MDPI

Article

Topological Machine Learning for Discriminative Spectral Band Identification in Raman Spectroscopy of Pathological Samples

Francesco Conti[1,2], **Davide Moroni**[1], **Maria Antonietta Pascali**[1]

[1] Institute of Information Science and Technologies "A. Faedo", National Research Council, Pisa, Italy
[2] National Institute for Research in Digital Science and Technology, DataShape, Sophia Antipolis, France
* Correspondence: francesco.conti@inria.fr

Abstract: In the field of Raman spectroscopy (RS), particularly when working with biological samples, identifying the chemical compounds most involved in specific pathologies is of critical importance for pathologists. The correlation between chemical substances present in biological tissue and pathology can contribute not only to a deeper understanding of the disease itself but also to the development of novel artificial intelligence-based diagnostic methodologies. Motivated by these clinical challenges, we propose a method to identify the most discriminative spectral bands by leveraging the synergy between Topological Machine Learning (TML) and Raman Spectroscopy. The intrinsic explainability of part of the TML pipeline can indeed play a key role in the detection of such spectral bands, e.g. the proteins most associated with the disease. In order to evaluate the performance of our method, we apply it to three case studies: the RS of biological tissue related to the chondrogenic bone tumors, the RS of cerebrospinal fluid associated with Alzheimer's disease and the RS of pancreatic tissue. The results obtained with our method are promising in pinpointing which spectral bands are most relevant for diagnosis, but they also highlight the need for further investigation.

Keywords: Raman, Machine Learning, Topological Data Analysis, Explainable Artificial Intelligence

1. Introduction

Raman spectroscopy (RS) is a spectroscopic technique used to provide a structural fingerprint of a sample. It is based on the evaluation of the inelastic scattering process in which photons incident on the sample transfer energy to or from molecular vibrational modes. Since the involved energies are relatively low, RS is applicable for non-destructive analysis; hence, it is compatible with in vivo or in vitro, making it suitable for many applications in the biological realm [1–4].

In a nutshell, different bands of the Raman spectra represent specific molecular movements and rotational states, providing an insight into molecular behaviour and composition; on the other hand, despite the chemical coherence of RS, when imaging biological samples, the sources of information are always multiple, and the most prominent ones may be hidden or obscured by other spurious signals.

Therefore, advanced data filtering and processing methods have been used to achieve a fast and robust interpretation of the spectra in various application fields, as well as machine learning methods to better understand RS and extract meaningful features from them. In this perspective, we propose to increase the interpretability of the models of Topological Machine Learning (TML, [5]) trained on three specific case studies through a band importance analysis for data of RS.

Received:
Revised:
Accepted:
Published:

Citation: . Topological Machine Learning for Discriminative Spectral Band Identification in Raman Spectroscopy . *Journal Not Specified* **2025**, *1*, 0. https://doi.org/

2. Materials and Methods

A framework inspired by RISE [6], which is a post hoc AI interpretability method, has been developed to identify the most discriminative spectral bands in Raman spectroscopy data. Our approach empirically evaluates spectral band importance through the quantitative assessment of their contribution to model performance:

- **Input:** Raman spectra of biological samples;
- **Preprocessing and feature extraction:** Same as original model (when applicable);
- **Sliding window masking:** Iterative masking with 10-width window and 5 stride to identify critical spectral regions;
- **Importance estimation:** Performance drop on masked spectra indicates diagnostically relevant regions.

Figure 1 sketches the framework overview. We used a Topological Machine Learning pipeline introduced in [5], and successfully applied in Raman spectroscopy [7,8], leveraging its input dimensionality invariance.

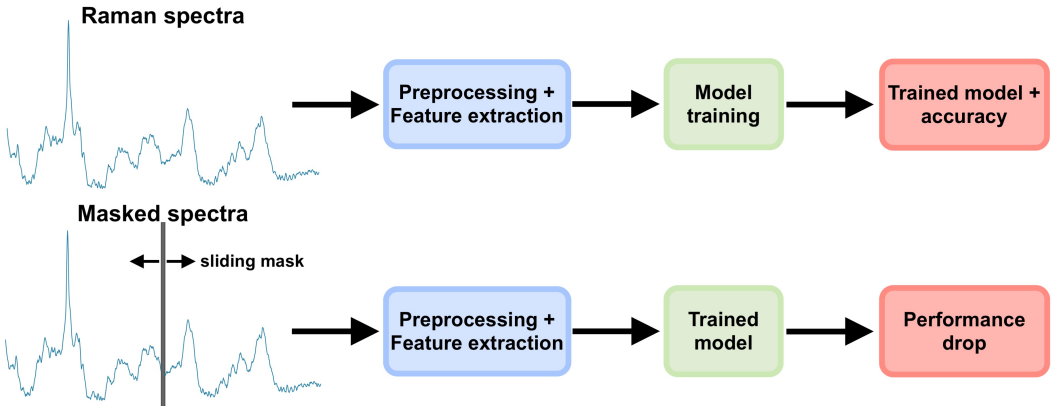

Figure 1. Identifying critical spectral regions in Raman data via sliding band masking, with band width of 10 cm^{-1} and stride of 5 cm^{-1}. First row: classical ML pipeline for RS. Second row: masked regions causing model performance drops reveal diagnostically relevant bands.

3. Results

We present the discriminative spectral band identification for our three case studies: the Alzheimer's disease detection in Figure 2, the chondrogenic cancer grading in Figure 3, and the pancreatic ductal adenocarcinoma (PDA) classification in Figure 4. For chondrogenic and pancreatic cases, we show critical regions for both binary and multi-label classification. Notably, the binary classification's critical regions remain important in multi-label scenarios (colored accordingly), demonstrating model consistency. In these figures, the results are plotted combining:

- A representative Raman spectrum (intensity a.u. *vs.* wavenumber cm^{-1});
- Performance drop curve (accuracy loss percentage *vs.* wavenumber cm^{-1}).

This dual representation links spectral features to diagnostic importance, potentially identifying relevant biomolecules. For each experiment, we set an arbitrary threshold marking significant performance degradation.

- **Alzheimer's:** Performance drop threshold: 30%. Key regions: 904-956 cm^{-1}, 1146-1198 cm^{-1}, and 1224-1313 cm^{-1} peaks.
- **Chondrogenic:** Performance drop threshold: 3%. Primary peak: 997-1016 cm^{-1}. For 4-class: additional peaks at 725-748 cm^{-1}, 997-1072 cm^{-1}, and 1384-1394 cm^{-1}.
- **PDA:** Performance drop threshold: 30% (3-class) and 45% (binary)s. Key features: 1250-1302 cm^{-1} peak and 958-973 cm^{-1} valley.

Minor graphical adjustments were made for better visualization. 69

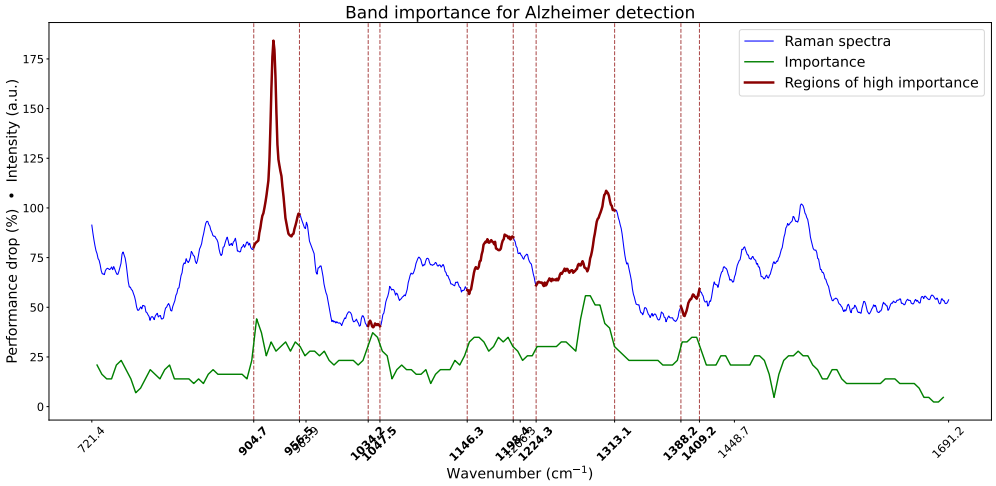

Figure 2. Discriminative spectral bands for Alzheimer's detection (904-956, 1146-1198, and 1224-1313 cm^{-1}) showing >30% performance drop when masked. Important bands are colored in red.

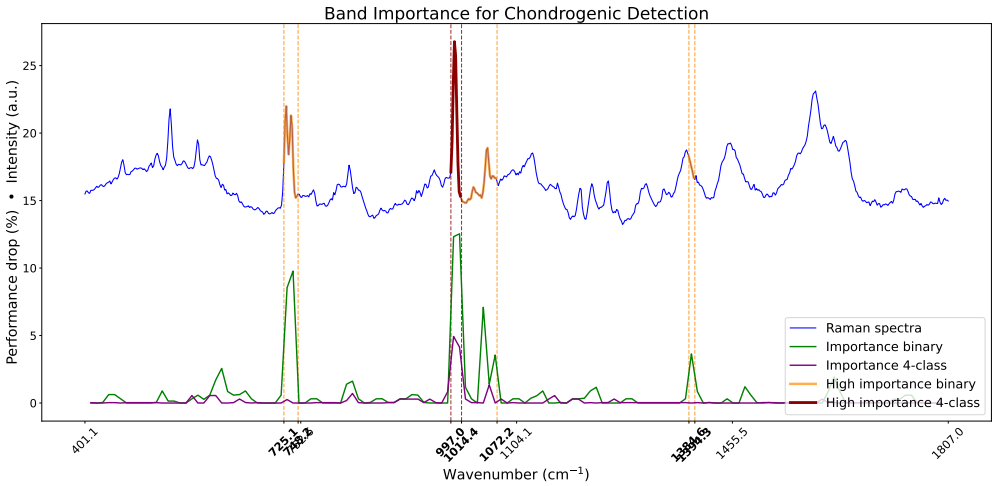

Figure 3. Critical Raman peaks for chondrogenic cancer detection (997-1016 cm^{-1} in binary classification, extending to 725-748 and 1384-1394 cm^{-1} for multi-class) showing >3% performance drop when masked. Overlapping important regions are shown in binary classification colors.

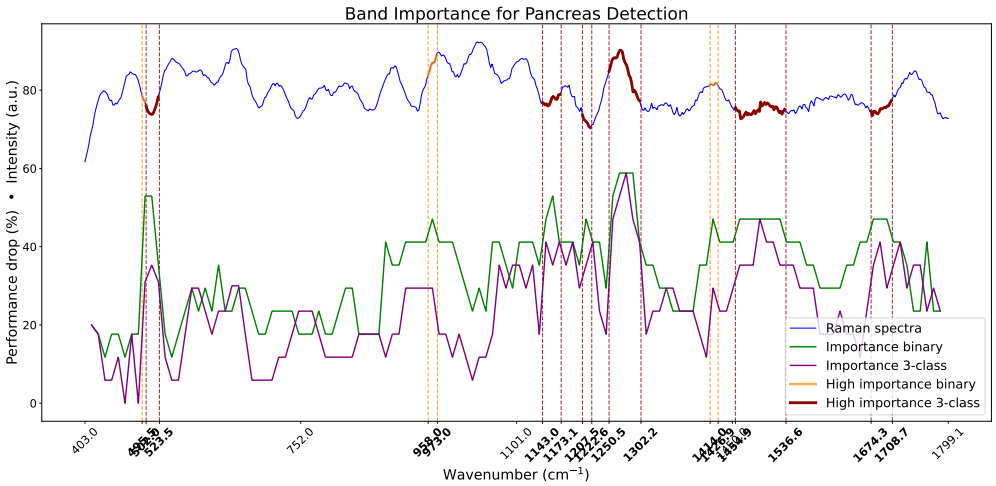

Figure 4. Diagnostically relevant spectral features for pancreatic adenocarcinoma classification (1250-1302 cm^{-1} peak and 958-973 cm^{-1} valley) showing >30% (multi-class) and >45% (binary) performance drop when masked.

4. Conclusions

This preliminary study presents an AI framework for Raman spectroscopy-based disease detection, able to enhance the explainability of the classification model, assessing the band importance of RS for each case study. Unlike conventional peak analysis, our approach reveals that diagnostic information may be distributed across broad spectral regions rather than isolated peaks, suggesting complex biomolecular interactions underlying disease signatures. The consistent identification of important spectral features across different classification tasks demonstrates the robustness of the method.These findings highlight the potential of combining topological data analysis with explainability techniques to bridge the gap between machine learning predictions and clinical interpretability in spectroscopic diagnostics. While promising, further validation is needed to establish the clinical relevance of the identified spectral patterns.

Author Contributions: Conceptualization and methodology, all; software, FC; validation and data curation, all; writing, FC and MAP; visualization, FC; supervision, DM; All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Regione Toscana through the PRAMA Project under Grant Ricerca Salute 2018.

Acknowledgments: The joint laboratory BIOICT Lab, Mario D'Acunto (IBF-CNR), Gianmarco Lazzini (ISTI-CNR), Paolo Matteini (IFAC-CNR), and Marella De Angelis (IFAC-CNR) are warmly acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Haka, A.S.; Shafer-Peltier, K.E.; Fitzmaurice, M.; Crowe, J.; Dasari, R.R.; Feld, M.S. Diagnosing breast cancer by using Raman spectroscopy. *Proceedings of the National Academy of Sciences* **2005**, *102*, 12371–12376.
2. Eberhardt, K.; Stiebing, C.; Matthäus, C.; Schmitt, M.; Popp, J. Advantages and limitations of Raman spectroscopy for molecular diagnostics: an update. *Expert Review of Molecular Diagnostics* **2015**, *15*, 773–787. https://doi.org/10.1586/14737159.2015.1036744.
3. Polykretis, P.; Banchelli, M.; D'Andrea, C.; de Angelis, M.; Matteini, P. Raman Spectroscopy Techniques for the Investigation and Diagnosis of Alzheimer's Disease. *FBS* **2022**, *14*, 22–null. https://doi.org/10.31083/j.fbs1403022.
4. Lazzini, G.; Gaeta, R.; Pollina, L.E.; Comandatore, A.; Furbetta, N.; Morelli, L.; D'Acunto, M. Raman spectroscopy based diagnosis of pancreatic ductal adenocarcinoma. *Scientific Reports* **2025**, *15*, 13240.
5. Conti, F.; Moroni, D.; Pascali, M.A. A topological machine learning pipeline for classification. *Mathematics* **2022**, *10*, 3086.
6. Petsiuk, V.; Das, A.; Saenko, K. RisE: Randomized input sampling for explanation of black-box models. In Proceedings of the British Machine Vision Conference 2018, BMVC 2018, 2019.
7. Conti, F.; D'Acunto, M.; Caudai, C.; Colantonio, S.; Gaeta, R.; Moroni, D.; Pascali, M.A. Raman spectroscopy and topological machine learning for cancer grading. *Scientific reports* **2023**, *13*, 7282.
8. Conti, F.; Banchelli, M.; Bessi, V.; Cecchi, C.; Chiti, F.; Colantonio, S.; D'Andrea, C.; de Angelis, M.; Moroni, D.; Nacmias, B.; et al. Harnessing topological machine learning in Raman spectroscopy: Perspectives for Alzheimer's disease detection via cerebrospinal fluid analysis. *Journal of the Franklin Institute* **2024**, *361*, 107249.