://doi.org/10.1016/j.microc.2015.09.006.

(10) Marques, A. de S.; Melo, M. C. N. de; Cidral, T. A.; Lima, K. M. G. de. Feature Selection Strategies for Identification of Staphylococcus Aureus Recovered in Blood Cultures Using FT-IR Spectroscopy Successive Projections Algorithm for Variable Selection: A Case Study. *J. Microbiol. Methods.* **2014**, *98* (1), 26–30. https://doi.org/10.1016/j.mimet.2013.12.015.

(11) Araújo, M. C. U.; Saldanha, teresa C. B.; Galvão, R. K. H.; Yoneyama, T.; Chame, H. C.; Visani, V. The Sucesive Projections Algorithm for Variable Selection in Spectroscopic Multicomponent Analysis. *Chemometrics Intell. Lab. Syst.* **2001**, *57* (2), 65–73.

(12) Marini, F.; Walczak, B. Particle Swarm Optimization (PSO). A Tutorial. *Chemometrics Intell. Lab. Syst.* **2015**, *149*, 153–165.
https://doi.org/10.1016/j.chemolab.2015.08.020.

(13) Brereton, R. G.; Lloyd, G. R. Partial Least Squares Discriminant Analysis: Taking the Magic Away. *J. Chemometrics* **2014**, *28* (4), 213–225.
https://doi.org/10.1002/cem.2609.
(14) Ballabio, D.; Consonni, V. Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA. *Anal. Methods* **2013**, *5* (16), 3790–3798. https://doi.org/10.1039/c3ay40582f.

(15) Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemometrics* **2003**, *17*, 166–173. https://doi.org/10.1002/cem.785.

(16) Chen, L. F.; Liao, H. Y. M.; Ko, M. T.; Lin, J. C.; Yu, G. J. A New LDA-Based Face Recognition System Which Can Solve the Small Sample Size Problem. *Pattern Recognit.* **2000**, *33*, 1713–1726.

(17) Martõ Ânez, A. M.; Kak, A. C. PCA versus LDA. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *2* (2), 228–233. https://doi.org/10.1109/34.908974.