Communication regarding the article “An efficient primary screening COVID-19 by serum Raman spectroscopy”

Felipe Soares ©  |  Gabrielli Yamashita

Industrial Engineering, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

Correspondence
Felipe Soares, Industrial Engineering, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.
Email: felipe.soares@ufrgs.br

Abstract
When performing computational modeling and machine learning experiments, it is imperative to follow a protocol that minimizes bias. In this communication, we share our concerns regarding the article “An efficient primary screening COVID-19 by serum Raman spectroscopy” published in this journal. We consider that the authors may have inadvertently biased their results by not guaranteeing complete independence of test samples from the training data. We corroborate our point by reproducing the experiment with the available data, showing that if full independence of the test set was ensured, the reported results should be lower. We ask the authors to provide more information regarding their article, as well as making available all code used to generate their results. Our experiments are available at https://doi.org/10.6084/m9.figshare.14124356.

KEYWORDS
COVID-19, machine learning, Raman spectroscopy, scientific rigor, screening

1 | INTRODUCTION

Due to the COVID-19 pandemic, research and publications have aimed at providing new diagnostic tools for this disease,\cite{1,2} The gold standard reverse transcription polymerase chain reaction (RT-PCR) may not be extensively available in some countries or not affordable in developing countries. We salute any effort that may help tackle this pandemic, but we as scientists should always be aware of scientific rigor and be sure that bias is reduced during experimentation.

In light of that, we would like to address some methodological aspects regarding the article “An efficient primary screening COVID-19 by serum Raman spectroscopy”,\cite{3} published in this journal. The authors proposed using Raman spectra from human blood serum to provide a screening tool regarding COVID-19. They enrolled 177 patients for this study from three different groups: healthy individuals, suspected cases, and COVID-19 positive.

For the analysis, they recorded spectra in the range of 600–1800 cm\(^{-1}\). Three experimenters recorded five times each sample, resulting in 15 spectra for each subject. The authors mentioned a total of 2355 spectra. After that, 30\% of spectra (not patients) were set aside for testing, while the remaining 70\% for training. The authors performed feature selection using analysis of variance (ANOVA) to identify the most relevant wavelengths and then used an SVM classifier to build the discriminant system. They reported accuracy values of 0.87 for COVID-19 versus suspected and 0.91 for COVID-19 versus healthy control.

The previous results on the hold-out set would be impressive. However, we consider that they are likely to be biased due to methodological concerns, which we will now describe, mainly related to violation of training and development/test set independence. Our analysis is based on the published article and their open-access data and code made available at Figshare (https://doi.org/10.6084/m9.figshare.12159924.v1). We would like to draw special attention to this assumption that the codes found in the...
Figshare are fragments of the proposed approach by the original authors. We tried contacting the original authors, but we received no answer. Thus, we decided to replicate the analysis by following the fragments and reimplementing it, then comparing the results. In addition, the code at (https://doi.org/10.6084/m9.figshare.12159924.v1) was not even running without errors, thus we had to extensive rely on cross-referencing the code fragments and the paper description. We also tried to keep traceability between our implementation and the original released one.

2 | **CONCERN 1: SUPERVISED WAVELENGTH SELECTION WITH ALL DATA OR SAMPLE LEAKAGE**

From what was described in the paper (Section 2.4) and the code on Figshare, the ANOVA is carried out on all available data, not only the training set. That means that information from samples on the test set is being “seen” to the feature selection process, which could by itself introduce a bias towards enhancing classification accuracy.

Hastie et al.\cite{3} explicitly state that first screening the predictors, selecting the most relevant ones, and then cross-validate the classification model gives an unfair advantage. Hastie et al.\cite{3} say that “Leaving samples out after the variables have been selected does not correctly mimic the application of the classifier to a completely independent test set, because the predictors ‘have already seen’ the left-out samples.”

However, when inspecting the code on Figshare, it does seem that the authors performed the hold-out process before conducting the ANOVA; however, later on in the code, they reshuflle the complete dataset and perform a new hold-out for the classification. Thus, the same case explained above occurs; it is not ensured that the test set samples are not “seen” during the ANOVA-based feature selection process.

3 | **CONCERN 2: DATA FOR CROSS-VALIDATION MAY NOT GROUPED BY PATIENTS**

The authors stated that a 70/30% cross-validation was performed and repeated 50 times. They also said that “To ensure the independence of the data, the random sampling process guaranteed that the spectra data were used to establish the model and for model test from completely different samples.” By checking the published code, we found that the data are being randomly assigned at spectra level, not patient level. Thus, if the initially released code does reflect the article’s description, we consider that the way they validated their model goes against their assertion of guaranteeing spectra data independence, and it is not enough to mitigate bias.

The article “Common mistakes in cross-validating classification models,”\cite{5} which the authors cited to corroborate their course of action, shows exactly their method described in Sections 2.4 and 3.1 as an example of bias.

To guarantee the independence of the test set, data should be grouped by patients, then patients split into training and test, and finally, their spectra assigned to each partition.

By shuffling all spectra and then performing the cross-validation sampling, they did not ensure that information from the test set was not leaked to the training one via replicates of the same serum, leading to overfitting. Even though due to the heterogeneous nature of human serum and spectra acquisition, a high correlation is expected between different samples from the same subject taken at the same time.

Our point is also corroborated by empirical results from Guo et al.\cite{5} who state that “[…] the dataset should be split at the highest hierarchical level to avoid the overestimation of classification models.” In this study, the highest level would be patient level, not sample or spectra level. Besides, Guo et al.\cite{5} state that by not following a replicate (or this case patient) level cross-validation, the information within the validation dataset is implicitly used during the model construction, violating the independence condition of training and validation.

Considering a hypothetical patient A, with three replicates (A1, A2, and A3), our point is that all three replicates should either lie in training or development, not leaving room for A1 being on training, and (A2 and A3) in testing, since they come from the same patient.

In Figure 1, we depict what is our view from the published article and the code fragments on Figshare of the employed process. Then, we also show what we consider would be a more appropriate way of performing validation, ensuring that there is no sample leakage (i.e., one spectrum of patient A in training set, and another spectrum from the same patient on development set, since the authors mentioned that there were replicates from the same serum sample).

4 | **SIMULATION**

We now try to replicate the authors’ experiments with their publicly available data and show the impact that data leakage and lack of independence of training/test
had. Given that the code provided by the authors that is possibly linked to the publication in question did not run at all (nor completely mirrored what was described in the research paper), we have to reimplement it. We tried to maintain fidelity as much as possible to both the original code and the research paper. In addition, we tried to provide traceability from both code and paper, such that readers can better understand our design choices and to help clarify the divergences and the weak points we try to shed light on this communication.

Although the authors did not specifically assign patient identification for each spectrum on their published data, they provided information on the number of spectra for each patient and which patients had fewer than three spectra. Thus, we implicitly generated the patient ID for the highest level splitting. We corroborated our assumption by experimenting with an LDA classifier aiming at patients’ IDs as labels.

Table 1 shows the results obtained for accuracy, sensitivity, and specificity, as these were the performance measures used by the original authors. The column “reproduced” is the reproduction of the experiment according to the original article’s description. The column “blocked by patients” aims at addressing both concerns 1 and 2 by performing cross-validation at the patient level, that is, first assigning patients to either train or test and then retrieve their spectra. The last column contains the values reported by the authors in their article. We performed a nonparametric Wilcoxon rank-sum test to assess statistical significance between the reproduced and blocked columns over 500 repetitions. All data are available at https://doi.org/10.6084/m9.figshare.14124356.

When comparing the three columns, one can notice that the larger differences in the results are between the values reported on the original paper and the reproduced ones. When considering only accuracy, the difference in COVID versus healthy is around 0.10, which is already a salient deviation. However, when focusing on sensitivity and specificity, we can see that they greatly diverge from the originally reported, in both reproduced experiments and when blocking by patient. When looking only at specificity, which is the ability to rule-out having the disease, the performance for the three comparison groups is strikingly different, going from 0.93 to 0.64, in COVID versus healthy, and from 0.86 to 0.66 in COVID versus suspected.

**Figure 1** Graphical demonstration of how we consider the appropriate way of conducting validation. Please notice that the feature selection step should be included in training, not as a separate step. In panel (a), we show how we understood that the authors performed evaluation, with leakage during feature selection and possibly during model training, and in panel (b) how we consider it should be carried out [Colour figure can be viewed at wileyonlinelibrary.com]
Looking at the reproduced results and the experiments blocked by patient, we can see that for COVID versus healthy, only sensitivity was found to be significantly different from the reproduced experiments. This does not replicate when looking at COVID versus suspected, and suspected versus healthy, where all metrics, but sensitivity in COVID versus suspected, were significantly different. Our experiments provide evidence to our claim that the authors’ experiments may be biased due to the non-independence of the test set. Furthermore, we can see that specificity values found in our experiments are not in pair with the ones reported by the authors.

As an additional experiment, we also conducted a similar analysis, by averaging the spectra at patient level, having the same number of spectra as patients. This was carried out to investigate if there is any additional benefit by averaging the spectra at patient level, rather than sample level. In Table 2, we report the results.

We can see that the results shown in Table 2 are very similar to the ones reported in Table 1, which means that further averaging the spectra at patient level may not provide additional benefit. This may be related to the nature of SVM, which tends to average support vectors, which may be implicitly already “averaging” spectra during training, if such samples are support vectors. Thus, even if that was the path followed by the original authors, it would still not have accounted for the discrepancy in the published results.

Overall, we consider that one of the main outcomes of our analyses is that we were unable to completely reproduce the original results published in Guo et al.,[5] although we still found evidence that there’s a significant difference in the reported metrics when considering their described feature selection method and by not blocking the experiments by patient. This happens primarily in COVID versus suspected, and suspected versus healthy group. The reason for that, we hypothesize, is that the signal-to-noise ratio is greater in infected patients, making it easier to both identify the most relevant features, and also perform the final classification.

### Table 1
Classification performance on reproduced experiment and unbiased cross-validation splitting with experiments blocked at patient level

| Class          | Parameter | Reproduced      | Blocked by patient | Original reported |
|----------------|-----------|------------------|--------------------|-------------------|
|                | Accuracy  | 0.81 ± 0.12      | 0.82 ± 0.10        | 0.91 ± 0.04       |
| COVID vs. Healthy | Sensitivity | 0.97 ± 0.06      | 0.95 ± 0.07        | 0.89 ± 0.07       |
|                | Specificity | 0.64 ± 0.26      | 0.68 ± 0.21        | 0.93 ± 0.06       |
| COVID vs. Suspected | Accuracy | 0.81 ± 0.12      | 0.76 ± 0.12        | 0.87 ± 0.05       |
|                | Sensitivity | 0.97 ± 0.07      | 0.96 ± 0.08        | 0.89 ± 0.08       |
|                | Specificity | 0.66 ± 0.25      | 0.56 ± 0.24        | 0.86 ± 0.09       |
| Suspected vs. Healthy | Accuracy | 0.73 ± 0.11      | 0.64 ± 0.09        | 0.69 ± 0.05       |
|                | Sensitivity | 0.88 ± 0.11      | 0.82 ± 0.20        | 0.70 ± 0.09       |
|                | Specificity | 0.57 ± 0.26      | 0.45 ± 0.27        | 0.66 ± 0.09       |

Note: Values in bold represent statistically significant difference between the reproduced and blocked columns. All p values were less than 0.01.

### Table 2
Classification performance on patient average spectra for the replication of the original article and the experiments with blocking by patients

| Class          | Parameter | Reproduced      | Blocked by patient |
|----------------|-----------|------------------|--------------------|
| COVID vs. Healthy | Accuracy | 0.82 ± 0.12      | 0.83 ± 0.12        |
|                | Sensitivity | 0.93 ± 0.08      | 0.94 ± 0.10        |
|                | Specificity | 0.71 ± 0.26      | 0.72 ± 0.25        |
| COVID vs. Suspected | Accuracy | 0.81 ± 0.12      | 0.84 ± 0.09        |
|                | Sensitivity | 0.83 ± 0.22      | 0.76 ± 0.22        |
|                | Specificity | 0.80 ± 0.24      | 0.92 ± 0.08        |
| Suspected vs. Healthy | Accuracy | 0.65 ± 0.11      | 0.66 ± 0.10        |
|                | Sensitivity | 0.80 ± 0.16      | 0.80 ± 0.16        |
|                | Specificity | 0.50 ± 0.28      | 0.50 ± 0.26        |

5 | CONCLUSION

In this communication, we aimed at questioning the methodological procedures followed by Yin et al.[3] in their
published article. Our main goal is to shed light on the importance of following strict protocols that guarantee test set independence when training machine learning algorithms. Due to their ability to learn complex relationships, it is not hard to overfit a model and bias the final result. The authors did not explicitly define their data processing pipeline, as carried out in related works,\textsuperscript{[6,7]} thus leaving room for multiple interpretations. Besides, the early code released on Figshare points towards a different direction from what was described in the article.

To the best of our ability, we tried to reproduce the methods laid out in their original paper, because we had to rewrite the whole code as it was not even running. We found results pointing out that our reproduced experiments perform much lower than what was originally presented. When comparing our reproduced experiments with the approach we consider would reduce the bias (i.e., feature selection inside the training data and blocking the experiments by patients), the difference in results is not that accentuated.

We ask the original authors\textsuperscript{[3]} to provide more information about their methodology, possibly the final code used to generate the data on the paper, and the spectra of all subjects, with their respective identification. We believe that proper scrutiny on research can lead to reliable experiments and beneficial results.

ACKNOWLEDGMENTS
Felipe Soares would like to acknowledge AWS Diagnostic Development Initiative (DDI) initiative for providing computational resources.

DATA AVAILABILITY STATEMENT
Our experiments and data are available at https://doi.org/10.6084/m9.figshare.14124356.

ORCID
Felipe Soares © https://orcid.org/0000-0002-2837-1853

ENDNOTE
* Post-acceptance note: The original authors mentioned that the data on that particular Figshare link was not of the same batch as the one used in the published article in this journal. We would like, however, to point out that our concerns still apply and that we should always seek to ensure our data processing is as unbiased as possible. In addition, to minimize ambiguities in the description of experiments, authors should seek to publish the source code alongside the manuscript or provide pseudocodes for better explanation.

REFERENCES
\[1\] T. Hou, W. Zeng, M. Yang, W. Chen, L. Ren, J. Ai, J. Wu, Y. Liao, X. Gou, Y. Li, X. Wang, H. Su, B. Gu, J. Wang, T. Xu, \textit{PLoS Pathog.} \textbf{2020}, \textit{16}, e1008705.
\[2\] F. Soares, A. Villavicencio, F. S. Fogliatto, M. H. Pitombeira Rigatto, M. José Anzanello, M. A. P. Idiart, M. Stevenson, \textit{bioRxiv}, \textit{2020}.
\[3\] G. Yin, L. Li, S. Lu, Y. Yin, Y. Su, Y. Zeng, M. Luo, M. Ma, H. Zhou, L. Orlandini, D. Yao, G. Liu, J. Lang, \textit{Journal of Raman Spectroscopy} \textbf{2021}, 52(5), 949. https://doi.org/10.1002/jrs.6080
\[4\] T. Hastie, R. Tibshirani, J. Friedman, \textit{The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition}, 2, 1. Springer Science & Business Media, New York, NY 2009.
\[5\] S. Guo, T. Bocklitz, U. Neugebauer, J. Popp, \textit{Anal. Methods}, \textbf{2017}, \textit{9}, 4410.
\[6\] F. Soares, K. Becker, M. J. Anzanello, \textit{Artif. Intell. Med.} \textbf{2017}, \textit{82}, 1.
\[7\] S. Feng, D. Lin, J. Lin, B. Li, Z. Huang, G. Chen, W. Zhang, L. Wang, J. Pan, R. Chen, H. Zeng, \textit{Analyst} \textbf{2013}, \textit{138}, 3967.

\textbf{How to cite this article:} F. Soares, G. Yamashita, \textit{J Raman Spectrosc} \textbf{2022}, \textit{53}(7), 1342. https://doi.org/10.1002/jrs.6330