Article type: Full Article

Identification of COVID-19 virus (SARS-CoV-2) in human sera by Raman Spectroscopy and Multi-class Support Vector Machines

Edgar Guevara[1]*

[1] CONACYT-Universidad Autónoma de San Luis Potosí, 78210 San Luis Potosí, Mexico

*Correspondence
Edgar Guevara, CONACYT-Universidad Autónoma de San Luis Potosí, 78210 San Luis Potosí, Mexico
Email: eguevara@conacyt.mx

Abstract

With over 33 million confirmed cases and more than 1 million deaths worldwide, there is an urgent need to implement rapid screening of severe acute respiratory syndrome coronavirus (SARS-CoV-2) to decrease the infection rate. A support vector machine (SVM) with a two-layer cross-validation scheme is used to classify Raman spectra of human sera into one of three classes: COVID-19, suspected, and healthy. A total of 465 human sera samples were probed with Raman spectroscopy and subjected to SVM classification, which yielded an overall accuracy of 90.32%. In patients with confirmed COVID-19, sensitivity, specificity, and area under the receiver-operating curve (AUC) were .95, .99, and .99, respectively. In the suspected group were .89, .95, and .97, and in healthy controls .93, .90, and .97. These results suggest that

our model can be used as a fast screening tool to accelerate the detection of COVID-19 positive patients to fight the ongoing pandemic.

Keywords: Raman spectroscopy; COVID-19; SARS-CoV-2; machine learning; support vector machine

Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus; **COVID-19**, coronavirus disease; **ROC**, receiver-operating curve; AUC, area under the receiver-operating curve; **SVM,** support vector machine; **Acc,** accuracy; **Se,** sensitivity; **Sp,** specificity.

1 INTRODUCTION

Coronavirus disease (COVID-19) is the latest global outbreak to threaten human health, economy, and society all around the world [1]. This infectious disease is caused by severe acute respiratory syndrome coronavirus (SARS-CoV-2) strain [2]. SARS-CoV-2 is a relatively large virus with an RNA genome encapsulated in a lipid membrane and spiked with glycoproteins [3]. COVID-19 can be transmitted through respiratory droplets, contact with nose, eye, and mouth mucosa [4] or contaminated surfaces [5]. Furthermore, there is surmounting evidence on the possibility of airborne transmission for the spreading of COVID-19 [6].

With over 33 million confirmed cases and more than 1 million deaths worldwide, as of September 28^{th,} 2020 [7,8], there is a crucial and urgent need to implement rapid detection of SARS-CoV-2 to decrease the rate of infection [9]. Some of the proposed optical techniques include infrared spectroscopy and Raman spectroscopy [10]. Both techniques are sensitive to low molecular concentrations and complement each other due to their sensitivity to changes in either the molecule polarizability or its dipole moment [11].

Raman spectroscopy is a light-based technique that can identify molecular signatures from a large variety of samples with minimal preparation and has shown promise as a universal diagnostics tool [12]. This spectroscopic technique has been proven to be an invaluable tool in the study of a large assortment of viral diseases, including influenza type A [13,14], avian influenza subtype H5N1 [15], hepatitis C [16], human immunodeficiency viruses [17] and the Middle East respiratory syndrome (MERS) [18]. Further studies have explored Raman spectroscopy to detect SARS-CoV-2 in human sera[19], saliva [20], and environmental specimens

[21], using univariate statistical inferences [20] and an assortment of multivariate analysis, such as principal component analysis [20,21], support vector machines (SVM) [19] and linear discriminant analysis [20]. There is also a registered study that plans to use Raman spectroscopy as a diagnostic tool for COVID-19 [22].

Whereas these recent studies indicate the feasibility of detecting SARS-CoV-2 in bodily fluids, there are common mistakes in their classification models, mainly the placement of the parameters optimization step in the cross-validation loop [23]. This issue overestimates the performance of the classifier and leads to models that fail to generalize in a new, completely independent, dataset.

I, therefore, investigate a two-layer cross-validation model where the test data is completely independent of both the training and the optimization of the classification model. This will result in a robust and reliable method to rapidly screen SARS-CoV-2 with high sensitivity and specificity.

2 EXPERIMENTAL SECTION

2.1 Sample collection

A total of 465 Raman spectra from human sera were collected from the first publicly available database with COVID-19 Raman spectra [19]. The dataset is comprised of 159 samples labeled as confirmed cases of COVID-19, 156 suspected samples, i.e. with fever but not COVID-19, and 150 healthy controls. A spectrometer consisting of a cooled charge-coupled device, a laser, and Volume Phase Holographic (VPH) spectrograph was used, as described by Yin *et al.* [19]

2.2 Pre-processing of Raman spectra

Before the construction of the classification model, Raman spectra were pre-processed to remove artifacts and interference from auto-fluorescence. All analyses were carried out in MATLAB (The MathWorks, Natick, MA). First, the spectra were cropped to the fingerprint region between the wavenumber range from 400 to 1800 cm^{-1} [24]. Thereafter, an automated iterative polynomial fitting method was used to remove the influence of the intrinsic fluorescence of the sample in the Raman spectra [25]. The function `vancouver.m` was utilized [26] and the polynomial order was set to polyOrder = 5, the error threshold was limited errThreshold = 0.02 and the maximum number of iterations was fixed to nIter = 500. Finally, a moving

average filter was applied to smooth the data, while preserving its spectral features, with a window width of nPoints = 2. As a last step, vector normalization was used to correct for variability in viral concentrations and sample thickness [27,28].

2.3 Multi-class Support Vector Machines

A support vector machine is a binary classifier, of linear nature [29]. Considering the multivariate dataset x, then a linear decision boundary $g(x)$ is found between the samples that are closest to the border of the neighboring class as shown in Equation (1). These samples are the support vectors $\boldsymbol{sv}$.

$$g(\boldsymbol{x}) = sgn(\boldsymbol{w}\boldsymbol{x}' + b) \tag{1}$$

Where $\boldsymbol{w}$ and b are the weight and bias computed from the training set. The sign of $g(\boldsymbol{x})$ dictates where a sample belongs to which one of the two classes.

Very often classes are not linearly separable and hence a kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$ must be used to transform the spectral components of our Raman data $\boldsymbol{x}$ into a feature space $f(\boldsymbol{x})$ where the linear classification rule can be successfully applied [28].

$$f(\boldsymbol{x}) = sgn(\alpha k(\boldsymbol{x}, \boldsymbol{x}') + b) \tag{2}$$

In Equation (2) the parameter α is a Lagrange multiplier, used to optimize $f(\boldsymbol{x})$ subject to constraints[30]. The hyper-parameter C is set as the upper bound of α:

$$0 \leq \alpha_i \leq C \tag{3}$$

In this work, the radial basis function (RBF) kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ was used due to the data distribution.

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma \|\boldsymbol{x} - \boldsymbol{x}'\|^2) \tag{4}$$

Where the hyper-parameter σ controls the width of the Gaussian, playing a similar role as the degree of a polynomial fitting function [31].

The optimization function $\phi(\boldsymbol{\alpha})$ can now be written in terms of the kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$ as:

$$\phi(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{i \in sv} \sum_{j \in sv} \alpha_i c_i k(x_i, x_j) c_j \alpha_j - \sum_{i \in sv}^{j} \alpha_i \tag{5}$$

Where c designates the class membership $c = +1$ for positive class and $c = -1$ for negative class. And finally, the decision boundary $g(x)$: $x \mapsto y$ can be expressed as:

$$g(\boldsymbol{x}) = sgn\left(\sum_{i \in sv} \alpha_i c_i k(\boldsymbol{s_i}, \boldsymbol{x}) + b\right) \tag{6}$$

Which assigns any input spectra x to a class y and displays a dependency on the support vectors $\boldsymbol{sv}$.

A one-vs-all multi-class implementation was used to transform this binary classifier into a multiple class discrimination model [30,32]. Briefly, a binary SVM can be extended to discriminate between multiple classes by training an SVM model with all the samples in a certain class y labeled as $c = +1$, while all the other samples are labeled as negative $c = -1$. The decision boundary $g(\boldsymbol{x})$ can then be determined by finding the class that maximizes the decision function $\sum_{i \in \boldsymbol{sv_y}} \alpha_i c_i k(\boldsymbol{s_{iy}}, \boldsymbol{x}) + b_y$ for class y:

$$g(\boldsymbol{x}) = \max\left(\sum_{i \in \boldsymbol{sv_y}} \alpha_i c_i k(\boldsymbol{s_{iy}}, \boldsymbol{x}) + b_y\right) \tag{7}$$

The one-vs-all extension method was implemented using the function `fitcecoc.m`.

Loss function L is defined as the fraction of misclassified observations:

$$L = \sum_{i \in sv} \boldsymbol{w_i}\{g(x_i) \neq y_i\} \tag{8}$$

Furthermore, the optimal hyper-parameters $C$and σ were found through a Bayesian optimization process, that aims to maximize the expected improvement EI [33]. EI is defined as the amount of improvement that we can expect over some objective when we evaluate the target function at x:

$$L = \sum_{i \in sv} \boldsymbol{w_i}\{g(x_i) \neq y_i\} \tag{9}$$

2.4 Cross-validation

A dual-layer cross-validation scheme was implemented to avoid both overfitting and overestimation of the classifier performance [23]. The pseudo-code of this layered cross-validation system is displayed in **Figure 1**. In the external cross-validation, loop data were split into $N = 10$ folds and each fold was used only once as a completely independent test set. The

remaining $N-1$ folds were used as the external training set and split into $M=5$ folds in the internal cross-validation loop. Within this internal loop, data were further divided into the internal training data set ($M-1$ folds) and the validation data set (1 fold). The hyperparameter optimization method was built only with the internal training set and validated by the validation dataset, resulting in $M=5$ validation accuracies. Then, the optimal model with C and σ hyperparameters showing the highest validation accuracy was chosen to build the SVM model with the external training set. Finally, this optimal model was used to predict the external training set. The process is repeated until all the samples are independently tested and the testing accuracy is computed over the average of $N=10$ folds.

```
for iFolds = 1:nFolds
  // External Cross-Validation Loop
  split(fullDataSet, nFolds)
  // Split randomly in nFolds=10
  trainSet = Training dataset (N-1 folds)
  testSet = Test dataset (1 fold)
  tSVM = templateSVM
  //Create SVM template
  // Internal cross-validation to optimize
  // SVM hyper-parameters (C, sigma)
  for jFolds = 1:mFolds
    // Training set from the ith iteration
    // of external CV
    split(trainSet(iFolds), mFolds)
    // Split randomly in mFolds=5
    internalTrainSet = Internal training
    dataset (M-1 folds)
    validationSet = Validation dataset (1
    fold)
    // Bayesian optimization
    for kIter = 1:maxIter
      classifier = train(tSVM,
      internalTrainSet(jFolds),
      validationSet(jFolds))
      // Update the loss function using
      // previous observed values
      f = loss(classifier, C, sigma)
      // Find the new set of parameters
      // that maximize expected
      // improvement
      (C, sigma)_new = argmax{EI(C,
       sigma)}
      // Compute the loss function
      f_new = loss(classifier, (C,
      sigma)_new)
    end for
    optSVM = SVM model with the optimal
    hyper-parameters (C, sigma)
    accOptim(jFolds) =
    accuracy(predict(classifier(optSVM,
    validationSet(iFolds))))
  end for
optSVM(C, sigma) = SVM model with the
maximum accuracy in accOptim
// Predict a completeley independent test
testAcc(iFolds) = accuracy(predict
(classifier(optSVM, testSet(iFolds))))
end for
Acc = mean(testAcc)
// Average accuracy over all nFolds
```

Figure 1. Pseudocode of the nested cross-validation.

2.5 Performance metrics

The performance of the proposed SVM model was assessed by using overall accuracy (Acc), sensitivity (Se), specificity (Sp), and area under the receiver-operating characteristic (ROC) curve (AUC) [34]. A true positive (TP) was defined as a sample from the positive class correctly classified, a true negative (TN) a sample from the negative class correctly labeled, a false negative (FN) a sample from the positive class labeled as negative, and a false positive (FP) is a negative sample classified as positive. Accuracy, sensitivity, specificity, ROC curve and area under the ROC curve are calculated using equations (10) - (14):

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$Se = \frac{TP}{TP + FN} \tag{11}$$

$$Sp = \frac{TN}{TN + FP} \tag{12}$$

$$ROC := Se\ vs.\ (1 - Sp) \tag{13}$$

$$AUC = \int_0^1 Se\ d(1 - Sp) \tag{14}$$

Each group was considered the positive class at a time for the ROC curve computation. Also, 95% confidence intervals were computed with bootstrap resampling technique, and averages of ROC curves were taken over 500 repetitions [35].

All metrics reported in this study were computed over the nested cross-validation process using the completely independent test dataset.

3 RESULTS AND DISCUSSION

Mean vector-normalized spectra of each group (healthy, suspected, and COVID-19) are depicted in **Figure 2** (a). Shaded error bars indicate one standard deviation about the mean and are meant to be used as aids to appreciate the variability of the data. All three groups present characteristic Raman bands at the same locations, hence univariate analysis based on peak or integrated intensity did not show any significant differences. The sixteen identified Raman bands and their molecular assignments are described in **Table 1**. Some bands are attributed to

genetic material, Large contributions from lipids [36] and proteins [37,38] - in particular Amide I and III - are also present in all groups, as they are non-specific and expected to appear across all samples.

Table 1. Main Raman bands observed in human serum samples and their assignment (bands labeled according to **Figure 1**). P = proteins, L = lipids, *ν* = stretch, *δ* = bending, *s* = symmetric.

Raman shift (cm^{-1})	Assignment
456	P, L-Proline [39]
527	ν P, Disulphide, [36,45]
808	Ring breathing, Cytosine, Uracil [39,40]
840	L -tryptophan [39]
897	ν_s, (CNC) amino group [39]
939	Ring breathing, Guanine, [39,44]
969	ν, PO_4 [42]
1001	Ring breathing, L-phenylalanine [44]
1033	Ribose sugars [20]
1150	δ, C-N P and DNA [37]
1167	Carotenoids [37]
1220	ν, P Amide III [37,38]
1328	ν, P diazo bond N=N [46]
1440	ν, L C=C [36]
1459	δ, L C-H [36]
1656, ~1660	ν, P, Amide I, [37,38]

Difference spectra are plotted in panel (b) of **Figure 2** and they show that spectra from the suspected and healthy groups have the most similar molecular fingerprint, therefore we expect our classifier to mislabel some of the samples from these groups. However, COVID-19 spectra show clear differences from the other two groups. COVID-19 shows the largest difference in the 808 cm^{-1} band, which is due to ring breathing of nitrogenous bases present in the virus ribonucleic acid (RNA), such as cytosine and uracil [39,40]. This suggests that the largest molecular differentiation issues from RNA specific to SARS-CoV-2. There is also a clear separation at the 840 cm^{-1} band, distinctive of lipids, that could be due to COVID-19 patients suffering from hyperlipidemia [41] or lipids from the viral membrane [3]. At a Raman shift of 969 cm^{-1}, vibrational mode from PO_4 [42] shows changes between COVID-19 and the other two groups. Positive peaks at 1150 cm^{-1} and 1328 cm^{-1} indicate increased DNA and proteins in the COVID-19 group [37]. The increased DNA synthesis may be the result of increased

lymphocytes as an immune response to the viral infection [43]. Negative peaks in the difference spectra at 1001 cm^{-1} and 1459 cm^{-1} indicate decreased L-phenylalanine and C-H bending modes from lipids in the COVID-19 group [36,44].

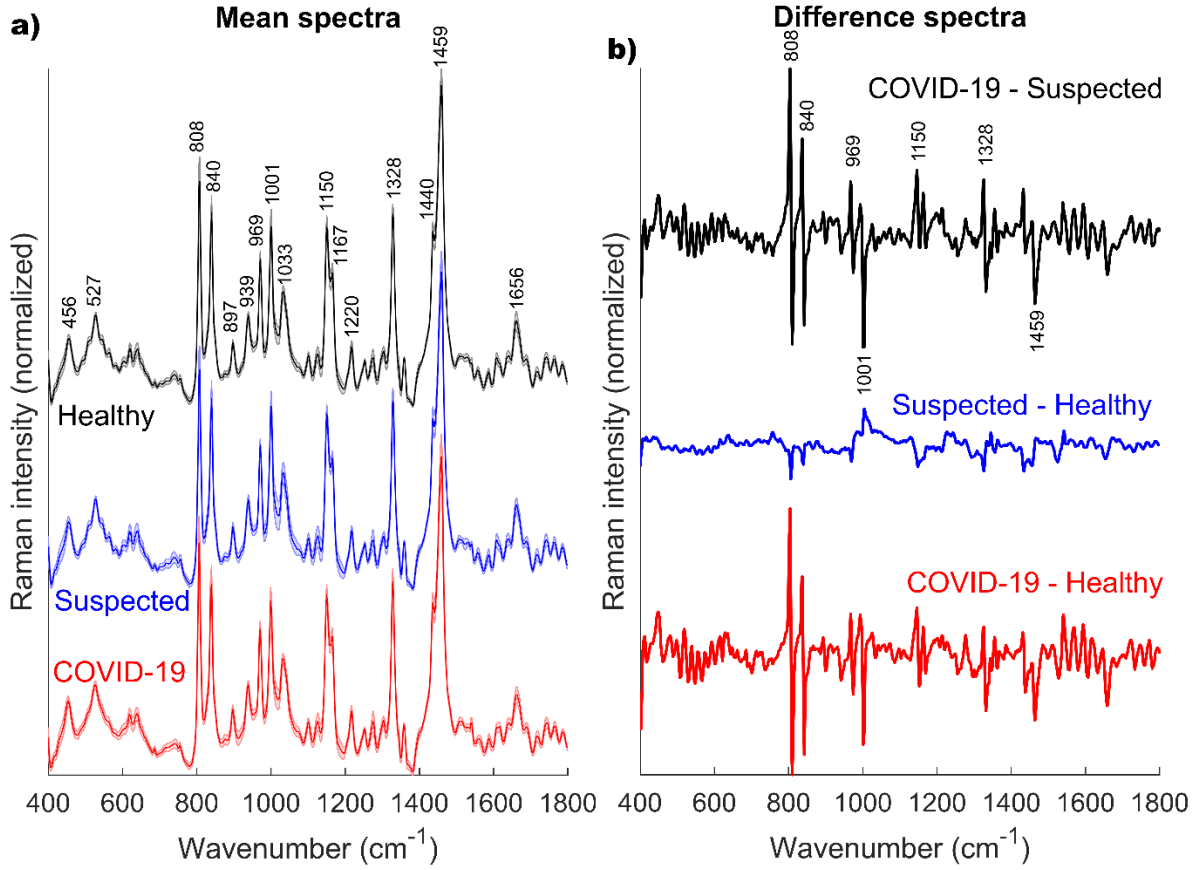

Figure 2. (a) Raman spectra (mean ± s.d.) for the three groups of human sera (COVID-19, suspected and healthy), with characteristic bands shown in the 400 -1800 cm^{-1} range. (b) Difference spectra are computed from the mean normalized spectrum from each group.

Figure 3 displays the classification results from our proposed SVM model that yielded an overall accuracy of 90.32%. From the 159 COVID-19 samples, only 8 of them were incorrectly classified, 3 were labeled as suspected and 5 as healthy. Among the 158 suspected cases, a total of 23 samples were mislabeled, 3 of them were classified as COVID-19 and 20 as healthy. From the 150 healthy controls, 14 of them were not categorized correctly, 3 were assigned the COVID-19 label, and 11 the suspected category. As the difference spectra suggested, there is a higher rate of misclassification between the healthy and the suspected group, because of the high similarity of their respective molecular fingerprints. 31 samples of the 45 misclassifications were suspected classified as healthy and vice versa. These results suggest that our model has high discriminatory power for COVID-19.

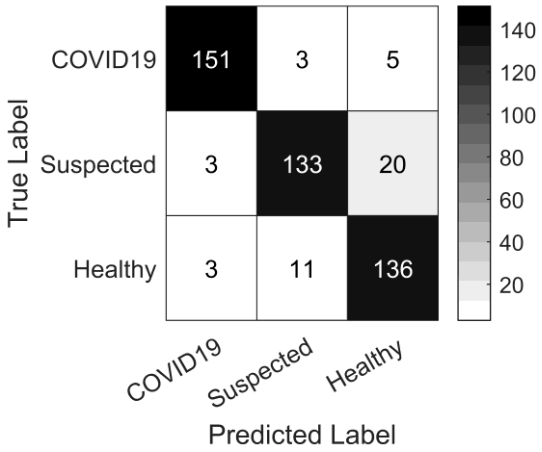

Figure 3. Confusion matrix of an SVM classifier applied to the spectral dataset. Every column of the matrix represents the occurrences in a predicted label, while every row indicates the occurrences in the actual label.

Figure 4 displays the ROC curve analysis for each group. Classification of COVID-19 spectra as the positive class yielded the highest AUC = 0.99 (0.97 – 1.00), while both suspected and healthy yielded the same AUC = 0.97, albeit the suspected group presented a slightly larger 95% interval of confidence (0.94 – 0.98), as compared to the healthy group (0.95 – 0.98).

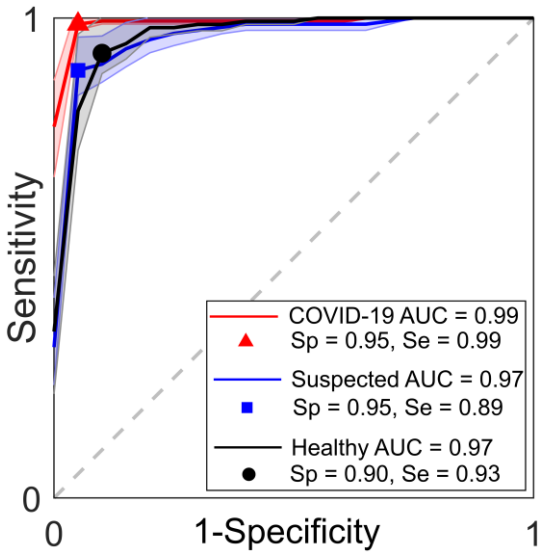

Figure 4. The ROC analysis for the multi-class SVM model. 95% confidence intervals are shown in shaded error bands. The red triangle, blue square, and black circle denote the optimal operating points of the COVID-19, suspected, and healthy classes, respectively.

Table 2 contains the sensitivity and specificity values for each class. Our model is most specific to COVID-19 and suspected Raman spectra (Sp=0.95) and least sensitive to the Raman spectra of healthy subjects (Sp=0.90). The single most advantageous feature of our classifier is its very high sensitivity to COVID-19 samples (Se=0.99) which is higher than that of other Raman spectroscopy-based studies [20]. Although the overall accuracy of our method (90.32%) is slightly lower than the 91.66% accuracy reported by Desai *et al.* [20], our two-layer cross-validation and its testing with samples completely independent of those used in optimization ensure a robust and repeatable method capable of generalizing to a real-world clinical setting. As a low-cost and rapid screening method, the seroconversion of the immunoglobulin G (IgG) and immunoglobulin M (IgM) antibodies against SARS-CoV-2 have shown a sensitivity of 87.5% and 77.3%, respectively, with 95% and 100% specificity in patients with confirmed COVID-19 infection [47,48]. Whereas in patients with suspected infection IgG and IgM have shown 70.8% and 87.5% sensitivity, respectively, while preserving 96.6% and 100% specificity [48]. Reverse transcriptase polymerase chain reaction (RT-PCR) has long been considered the gold standard to confirm COVID-19 infection [49], although it is debatable if it should be considered the gold standard [47]. Therefore, a direct comparison of our method with RT-PCR is not feasible.

Table 2. Results of the receiver operating characteristics (ROC) curve: AUC values with 95% confidence intervals and optimal operating points.

Positive class	AUC	Sp	Se
COVID-19	0.99 (0.97 – 1.00)	0.95	0.99
Suspected	0.97 (0.94 – 0.98)	0.95	0.89
Healthy	0.97 (0.95 – 0.98)	0.90	0.93

4 CONCLUSION

These results suggest that our model performs adequately with an unseen test dataset and can be used as a fast screening tool in high throughput applications. This will hasten the detection of COVID-19 positive patients, provide treatment, and enforce their subsequent isolation, which might help accelerate the end of the ongoing pandemic.

ACKNOWLEDGMENTS

The author would like to acknowledge support from “Consejo Nacional de Ciencia y Tecnología” (CONACyT) through grants “Cátedras CONACyT” project 528 and “Ciencia de Frontera” No. 20884/2020.

CONFLICT OF INTEREST

The author declares no financial or commercial conflict of interest.

DATA AVAILABILITY STATEMENT

All code generated and datasets analyzed in this paper are available at https://github.com/guevaracodina/RamanCOVID19.git

REFERENCES

[1] A. S. Fauci, H. C. Lane, R. R. Redfield, *New England Journal of Medicine* **2020**, *382*, 1268–1269.

[2] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, *Nature Medicine* **2020**, *26*, 450–452.

[3] C. Liu, Q. Zhou, Y. Li, L. V. Garner, S. P. Watkins, L. J. Carter, J. Smoot, A. C. Gregg, A. D. Daniels, S. Jervey, D. Albaiu, *ACS Cent. Sci.* **2020**, *6*, 315–331.

[4] X. Peng, X. Xu, Y. Li, L. Cheng, X. Zhou, B. Ren, *Int J Oral Sci* **2020**, *12*, 9.

[5] L. Fiorillo, G. Cervino, M. Matarese, C. D’Amico, G. Surace, V. Paduano, M. T. Fiorillo, A. Moschella, A. La Bruna, G. L. Romano, R. Laudicella, S. Baldari, M. Cicciù, *International Journal of Environmental Research and Public Health* **2020**, *17*, 3132.

[6] R. Zhang, Y. Li, A. L. Zhang, Y. Wang, M. J. Molina, *PNAS* **2020**, *117*, 14857–14863.

[7] E. Dong, H. Du, L. Gardner, *The Lancet Infectious Diseases* **2020**, *20*, 533–534.

[8] Center for Systems Science and Engineering (CSSE), “COVID-19 Map,” can be found under https://coronavirus.jhu.edu/map.html, **n.d.**

[9] G. Seo, G. Lee, M. J. Kim, S.-H. Baek, M. Choi, K. B. Ku, C.-S. Lee, S. Jun, D. Park, H. G. Kim, S.-J. Kim, J.-O. Lee, B. T. Kim, E. C. Park, S. I. Kim, *ACS Nano* **2020**, *14*, 5135–5142.

[10] L. F. das C. e S. de Carvalho, M. S. Nogueira, *Photodiagnosis Photodyn Ther* **2020**, *30*, 101765.

[11] K. Hashimoto, V. R. Badarla, A. Kawai, T. Ideguchi, *Nature Communications* **2019**, *10*, 4411.

[12] N. M. Ralbovsky, I. K. Lednev, *Spectrochim Acta A Mol Biomol Spectrosc* **2019**, *219*, 463–487.

[13] G. Pezzotti, W. Zhu, T. Adachi, S. Horiguchi, E. Marin, F. Boschetto, E. Ogitani, O. Mazda, *Journal of Cellular Physiology* **2020**, *235*, 5146–5170.

[14] K. Dardir, H. Wang, B. E. Martin, M. Atzampou, C. B. Brooke, L. Fabris, *J. Phys. Chem. C* **2020**, *124*, 3211–3217.

[15] W. S. Khan, M. Z. Iqbal, in *Nanobiosensors*, John Wiley & Sons, Ltd, **2020**, pp. 289–310.

[16] H. Cheng, C. Xu, D. Zhang, Z. Zhang, J. Liu, X. Lv, *Photodiagnosis and Photodynamic Therapy* **2020**, *30*, 101735.

[17] S. M. Kim, T. Lee, Y.-G. Gil, G. H. Kim, C. Park, H. Jang, J. Min, *Materials* **2020**, *13*, 3234.

[18] H. Kim, J. Hwang, J. H. Kim, S. Lee, M. Kang, in 2019 IEEE 14th International Conference on Nano/Micro Engineered and Molecular Systems (NEMS), **2019**, pp. 498–501.

[19] G. Yin, L. Li, S. Lu, Y. Yin, Y. Su, Y. Zeng, M. Luo, M. Ma, H. Zhou, D. Yao, G. Liu, J. Lang, *figshare* **2020**, *Dataset*, DOI 10.6084/m9.figshare.12159924.v1.

[20] S. Desai, S. V. Mishra, A. Joshi, D. Sarkar, A. Hole, R. Mishra, S. Dutt, M. K. Chilakapati, S. Gupta, A. Dutt, *Journal of Biophotonics* **2020**, *Online ahead of print*, DOI 10.1002/jbio.202000189.

[21] D. Zhang, X. Zhang, R. Ma, S. Deng, X. Wang, X. Zhang, X. Huang, Y. Liu, G. Li, J. Qu, Y. Zhu, J. Li, *medRxiv* **2020**, 2020.05.02.20086876.

[22] L. Jacobi, V. H. Damle, B. Rajeswaran, Y. R. Tischler, *RSOS Registered Reports* **2020**, DOI 10.17605/OSF.IO/Y54H3.

[23] S. Guo, T. Bocklitz, U. Neugebauer, J. Popp, *Anal. Methods* **2017**, *9*, 4410–4417.

[24] W. Lee, A. T. M. Lenferink, C. Otto, H. L. Offerhaus, *Journal of Raman Spectroscopy* **2020**, *51*, 293–300.

[25] J. Zhao, H. Lui, D. I. McLean, H. Zeng, *Appl. Spectrosc.* **2007**, *61*, 1225–1232.

[26] E. Guevara, F. J. González, Algoritmo de Reducción de Fluorescencia En Señales Raman, **2019**.

[27] H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, M. R. McAinsh, N. Stone, F. L. Martin, *Nat Protoc* **2016**, *11*, 664–687.

[28] C. L. M. Morais, K. M. G. Lima, M. Singh, F. L. Martin, *Nature Protocols* **2020**, *15*, 2143–2162.

[29] C. Cortes, V. Vapnik, *Mach Learn* **1995**, *20*, 273–297.

[30] R. G. Brereton, G. R. Lloyd, *Analyst* **2010**, *135*, 230–267.

[31] A. Ben-Hur, J. Weston, in *Data Mining Techniques for the Life Sciences* (Eds.: O. Carugo, F. Eisenhaber), Humana Press, Totowa, NJ, **2010**, pp. 223–239.

[32] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Muller, E. Sackinger, P. Simard, V. Vapnik, in *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)*, **1994**, pp. 77–82 vol.2.

[33] P. I. Frazier, arXiv:1807.02811 [cs, math, stat] **2018**.

[34] N. Hameed, A. M. Shabut, M. K. Ghosh, M. A. Hossain, *Expert Systems with Applications* **2020**, *141*, 112961.

[35] B. Efron, R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman And Hall/CRC, New York, **1993**.

Graphical Abstract

A support vector machine model based on Raman spectra of human sera was built to screen for coronavirus disease COVID-19 with 90.32% overall accuracy.