nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-0644-7

Detection of SARS-CoV-2 in nasal swabs using MALDI-MS

Fabiane M. Nachtigall[1]✉, Alfredo Pereira[2], Oleksandra S. Trofymchuk[3] and Leonardo S. Santos[2]✉

Detection of SARS-CoV-2 using RT–PCR and other advanced methods can achieve high accuracy. However, their application is limited in countries that lack sufficient resources to handle large-scale testing during the COVID-19 pandemic. Here, we describe a method to detect SARS-CoV-2 in nasal swabs using matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) and machine learning analysis. This approach uses equipment and expertise commonly found in clinical laboratories in developing countries. We obtained mass spectra from a total of 362 samples (211 SARS-CoV-2-positive and 151 negative by RT–PCR) without prior sample preparation from three different laboratories. We tested two feature selection methods and six machine learning approaches to identify the top performing analysis approaches and determine the accuracy of SARS-CoV-2 detection. The support vector machine model provided the highest accuracy (93.9%), with 7% false positives and 5% false negatives. Our results suggest that MALDI-MS and machine learning analysis can be used to reliably detect SARS-CoV-2 in nasal swab samples.

The outbreak of coronavirus disease 2019 (COVID-19) is a crisis that affects rich and poor countries alike[1]. Detection of SARS-CoV-2 in patient samples is a critical tool for monitoring spread of the disease, guiding therapeutic decisions and devising social distancing protocols[2]. Detection assays based on RT–PCR are the most effective and sensitive method for diagnosis of SARS-CoV-2 infection and are used in laboratories around the world[3]. However, some countries lack the laboratory resources and access to PCR kits to conduct testing at the required levels. Therefore, other reliable diagnostic techniques are needed. Most clinical diagnostic laboratories have MALDI-MS equipment, which is used to identify bacterial and fungal infections. We propose to leverage the ease-of-use and robustness of MALDI-MS pathogen identification for large-scale SARS-CoV-2 testing in developing countries.

MALDI-MS-based assays rely on reference spectra of strains and bioinformatics for high-sensitivity and high-specificity species identification through proteomic profiling. This approach is well established and accepted in many countries for routine diagnostics of yeast and bacterial infections. However, no spectral libraries for SARS-CoV-2 identification using MALDI-MS are publicly available to our knowledge.

We first acquired MALDI mass spectra of nasal swab samples that had been tested for SARS-CoV-2 by RT–PCR and analyzed them using machine learning (ML). In this experiment (Fig. 1a), a total of 362 samples (211 SARS-CoV-2-positive and 151 negative, unequivocally confirmed by PCR), which came from three different countries, Argentina (Lab 1), Chile (Lab 2) and Peru (Lab 3), were placed on the MALDI plate without prior sample purification.

All spectra from the two groups under study were preprocessed, and peak detection was applied to obtain an intensity matrix (Fig. 1b). FS methods were used to select the most characteristic peaks distinguishing SARS-CoV-2-positive from SARS-CoV-2-negative samples. These data were then used for principal component analysis (PCA), where we could explore and compare the spectra in multidimensional space using peaks selected by FS methods. Finally, FS methods together with six different ML algorithms (decision tree, DT; *k*-nearest neighbors, KNN; naive Bayes, NB; random forest, RF; support vector machine with a linear kernel, SVM-L; support vector machine with a radial kernel, SVM-R) allowed us with a certainty of 0.947 sensitivity and 0.926 specificity to distinguish samples of the control group from patients with COVID-19 (see the description below for ML).

For the purposes of quality and reproducibility of the analysis, the MALDI-MS data must be standardized for inter-laboratory applications to afford reliable spectra. For each sample, at least 500 individual spectra (50 laser shots at 10 different spot positions) were accumulated and averaged. The parameters that must be optimized to achieve reliable and high-quality spectra include base peak intensity, signal-to-noise ratio, base peak minimum resolution, external calibration, preprocessed smoothing, baseline subtraction and internal peak alignment. The most important for typing is the internal recalibration that allows the reliable comparison of spectra. The 362 samples (Fig. 2a) under study were processed together, while Fig. 2b and Fig. 2c show a comparison of the SARS-CoV-2-positive and -negative mean spectra. With this approach, we established a mass analysis range between 3 and 15.5 kDa and a peak detection process, which resulted in an intensity matrix of 88 peaks (Supplementary Data 1–3). To identify peaks that differentiate SARS-CoV-2-positive from -negative samples, we applied a two-tailed Wilcoxon rank sum test (with false discovery rate corrections for multiple hypothesis testing), which afforded 31 peaks (Supplementary Table 1) with a P value < 0.05. Figure 2d shows the intensity comparison of the first six selected peaks together with one of the most relevant peaks (mass-to-charge ratio (m/z) 7,612). Differences in peak intensities between mean preprocessed MALDI-MS spectra are shown in Fig. 2e,f. The most substantial intensity difference was exhibited by the peak with m/z of 3,358 (Fig. 2e).

Next, two FS methods were used for peak selection: information gain-based (Ig) FS[4] and correlation-based FS (Cfs)[5]. Both methods were applied to all spectra, as well as separately to the spectra obtained from each laboratory. Of note, when selected peaks were compared between laboratories, only the m/z 7,612 peak was common to all. The difference in the intensities of this peak for SARS-CoV-2-positive and -negative samples is presented in Fig. 2f (selected Ig FS and Cfs peaks for each laboratory are presented in Supplementary Table 2).

[1]Instituto de Ciencias Químicas Aplicadas, Universidad Autónoma de Chile, Talca, Chile. [2]Instituto de Química de Recursos Naturales, Universidad de Talca, Talca, Chile. [3]Núcleo Científico Multidisciplinario-DI, Universidad de Talca, Talca, Chile. ✉e-mail: fabiane.manke@uautonoma.cl; lssantos@utalca.cl

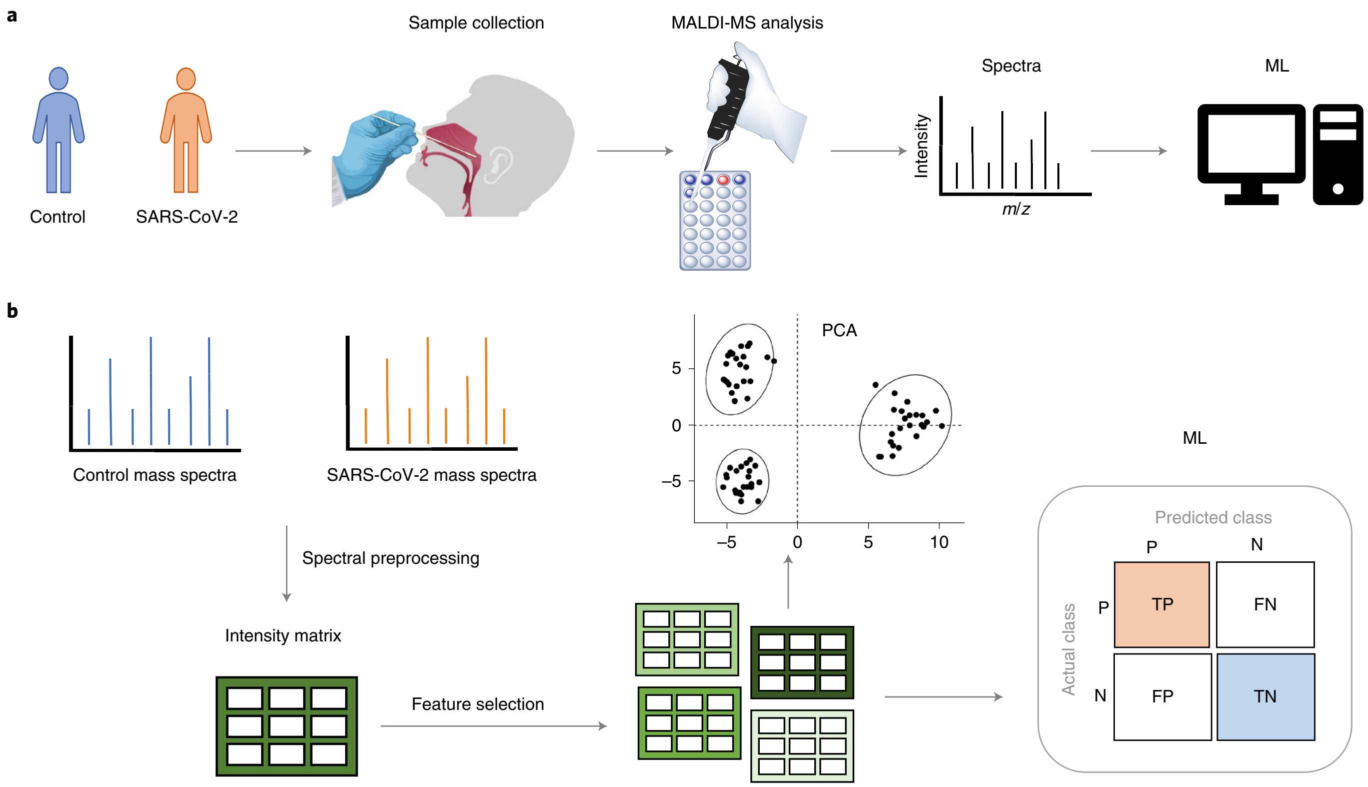


Fig. 1 | Coronavirus detection in nasal mucous secretion. a, A general scheme for SARS-CoV-2 detection by MALDI-MS and ML. The same samples of nasal mucous secretion swabs employed for PCR tests were used for MALDI mass spectra acquisition and the obtained spectra were analyzed using ML methods. **b**, Prediction of SARS-CoV-2 using ML. Spectral preprocessing was performed on the spectra obtained from SARS-CoV-2-positive and SARS-CoV-2-negative groups of patients. FS was applied to the obtained intensity matrices, and then PCA and ML were performed.

We then used PCA[6] to explore and compare SARS-CoV-2-positive and -negative samples (all samples and each laboratory separately) in a multidimensional space using all 88 peaks and selected peaks for FS methods. PCA for Cfs peaks is shown in Fig. 2g–j, and PCA for Ig peaks and all 88 peaks is shown in Supplementary Fig. 1. When PCA was performed with samples from the three laboratories, the data did not completely separate (Fig. 2g); however, when samples from each country were handled independently, the SARS-CoV-2-positive group was well separated from the control group (Fig. 2h–j). For example, for samples from Lab 1, it was possible to show a better separation between groups (Fig. 2h).

Using data from Lab 2 and Lab 3 samples we could visualize better data separation into two clusters, as depicted in Fig. 2i,j. Although the Cfs method used together with PCA helps to interpret and compare the data under study, complete sample separation was not achieved, especially when samples from all three laboratories were used, and therefore more advanced methods are required to discriminate SARS-CoV-2-positive and -negative samples.

Next, we applied ML approaches to different sets of peaks to detect SARS-CoV-2 in a smaller set of 80 SARS-CoV-2-positive and -negative samples ($n=40$ per group, data from Lab 2 and Lab 3). We used two different FS methods (Ig FS and Cfs; Supplementary Table 3) or no FS method. They were evaluated using six different ML methods (DT, NB, KNN, RF, SVM-L and SVM-R), thus affording a total number of 18 trained and tested ML models (Methods). Fourfold (outer) nested repeated (five times) tenfold (inner) cross-validation was used to train and test ML models and the hyperparameters of each algorithm were optimized (Supplementary Table 4).

A receiver operating characteristic (ROC) curve indicates the relationship of the true-positive rate (TPR) and the false-positive rate (FPR) and takes the uncertainty of each prediction into account when evaluating the performance of a ML model (Fig. 3a–c)[7]. The precision–recall (PR) curves for the six different ML algorithms are shown in Fig. 3d–f. The area under the ROC curve is equal to the probability that a classifier sorts a randomly selected positive sample higher than a randomly selected negative one[8]. The SVM-R model demonstrates the highest area under curve (AUC) values of 0.98 and 0.95 for ROC and PR curves, respectively (among six ML algorithms and with no FS used), indicating the highest true-positive numbers.

Performance analysis demonstrates that the models' sensitivities, specificities and accuracies do not vary substantially among the ML methods (Fig. 3g–i). Here, SVM-R with no FS proved to be the best method, detecting 36/40 SARS-CoV-2-positive samples and 36/40 SARS-CoV-2-negative samples (Fig. 3j). It is worth mentioning that NB, RF and SVM-L achieved high performance metrics as well, while DT and KNN showed lower performance.

Finally, we tested the ML approach for detection of SARS-CoV-2-positive samples among all 362 samples from all 3 laboratories using the SVM-R ML algorithm, together with two FS methods (Ig and Cfs) and a model without FS (Fig. 4). Each model was trained and tested with fourfold (outer) nested repeated (five times) tenfold (inner) cross-validation. In the inner cross-validation, the FS methods were applied (peaks in Supplementary Table 5) and the hyperparameters of each algorithm were optimized (Supplementary Table 6). To test each of the trained models, these were evaluated in four independent datasets (outer cross-validation). This stage allows evaluation of the models' robustness and demonstration of

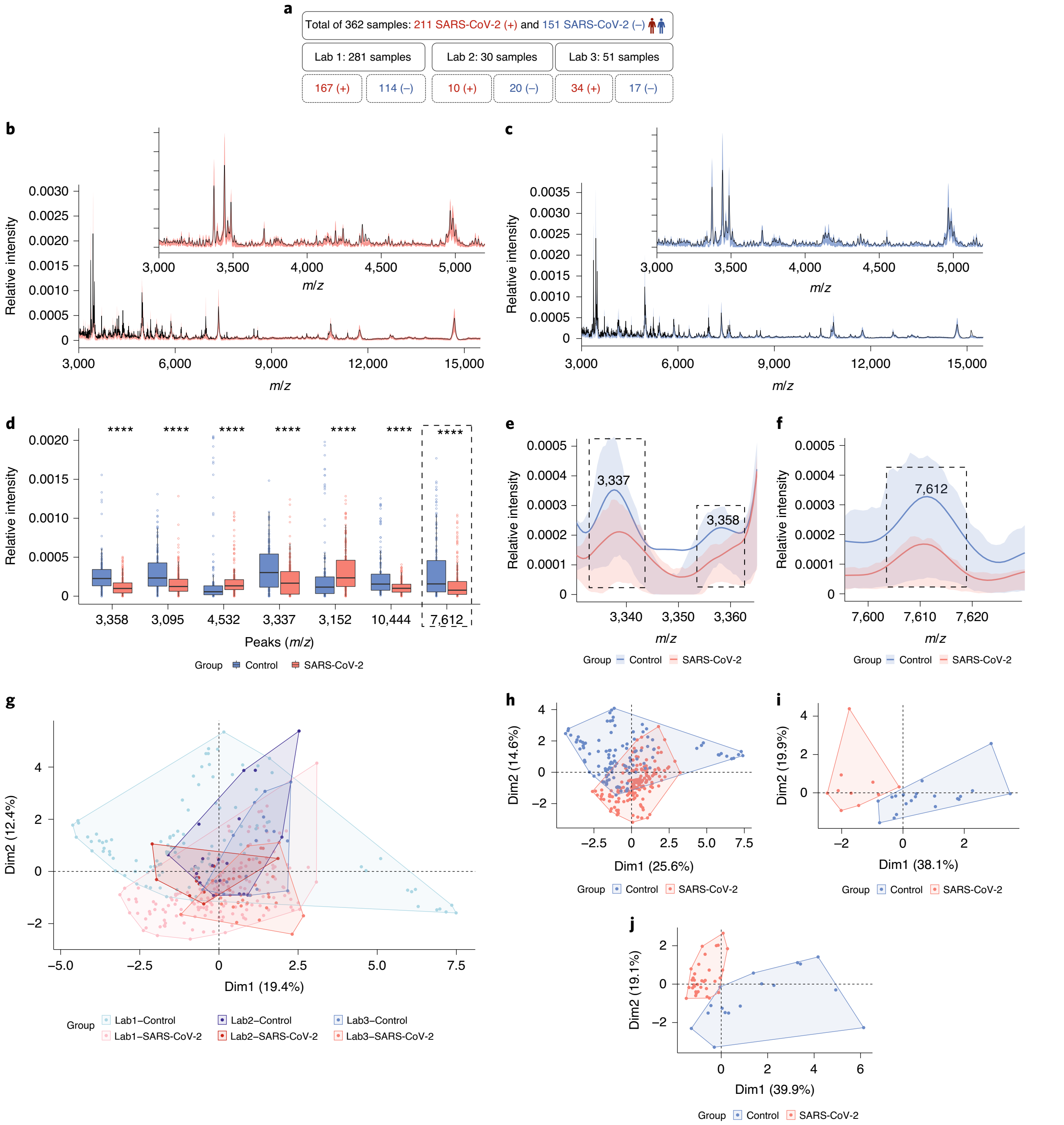


Fig. 2 | Comparative analysis of SARS-CoV-2-positive and -negative samples. a, Summary of samples by country. **b**, Mean MALDI-MS spectra and the respective interquartile range (IQR) (denoted by orange, see the inset) processed with the MALDIquant R package, obtained from a nasal mucous secretion in the SARS-CoV-2 group. **c**, Mean MALDI-MS spectra and the respective IQR (denoted by blue, see the inset) processed with the MALDIquant R package, obtained from a nasal mucous secretion in the control group. **d**, Intensity comparison of seven selected peaks of distinct *m/z* in the dataset under study (the SARS-CoV-2 group is denoted as orange and the control group is denoted as blue). One of the most relevant peaks (of *m/z* 7,612) is highlighted by a dashed box. The boxes show the 25th–75th percentile with the median; the whiskers correspond to 1.5 × IQR. $n = 362$ independent samples (211 SARS-CoV-2-positive samples and 151 SARS-CoV-2-negative) were used. ****$P < 0.0001$ by two-tailed Wilcoxon rank sum test. Exact *P* values for each peak are presented in Supplementary Table 1. **e**, Comparison of the intensities of one of the most relevant peaks (of *m/z* 3,358) to differentiate the control group from the SARS-CoV-2 group; data are presented as mean values (line) ± IQR (shadow). $n = 362$ independent samples (211 SARS-CoV-2-positive samples and 151 SARS-CoV-2-negative) were used. **f**, Comparison of the intensities of one of the most relevant peaks (of *m/z* 7,612) to differentiate the control group from the SARS-CoV-2 group; data are presented as mean values (line) ± IQR (shadow). $n = 362$ independent samples (211 SARS-CoV-2-positive samples and 151 SARS-CoV-2-negative) were used. **g**, PCA of the mass spectra of the SARS-CoV-2 and control samples from three countries using peaks selected with the Cfs method. **h**, PCA of the mass spectra of the SARS-CoV-2 and control samples from Lab 1 using peaks selected with the Cfs method. **i**, PCA of the mass spectra of the SARS-CoV-2 and control samples from Lab 2 using peaks selected with the Cfs method. **j**, PCA of the mass spectra of the SARS-CoV-2 and control samples from Lab 3 using peaks selected with the Cfs method.

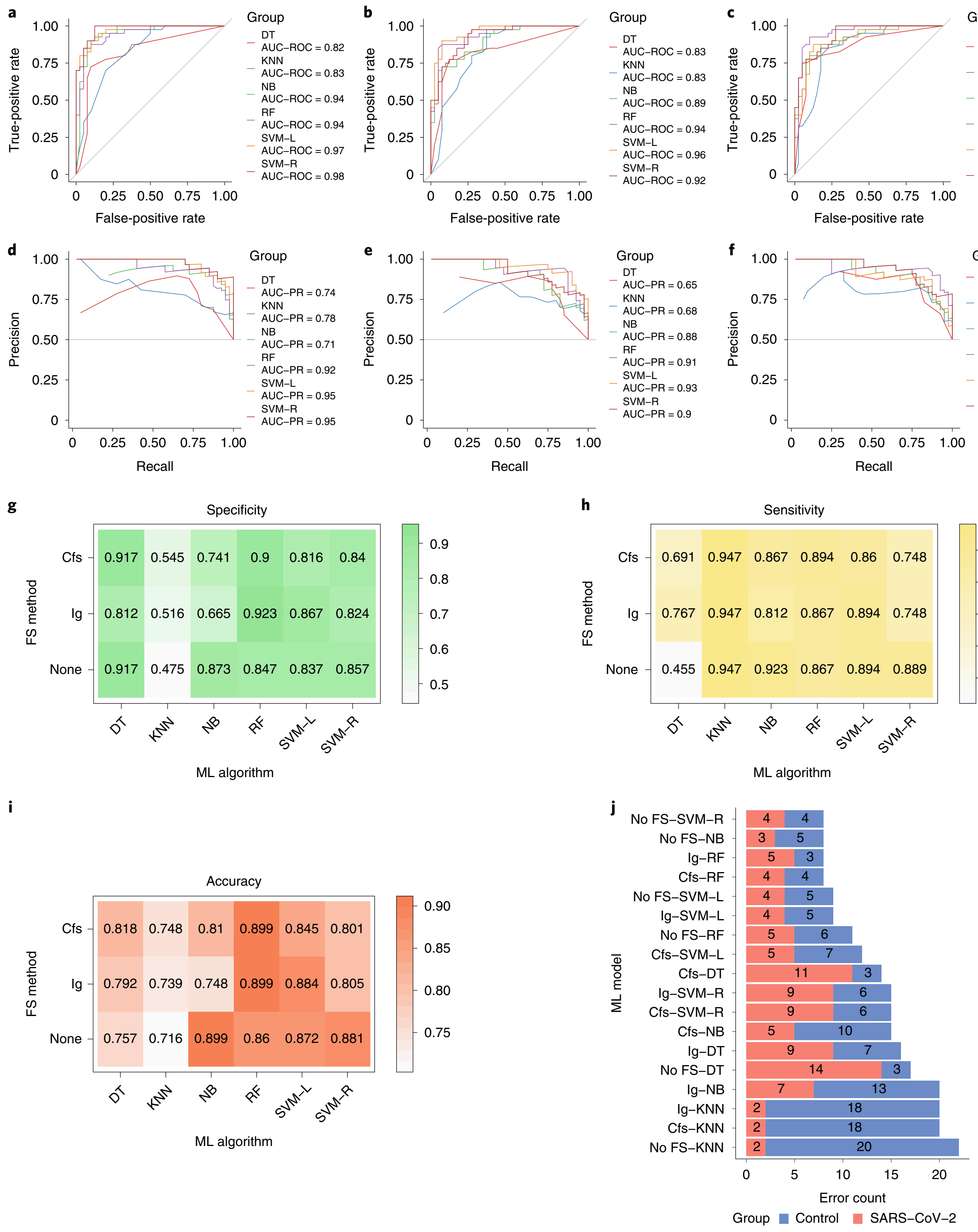
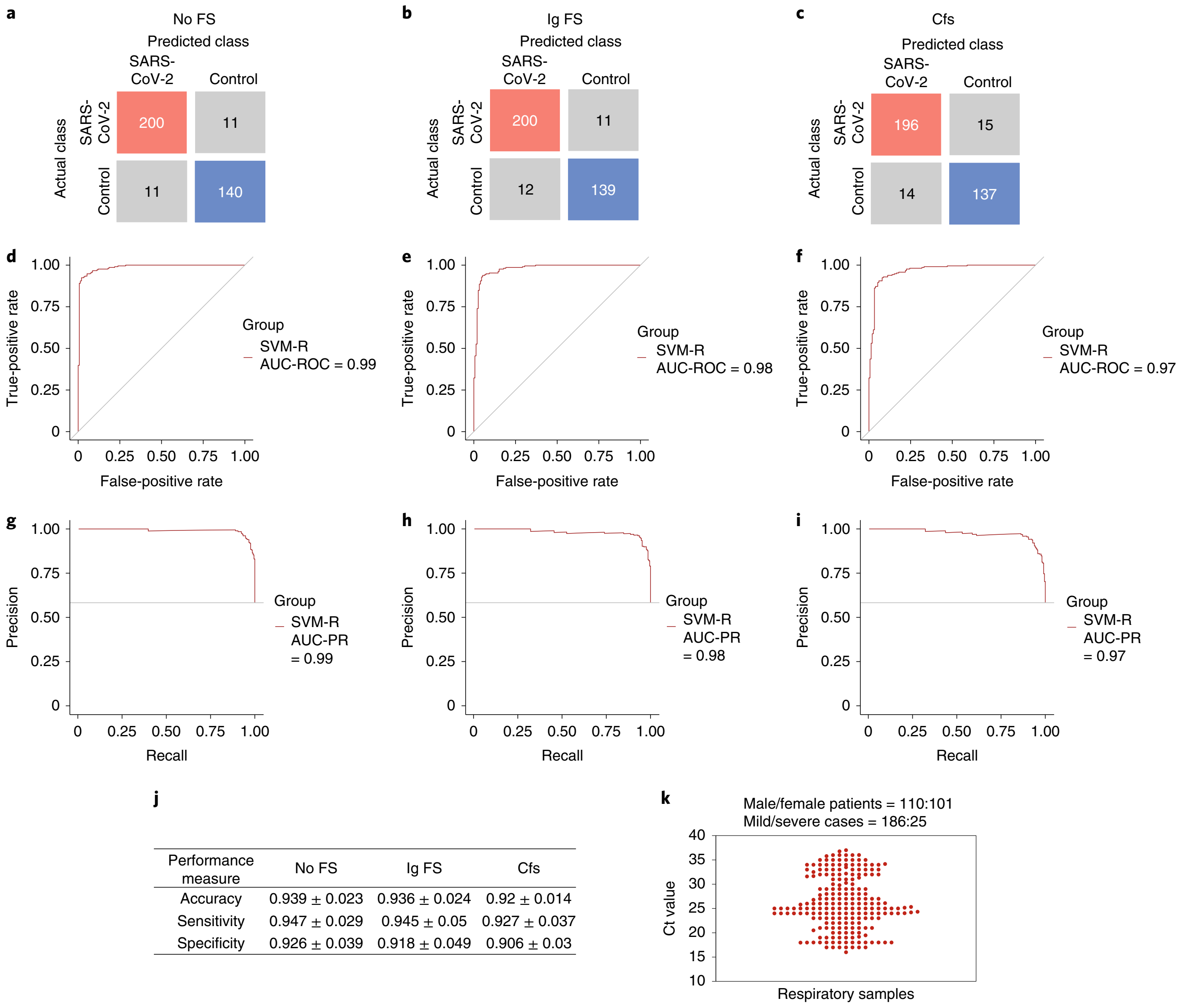


Fig. 3 | Results of testing of 18 ML models. Each ML model is the combination of an FS method and an ML algorithm (40 SARS-CoV-2 samples and 40 control samples were used). **a**, ROC curves (TPR-FPR) for no FS and six ML algorithms. **b**, ROC curves (TPR-FPR) for Ig FS and six ML algorithms. **c**, ROC curves (TPR-FPR) for Cfs FS and six ML algorithms. **d**, PR curves for no FS and six ML algorithms. **e**, PR curves for Ig FS and six ML algorithms. **f**, PR curves for Cfs FS and six ML algorithms. **g**, Summary of the specificities obtained for each model. **h**, Summary of the sensitivities obtained for each model. **i**, Summary of the accuracies obtained for each model. **j**, The error count for each ML model (SARS-CoV-2 errors are denoted in orange and control errors are in blue).



Performance measure	No FS	Ig FS	Cfs
Accuracy	0.939 ± 0.023	0.936 ± 0.024	0.92 ± 0.014
Sensitivity	0.947 ± 0.029	0.945 ± 0.05	0.927 ± 0.037
Specificity	0.926 ± 0.039	0.918 ± 0.049	0.906 ± 0.03

Fig. 4 | Performance metrics from fourfold nested cross-validation with 362 samples (inter-laboratory detection). **a**, A confusion matrix of the SVM-R model without FS. **b**, A confusion matrix of the SVM-R model with Ig FS. **c**, A confusion matrix of the SVM-R model with Cfs. **d**, The ROC curve for the model with all peaks (no FS). **e**, The ROC curve for the model with peaks selected with Ig FS. **f**, The ROC curve for the model with peaks selected with Cfs. **g**, The PR curve for the model with all peaks (no FS). **h**, The PR curve for the model with peaks selected with Ig FS. **i**, The PR curve for the model with peaks selected with Cfs. **j**, Performance measures for inter-laboratory detection. **k**, Ct values of all RT–PCR-positive sample distribution evaluated by MALDI-MS/ML confusion matrices.

their possible application for SARS-CoV-2 detection (Fig. 4). In addition, this testing approach can demonstrate whether a set of attributes from one group of patients can be used to determine the SARS-CoV-2 status in others.

Figure 4a–c presents the general confusion matrix (outer cross-validation) obtained from different models. Using no FS, 340 of 362 samples were correctly classified (Fig. 4a), achieving 0.939 ± 0.023 accuracy (Fig. 4j). Ig FS reached 0.936 ± 0.024 accuracy, with 339 of 362 samples correctly classified (Fig. 4b). When Cfs was used, 0.92 ± 0.014 accuracy was obtained, with 333 correctly classified samples. As can be seen from the ROC curves for these three approaches (no FS, Ig and Cfs, Fig. 4d–i), using SVM-R the AUC values are 0.99, 0.98 and 0.97, respectively (Fig. 4d–f). The PR curves were used as an additional indicator to assess model prediction performance (Fig. 4g–i). SVM-R achieved an AUC of 0.99 (no FS; Fig. 4g), 0.98 (Ig FS; Fig. 4h) and 0.97 (Cfs; Fig. 4i). The analysis of the ROC and PR curves indicates that it is not necessary to perform FS to have high performance, as the model without FS is highly robust. The metrics presented in Fig. 4j also support the robust performance of the SVM-R algorithm. For example, the model with the highest accuracy is SVM-R with no FS, which also achieves a sensitivity of 0.947 ± 0.029 and a specificity of 0.926 ± 0.039. Class imbalance is a common problem in ML tasks, causing instances of the minority class to be classified as the majority class[9]. It has also been reported that some algorithms tend to suffer more than others[10]. However, our model is imbalance-robust (for comparisons of ML models, see Supplementary Fig. 2).

Our approach could be used to determine SARS-CoV-2 positivity in samples screened by RT–PCR with Ct values ranging from 17 to 37 (Fig. 4k and Supplementary Table 8) from patients seeking medical care following symptoms consistent with COVID-19. Although several studies[11,12] have described the relationship between Ct values and disease severity, in this study we were unable to make such correlations given a lack of data regarding patient characteristics and disease severity.

Although molecular techniques such as RT–PCR and immunochromatography are undoubtedly useful for SARS-CoV-2 detection, MALDI-MS with ML is a promising alternative given its speed, simplicity and low cost, and the availability of equipment and expertise in many hospital laboratories in developing countries. In this work, we implemented this method according to standardized protocols for SARS-CoV-2. Nasal swabs used for RT–PCR tests were useful for acquisition of MALDI mass spectra without purification, and as such, samples testing positive for SARS-CoV-2 by MS could be used in a follow-up confirmation test using the RT–PCR gold standard. Comparing RT–PCR and MALDI/ML results on our set of samples, we found a concordance rate that is acceptable as a clinical diagnostic approach (>80%). We propose a step-by-step diagnostic method that can be implemented through a ranking system combining the most robust ML models, which could be rapidly validated in other laboratories before being adopted as a fast screening assay for SARS-CoV-2.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0644-7.

Received: 12 May 2020; Accepted: 14 July 2020;
Published online: 30 July 2020

References

1. Zhang, C. et al. Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. *J. Proteome Res.* **19**, 1351–1360 (2020).
2. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).
3. Bonetta, L. Prime time for real-time PCR. *Nat. Methods* **2**, 305–312 (2005).
4. Appavu, S., Rajaram, R., Nagammai, M., Priyanga, N. & Priyanka, S. Bayes theorem and information gain based feature selection for maximizing the performance of classifiers. in *Advances in Computer Science and Information Technology* (eds Meghanathan N., Kaushik B. K. & Nagamalai D.) *Communications in Computer and Information Science* Vol. 131 (Springer, 2011).
5. Hall, M. A. Correlation-based feature selection for discrete and numeric class machine learning. in *Proceedings of Seventeenth International Conference on Machine Learning (ICML)* 359–366 (Morgan Kaufmann, 2000).
6. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 37–52 (1987).
7. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* (Wiley, 1966).
8. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
9. Haixiang, G. et al. Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* **73**, 220–239 (2017).
10. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
11. Zheng, S. et al. Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January–March 2020: retrospective cohort study. *Br. Med. J.* **369**, m1443 (2020).
12. Wishaupt, J. O. et al. Pitfalls in interpretation of CT-values of RT-PCR in children with acute respiratory tract infections. *J. Clin. Virol.* **90**, 1–6 (2017).

Methods

Data reporting. No statistical methods were used to determine the patient sample size—all 362 RT–PCR swab samples available were analyzed by MALDI-MS and employed to develop the ML platform of analysis. The investigators were blinded to discrimination during experiments and outcome assessment for the samples from Argentina.

Ethical and biosafety statement. The work followed the ethical protocols approved by the Hospital Regional de Talca Commission of Ethics and no identity of any sample was related to the name of the patient or other information that could lead to personal identification. Specific patient demographic information (disease severity, age and comorbidity) was limited during this work. It was not possible obtain according to administrative protocols from clinical laboratories and confidentiality agreements. All samples were spotted on the MALDI plate and were used according to Hospital de Talca Clinical Laboratory protocols. Before analyses in the mass spectrometer, the MALDI plates, after spotting the samples inside a biosecurity cabinet, were irradiated by an ultraviolet lamp for at least 20 min to prevent any source of contamination to the mass spectrometrists.

Sample collection. A total of 362 clinical (211 RT–PCR-positive and 151 RT–PCR-negative with Ct values above 40) nasal mucous secretion samples were obtained from patients suffering or not from symptoms of SARS-CoV-2 (Regional Talca Hospital, Chile; Instituto Nacional, Peru; Reference Respiratory Virus Laboratory INEI-ANLIS "Dr. Carlos G Malbran" in Argentina) during the pandemic outbreaks in March–April of 2020. The swabs were ABS plastic sticks with a flocked nylon tip end (break point at 3 cm and total stick of 15 cm), RNAse/DNAse-free with sterile tubes containing Cary–Blair transport medium. All patients were developing symptoms suggestive of COVID-19 and had a nasopharyngeal swab sample taken for RT–PCR analyses, according to the World Health Organization PCR clinical protocols. In brief, a nasopharyngeal patient swab was collected by inserting a swab into the nostril parallel to the palate. The swab was inserted to a location equidistant from the nostril and the outer opening of ear and was left in place for a few seconds to absorb secretions. The synthetic fiber swab with a plastic shaft was placed immediately into a sterile tube containing 3 ml of viral transport medium. The same samples of nasopharyngeal swab solutions employed for RT–PCR were used directly for MALDI-MS analysis without sample preparation.

MALDI-MS analysis. Mass spectrometric analyses were performed with a MALDI time-of-flight instrument (Autoflex, Bruker) with a pulsed nitrogen laser (337 nm), operating in positive-ion linear mode using a 19 kV acceleration voltage. The matrix solution was prepared with α-ciano-hydroxy-cinnamic acid (CHCA) at 1% in acetonitrile/0.1% trifluoroacetic acid (1:1). One microliter of each nasopharyngeal swab sample (previously used for RT–PCR analysis) was spotted on a MALDI steel plate followed by the addition of 1 µl of the matrix solution (CHCA) and air-drying. Before analyses in the mass spectrometer, the MALDI plates were finally irradiated by an ultraviolet lamp inside a biosecurity cabinet for at least 20 min to prevent any source of contamination to the mass spectrometrists. Spectra were generated by summing 500 single spectra (10 × 50 shots) in the range between 3 and 20 kDa by shooting the laser at random positions on the target spot.

Spectral preprocessing. MALDI-MS fid files (Bruker) were converted to mzML with MSconvert (version 3.0.19039) from the ProteoWizard suit[13], and subsequently preprocessed in R (version 4.0.0) using the MALDIquant[14] and MALDIquantForeign[14,15] packages. All spectra were trimmed to a range from 3 to 15.5 kDa. Square root transformation was applied, and smoothing was realized by the Savitzky–Golay method[16]. The baseline correction was performed using the TopHat algorithm[17], and the intensity was normalized using the total ion current calibration method[18]. To correct the calibration differences between the samples obtained from Lab 2 (Talca) measured with a Bruker Autoflex, and Lab 1 (Argentina) and Lab 3 (Peru) measured with a Shimadzu and Bruker Microflex, respectively, we used the MALDIquant warpMassSpectra command. The applied calibration function was calculated using 14 high-intensity peaks obtained from Lab 1 and Lab 2 (Supplementary Table 8) affording the equation $-(7\times10^{-11})x^3+(2\times10^{-6})x^2+0.973x+51.611$. The selected peaks were manually verified to be the same. Peak detection was carried out applying a signal-to-noise ratio of 2 and a halfWindowSize of 10. Peaks were binned with the binpeaks command with a tolerance of 0.003. To avoid any additional calibration differences, the peak binning was carried out in two stages. First, SARS-CoV-2-positive and control group spectra for each laboratory were separately binned (six spectral subgroups); additionally at this stage peak filtration was performed, keeping only those peaks that were present in 80% of the spectra of each subgroup. Next, all peaks were binned together. The resulting matrix of peak intensities was used for FS, PCA and ML analyses.

Statistical analysis and FS. To compare COVID-19 and control samples, a two-tailed Wilcoxon rank sum test with false discovery rate corrections for multiple hypothesis testing was performed. Previously, sample normality was discarded by a Shapiro–Wilk test, and P values < 0.05 were considered to indicate a significant difference. The number of features is critical to determine the performance of ML-based methodologies. The MALDI-MS data highlight successful FS methods to find the most important peaks and to discriminate two or more different samples. In this work we have applied two different FS methods, Ig FS and Cfs. Both were performed using the FSelector[19] package in R. We have chosen two FS filter methods, since they are faster than wrapper/embedded methods; also they are classifier independent, which is a major advantage when evaluating various algorithms. Ig FS was chosen for being a univariate filter method and Cfs for being a multivariate filter method, which evaluates dependencies between attributes (correlations).

PCA. PCA is an unsupervised ML method to reduce the dimensionality of a dataset comprising many interrelated variables. To explore and compare spectra in multidimensional space, PCA was conducted using the R FactoMineR[20] and factoextra[21] packages (data were scaled to unit variance). PCA was performed on all samples, and for each laboratory separately, using FS methods and without FS.

ML. An initial ML classification was performed with 80 samples to find the best algorithm. Lab 3 samples (34 SARS-CoV-2 and 17 control) together with Lab 2 samples (6 SARS-CoV-2 and 23 control) were used. Six different ML algorithms were chosen to classify SARS-CoV-2 and control samples, namely NB[22], SVM-L[23], SVM-R[23], KNN[24], DT[25] and RF[25]. We implemented 18 ML models: combining no FS, Ig FS and Cfs methods together with 6 ML algorithms. The training and testing stages were carried out through fourfold (outer) nested repeated (five times) tenfold (inner) cross-validation (with randomized stratified splitting) in R with the Caret package. The hyperparameters of each algorithm were optimized in the inner loop by random search among 20 parameters (hyperparameters are reported in Supplementary Table 4). In this way, repeated tenfold cross-validation was performed 20 times and the models obtained with the best results were reported, according to their AUC.

ROC and PR curves were constructed using the MLeval R package, and respective AUCs were calculated. Besides, the overall cross-validation prediction performance was summarized by the accuracy, sensitivity and specificity performance measures. The equations used to quantify these performance measures are presented below (in which TP represents true positives, TN represents true negatives, FP represents false positives and FN represents false negatives):

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{3}$$

Inter-laboratory detection. Inter-laboratory detection of SARS-CoV-2 based on ML was developed with a total of 362 MALDI-MS spectra from 3 laboratories. The training and testing stages were carried out through fourfold (outer) nested repeated (five times) tenfold (inner) cross-validation (with randomized stratified splitting) utilizing a different random sampling each time. Since the best performance measures were obtained in the previous section, SVM-R was used to train and test the ML model using all peaks from the spectral processing section; also we have applied two different FS methods, Ig FS and Cfs, to compare the obtained results. During model training, the hyperparameters of each algorithm were optimized in the inner loop by random search among 20 parameters (hyperparameters are reported in Supplementary Table 7). In summary, each model was trained 20 times by repeated tenfold cross-validation (each time using different parameters). The AUC was estimated each time, and better hyperparameters were chosen for the final model. With fourfold nested cross-validation, four independent test datasets were generated, and then the total confusion matrix and the general ROC curve for the fold were reported. In addition, the mean and standard deviation of the accuracy, sensitivity and specificity were calculated through the folds. ML was carried out in R with the Caret package and ROC and PR curves were constructed using the R package MLeval.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available within the article and its Supplementary Information. Further information about the findings of this study is also available from the corresponding authors, upon reasonable request.

References

13. Gibb, S. & Strimmer, K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* **28**, 2270–2271 (2012).
14. Tammen, H. & Hess, R. in *Peptidomics: Methods and Strategies* (eds Schrader, M. & L. Fricker, L.) 187–196 (Springer, 2018).
15. Gorry, P. A. General least-squares smoothing and differentiation by the convolution (Savitzky–Golay) method. *Anal. Chem.* **62**, 570–573 (1990).
16. Savitzky, A. & Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639 (1964).
17. Hall, M. et al. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10–18 (2009).
18. Romanski, P. & Kotthoff, L. FSelector: Selecting attributes. R package. version 0.31 https://cran.r-project.org/web/packages/FSelector/index.html (2018).
19. Husson, F., Josse, J. & Lê, S. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
20. Kassambara, A. & Mundt, F. factoextra: Extract and visualize the results of multivariate data analyses. R package version 1.0.3 https://CRAN.R-project.org/package=factoextra (2016).
21. Kruse, R. & Borgelt, C. Data Mining with Graphical Models. in *Discovery Science* (eds Lange, S., Satoh, K. & Smith, C. H.) *Lecture Notes in Computer Science* Vol. 2534 (Springer, 2002).
22. Vapnik, V. N. *The Nature of Statistical Learning Theory* (Springer, 1995).
23. Fix, E. & Hodges, J. L.Jr. Discriminatory analysis. Nonparametric discrimination: consistency properties. *Int. Stat. Rev.* **57**, 238–247 (1989).
24. Fakiola, M. et al. Classification and regression tree and spatial analyses reveal geographic heterogeneity in genome wide linkage study of Indian visceral leishmaniasis. *PLoS ONE* **5**, e15807 (2011).
25. Kajdanowicz, T. & Kazienko, P. Boosting-based multi-label classification. *J. Univers. Comput. Sci.* **19**, 502–520 (2013).

Acknowledgements

We acknowledge the efforts of workers from Hospital Regional de Talca (Chile) to provide RT–PCR samples as well as Instituto Nacional de Salud (Peru) and the Reference Respiratory Virus Laboratory INEI-ANLIS "Dr. Carlos G Malbran" (Argentina) for providing access to MALDI-MS measurement data from their RT–PCR samples. This work was funded in part by FONDECYT (Regular number 1180084, L.S.S. and F.M.N.; Iniciación number 11180948, O.S.T.). A.P. thanks the National Agency for Research and Development (ANID) for a PhD grant (number 2016-21161314). We acknowledge Bruker Daltonics (L. F. A. Santos and D. M. Assis, Bruker Brazil) for providing prompt equipment assistance during the development of this work.

Author contributions

L.S.S. and F.M.N. conceptualized the project. MALDI-MS experiments were carried out by F.M.N. and L.S.S. ML analyses were performed by A.P. under the supervision of L.S.S. All authors were involved in interpretation of the data, discussed the results and commented on the manuscript. The paper was written with contributions from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0644-7.

Correspondence and requests for materials should be addressed to F.M.N. or L.S.S.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Leonardo S. Santos

Last updated by author(s): Jul 8, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☒ □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ □ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☒ □ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ □ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ □ Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Spectra format was converted using MSconvert (version 3.0.19039). Spectral Preprocessing, data analysis, Feature Selection and Machine Learning methods were performed in R (version 4.0.0) and RStudio (version Version 1.3.959). Excel for Microsoft 365 MSO (16.12730.20188) was used to infer the calibration function. Please see the methods section for further details.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw MALDI-MS spectra and the intensity matrix that support the findings of this study are available from the corresponding author upon reasonable request. Also, the authors declare that all other relevant data supporting the findings of this study are available within the paper and its supplementary information files.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed.
Data exclusions	No data were excluded from the analyses.
Replication	All attempts of replication were successful and reproducible in different laboratories (Lab 1, 2 and 3) during this proof-of-principle study.
Randomization	The MALDI-MS analyses were not randomized and all available samples were evaluated.
Blinding	Argentinian samples were blinded during the study. After characterization the samples by our technique, it was sent the positive/control samples table.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging