Conference on
Information Technology and
Data Science

Classifying Raman Spectroscopy Data Using Machine Learning Algorithms for Diagnosing Infection With Sars-Cov-2

Robert Istvan Oniga

Katholieke Universiteit Leuven, Belgium
oniga.robi@gmail.com

Proceedings of the 1st Conference on Information Technology and Data Science
Debrecen, Hungary, November 6–8, 2020
published at http://ceur-ws.org

Abstract

The rapid development of the corona crisis requires new methods and approaches that could help flatten the curve. For this reason, a possible alternative for a detection method is investigated to diagnose in a faster and more reliable way the disease and help prevent the spread. Raman spectroscopy on blood serum is a potential candidate for this issue and thus, research was done towards this direction. The data obtained from the spectrometer was further analyzed through linear discriminant analysis and a predictive model was achieved with an accuracy of 93.5%.

Keywords: SARS-COV-2, LDA, Raman spectroscopy, data processing

1. Introduction

In December 2019, a novel virus has emerged and in a brief period it has reached all the corners of the world. This virus called the corona virus has affected all people. Due to the continuous rise in the number of infected and deceased persons, a new approach must be taken in order to deal with this situation. Since the measures taken to isolate infected individuals had no significant result with the current testing results, perhaps the development of a new detection method that is more precise and rapid could prevent the further spread of the virus.[5, 9] A possible detection method that could satisfy the requirements is by means of Raman spectroscopy on blood serum and with the help of machine learning techniques.

2. Used Method

2.1. Raman Spectroscopy

Raman spectroscopy is a method that relies on a non ionizing laser that can excite a molecule if the energy of the incident photon matches the energy gap between the ground state and the excited state of said molecule. The phenomena of fluorescence will occur when the molecule relaxes and generates emission of photons in both the visible and near-infrared spectral ranges. The emission can happen either by means of a elastic scattering which has no relevant information, or plastic scattering which means that a part of the energy was absorbed and only a fraction of that was released back into the medium. That difference in terms of energy is studied in order to obtain relevant information that may be used for bio-medical applications but not only.[10] In this paper the differences in terms of energy absorption between healthy and infected individuals is studied for a possible diagnosis. The said differences are subtle, however, precise statistical algorithms can detected these trends leading to a clear diagnosis.

In many bio-medical applications that include the use of Raman spectroscopy, the probe being analyzed is blood serum. Many important features can be observed that prove to be relevant in diagnosis. More precisely, in the blood serum, different organic components are present such as proteins that can indicate the presence of a virus, or one can analyze the serum in order to diagnose different types of carcinoma. However, in order to be able to analyse blood serum, a series of preparatory steps must be taken to obtain a clear and reliable sample.[11] These steps include:

- Blood has to be drawn from the subject in question.
- The blood has to sit in the test tube for 15-30 minutes in order to clot.
- The blood has to be centrifuged in order to get rid of the clot.
- The serum has to be immediately transferred to another tube to preserve the purity.

Knowingly, the use of Raman spectroscopy on blood serum is proposed as a detection method of infection with the SARS-COV-2 virus. Further knowledge regarding the proteins that mark the presence of the virus in the human system was acquired in order to obtain an estimated location in the spectroscopic data of those elements for better detection by means of machine learning. The most important protein that characterizes the virus is the spike glicoprotein also called S-protein. Because of this particular protein the virus is able to enter our system by attaching the so called spikes to a receptor called angiotensin converting enzyme 2 (ACE2).[7]

Proteins in general are made of amino acid chains which all contain amine and carboxyl functional groups. Because in this paper the main focus is about processing raman spectroscopic data, the vibrational modes of the proteins have to be understood. The most characteristic bands in raman spectroscopy for proteins are associated with the CONH group that stretch up to 3100 cm^{-1} (amide A) and several other band for amide B: 1600–1690 cm^{-1}, 1480–1580 cm^{-1}, 1230–1300 cm^{-1}, 625–770 cm^{-1}, 640–800 cm^{-1}, 540–600 cm^{-1} and 200 cm^{-1}.[8] This information will be later compared with the results of the machine learning algorithm in order to identify with greater preciseness the component in the spectroscopic data that will yield the best differentiation rate.

2.2. Liner Discriminant Analysis (LDA)

For this application, a linear classification technique called LDA was used. This method is straightforward and it consists of analysis of statistical properties of data that are calculated for each class. The statistical properties of interest are the mean and covariance matrix over the multiple variables. The assumption made by the algorithm is that all data is Gaussian and each attribute of the data has equal variance and each value varies around the mean with the same amount overall. After these are taken into consideration, the algorithm estimates the mean and the variance for each class. Having these parameters, the next step can be made: classification.[2] In order to estimate the probability of a unknown sample to be part of a certain class, the model uses Bayes theorem which is governed by the following formula:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)},$$

where A and B are events, $P(A \mid B)$ is the probability of A given B is true, $P(B \mid A)$ is probability of B given A is true and finally, $P(A)$ and $P(B)$ are independent probabilities of A and B. [4]

2.3. Leave-One-Out Cross-Validation

The classification technique discussed above can yield accurate results on its own without the need of reassuring techniques such as leave-one-out cross-validation. However, being a delicate biomedical application, the maximum achievable accuracy is of interest. For this particular reason the sequential training and testing of different unique values is made. Precisely, each data point from the data-set will serve as test sample once. By doing this, all data points are thoroughly analyzed and assimilated with one of the classes with greater preciseness.

3. Particular Application of Methods

The data processing was done in MATLAB software by means of machine learning and statistical toolboxes. The data-set consisted of Raman spectra on blood serum of both healthy and infected subjects. The first step was to whiten the data by removing the DC component and then the data was visually inspected in order to observe any outstanding features that might impede the classification process. The whitened data can be observed in Figure 1.

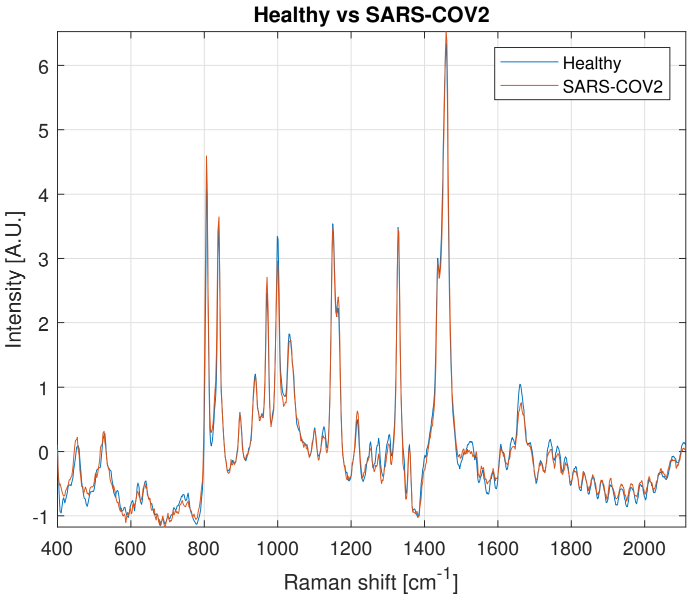


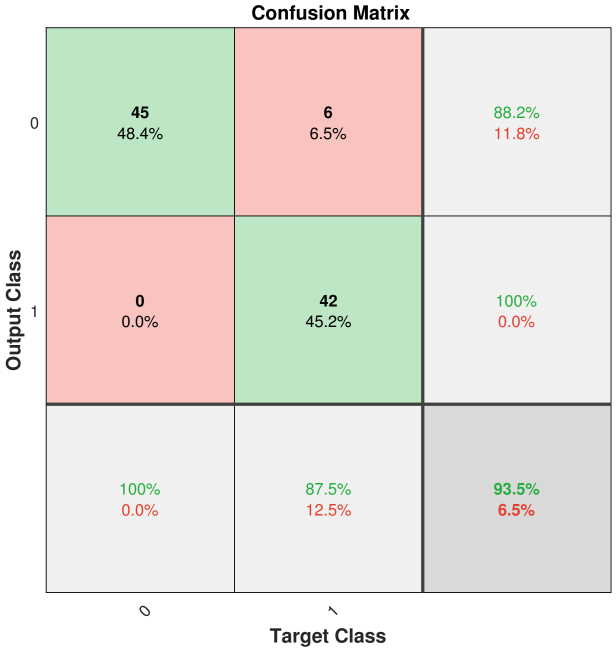
Figure 1. Normalized spectroscopic data.

The displayed data has subtle differences which cannot be picked up by the naked eye in order to make a precise classification. For this reason, after the algorithm successfully detects those features that are relevant for the differentiation between the two classes, the raman shift bands will be displayed for a better visual analysis.

After the normalization process, the data was randomized along with the labels and in the healthy and infected individuals were mixed together in order to prepare for the training process of said classifier. The division in terms of samples was done in the following manner: 70% of the data was used as training set and 30% of the data was used as test set. However, as it has been mentioned before, the training was done through LDA and through LDA with leave-one-out cross-validation technique in order to compare the results and to take into consideration the possibility of over-fitting of the model.[1]

4. Results

The data-set was composed of a total of 309 individuals of which 150 were healthy and 159 were infected with the corona virus[12]. Initially, the training process was done by using only the linear discriminant analysis method which yielded a 89% accuracy which means that 89% of the subjects in question were correctly identified as being either healthy of infected while 11% were wrongfully attributed to a class. In order to test for a possible improvement of the model, leave-one-out cross-validation was used. In this algorithm rather than using sample by sample to obtain the prediction, Raman shift ranges of length of approximately 50 cmˆ1 were used. The results of this alternative were better having a accuracy of 93.5%. Since the algorithm performed many operations with a small amount of data, the possibility of over-fitting was taken into consideration before applying the method. However, this was not the case with this data-set. To get a better understanding of the results, the confusion matrix was displayed and it is shown in Figure 2.



	0	1	
0	45 48.4%	6 6.5%	88.2% 11.8%
1	0 0.0%	42 45.2%	100% 0.0%
	100% 0.0%	87.5% 12.5%	93.5% 6.5%

Figure 2. Confusion matrix.

From this table it can be seen that from the total of 48 healthy individuals used for the testing process, 47 were correctly identified as being healthy (which accounts for 47 true negatives) and one was erroneously identified as infected (false positive). The infected individuals had a lower accuracy when it comes to detection as from the total of 45 infected, only 40 were correctly identified as indeed infected (true positive) and 5 were assumed to be healthy (false negative). Other parameters were taken into consideration for performance assessment of the classifier such as the accuracy, sensitivity and specificity of the model. All these parameters can be seen in Table 1.

Table 1. Performance of the classifier.

Accuracy	93.55%
Sensitivity	83.33%
Specificity	90.38%

As discussed in 2.1, certain proteins yield Raman shift ranges that are useful for the classification process. The predominant features in terms of Raman shift in this particular data-set were in the following ranges: [400–591]cm^{-1}, [647–673]cm^{-1}, [721–798]cm^{-1}, [820–896]cm^{-1} and [1003–1241]cm^{-1}. From all the ranges discussed in 2.1 the one that overlaps significantly with the predominant features in this dataset is the one in the 600–800 cm^{-1} range. The conclusion that can be drawn from this is that the spike glicoprotein that is responsible for the infection with the SARS-COV-2 virus has a high vibrational state in this particular range which can be used to detect infected individuals. Now, additional information has come to light and better approaches can be made in order to increase the accuracy of the model. The predominant features used to achieve the classification process are now reduced to one feature generally speaking. This single feature is represented by a range of Raman shifts in the 600–800 cm^{-1} that can be seen in Figure 4.

From a visual inspection it can be seen that the data from the two classes show little differences in terms of intensities. However, there are certain location in the specified ranges of raman shifts in which the differences in intensities go up to 0.5 [A.U.]. These values are high from the perspective of Raman spectroscopy accounting for a difference of 12.5% in terms of intensity. When analyzing these differences the most important characteristic that was considered at all times was the quantity of virus present in the sample which could have been very low for some of the subjects. The quantity is highly dependent on one parameter: days after infection (Figure 3). Because of this reason, and because of the fact that there was no information regarding this parameter, possible errors in classification might have occurred.

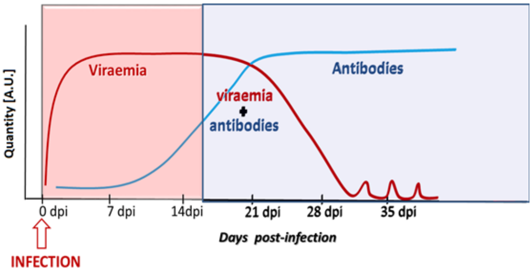


Figure 3. Level of virus and antibodies in the system.

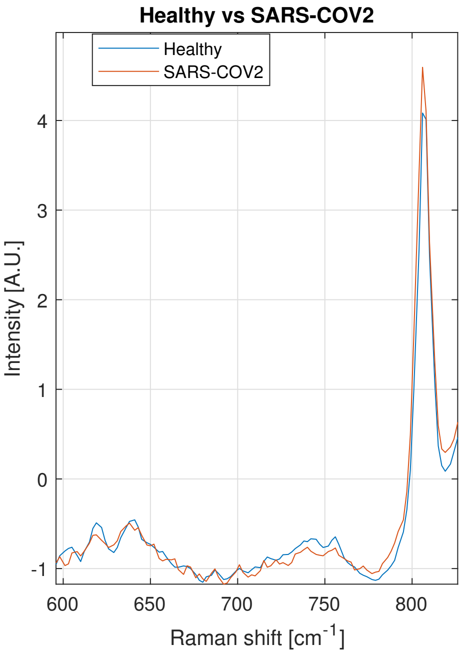


Figure 4. Range of interest from normalized data.

5. Conclusions

The proposed technique for detection of SARS-COV-2 infection proved itself to be useful and reliable. The potential of Raman spectroscopy in the detection of the virus is considerable especially due to its ability of observing even subtle differences that can be missed by the naked eye and which are key elements for the classification process. The LDA algorithm combined with leave-one-out cross-validation created a statistical model that is well suited for the task at hand. Rapid tests are time-efficient, however the accuracy of detection is only ∼80%. These tests can be done by the patient itself and they require no medical expertise.[6] The most-widely-used method is the PCR one and this requires a laboratory and trained personnel in order to obtain the test result. Even though it has a high accuracy, the time needed to obtain a result is long as it can take up to 48 hours.[3] Both methods have their own shortcomings; when compared to the Raman spectroscopy based detection method, it can be seen that this one has only the beneficial parts of both: does not require a laboratory as there are plenty of portable Raman spectrometers, this method is highly time efficient as the result can be passed to the patient in less than one hour, there is no need for extra training in order to perform the testing operation as the process can be done by a nurse, it has high accuracy (∼93.5%). All these advantages add up to a novel method that can impede the rapid spread of the virus thus contributing to a faster ending of the pandemic.

References

[1] S. rekha Hanumanthu: *Role of Intelligent Computing in COVID-19 Prognosis: A State-of-the-Art Review.* Chaos, Solitons and Fractals (2020),
doi: https://doi.org/10.1016/j.chaos.2020.109947.

[2] A. A. Hussain, O. Bouachir, F. Al-Turjman, M. Aloqaily: *AI Techniques for COVID-19* (2020),
doi: https://doi.org/10.1109/ACCESS.2020.3007939.

[3] T. Ishige, T. Murata, S. Taniguchi, T. Miyabe, A. Kitamura, K. Kawasaki, K. Nishimura, M. Igari, H. Matsushita: *Highly sensitive detection of SARS-CoV-2 RNA by multiplex rRT-PCR for molecular diagnosis of COVID-19 by clinical laboratories.* Clinica Chimica Acta (2020),
doi: https://doi.org/10.1016/j.cca.2020.04.023.

[4] A. J. Izenman: *Linear Discriminant Analysis. In: Modern Multivariate Statistical Techniques*, Springer Texts in Statistics. Springer, New York (2013),
doi: https://doi.org/10.1007/978-0-387-78189-1_8.

[5] S. Ludwig, A. Zarbock: *Coronaviruses and SARS-CoV-2*, Anesthesia & Analgesia (2020),
doi: https://doi.org/10.1213/ane.0000000000004845.

[6] G. Mak, K. Cheng, S. Lau, K. Wong, C. Lau, T. Lam, C. Chan, N. Tsang: *Evaluation of rapid antigen test for detection of SARS-CoV-2 virus*, Journal of Clinical Virology (2020),
doi: https://doi.org/10.1016/j.jcv.2020.104500.

[7] B. G. Pinto, A. E. Oliveira, Y. Singh, L. Jimenez, A. N. A. Gonçalves, R. L. Ogava, R. Creighton, J. P. S. Peron, I. Nakaya: *ACE2 Expression is Increased in the Lungs of Patients with Comorbidities Associated with Severe COVID-19*, The Journal of Infectious Diseases (2020),
doi: https://doi.org/10.1101/2020.03.21.20040261.

[8] A. Rygula, K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyka, M. Baranska: *Raman spectroscopy of proteins: a review*, Wiley Online Library (2013),
doi: https://doi.org/10.1002/jrs.4335.

[9] B. A. Taha, Y. Al Mashhadany, M. H. Hafiz Mokhtar, M. S. Dzulkefly Bin Zan, N. Arsad: *An Analysis Review of Detection Coronavirus Disease 2019 (COVID-19) Based on Biosensor Application*, Sensors (2020),
doi: https://doi.org/10.3390/s20236764.

[10] Q. Tu, C. Chang: *Diagnostic applications of Raman spectroscopy*, Nanomedicine: Nanotechnology, Biology and Medicine 8.5 (2012),
doi: https://doi.org/10.1016/j.nano.2011.09.013.

[11] M. K. Tuck, D. W. Chan, D. Chia, A. K. Godwin, W. E. Grizzle, K. E. Krueger, W. Rom, M. Sanda, L. Sorbara, S. Stass, W. Wang, D. E. Brenner: *Standard Operating Procedures for Serum and Plasma Collection*, Journal of Proteome Research (2020),
doi: https://doi.org/10.1021/pr800545q.

[12] G. Yin, L. Li, S. Lu, Y. Yin, Y. Su, Y. Zeng, et al.: *Data and code on serum Raman spectroscopy as an efficient primary screening of coronavirus disease in 2019 (COVID-19)*, Nanomedicine: Nanotechnology, Biology and Medicine (2020),
doi: https://doi.org/10.6084/m9.figshare.12159924.v1.