MDPI

Article

Topological machine learning for Raman spectroscopy: perspectives for pancreatic diseases

Francesco Conti[1,2], Gianmarco Lazzini[1], Raffaele Gaeta[4], Luca Emanuele Pollina[4], Annalisa Comandatore[5], Niccolò Furbetta[5], Luca Morelli[5], Mario D'Acunto[3], Davide Moroni[1], Maria Antonietta Pascali[1]

[1] Institute of Information Science and Technologies "A. Faedo", National Research Council, Pisa, PI, Italy
[2] National Institute for Research in Digital Science and Technology, DataShape, Sophia Antipolis, France
[3] Second Division of Surgical Pathology, University Hospital of Pisa, Pisa, Italy
[4] General Surgery Unit, Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy
[5] Institute of Biophysics, National Research Council of Italy, Pisa, PI, Italy

Abstract: The analysis of tissue samples from 17 subjects clinically diagnosed with chronic pancreatitis, ductal adenocarcinoma, or classified as controls has been collected and analyzed by Raman spectroscopy (RS). Such data are classified using a recent methodology which combines machine learning with advanced Topological Data Analysis (TDA) techniques, known as Topological Machine Learning (TML). A classification accuracy of 82% was achieved following a cross-validation scheme with patient stratification, suggesting that the combination of RS and topological data analysis holds significant potential for distinguishing between the three diagnostic categories. When restricted to binary classification (cancer *vs.* no cancer), performance increases to 88%. This approach offers a promising and fast method to support clinical diagnoses, potentially improving diagnostic accuracy and patient outcomes.

Keywords: Raman spectroscopy, Pancreas diseases, Topological machine learning, Topological data analysis

1. Introduction

Chronic pancreatitis (pc) and pancreatic ductal adenocarcinoma (dag) are severe pancreatic disorders with significant global morbidity and mortality. Globally, pc affects 5–12 per 100,000 individuals annually, while dag accounts for $\sim 495,000$ new cases each year, with projections suggesting pancreatic cancer may become the second-leading cause of cancer deaths by 2030 [1].

Furthermore, current diagnostic suffers from various limitations:

· Imaging (CT/MRI): Conventional imaging methods have limited sensitivity for early chronic pancreatitis and small pancreatic ductal adenocarcinoma, as current standards lack universal criteria to detect early parenchymal changes and rely heavily on ductal abnormalities captured by the Cambridge Classification, which misses subtle early-stage features [2];

· Biomarkers (CA 19-9): Despite its widespread clinical use, CA 19-9 exhibits stage-dependent sensitivity in dag, with pooled sensitivity of 78.2% in all-stage analysis, but only 48% for localized T1 tumors, while specificity is compromised by false positives in $15-20\%$ of benign biliary obstructions and undetectable in $5-10\%$ of Lewis antigen-negative individuals [3];

Received:
Revised:
Accepted:
Published:

Citation: . Topological machine learning for Raman spectroscopy. *Journal Not Specified* **2025**, *1*, 0. https://doi.org/

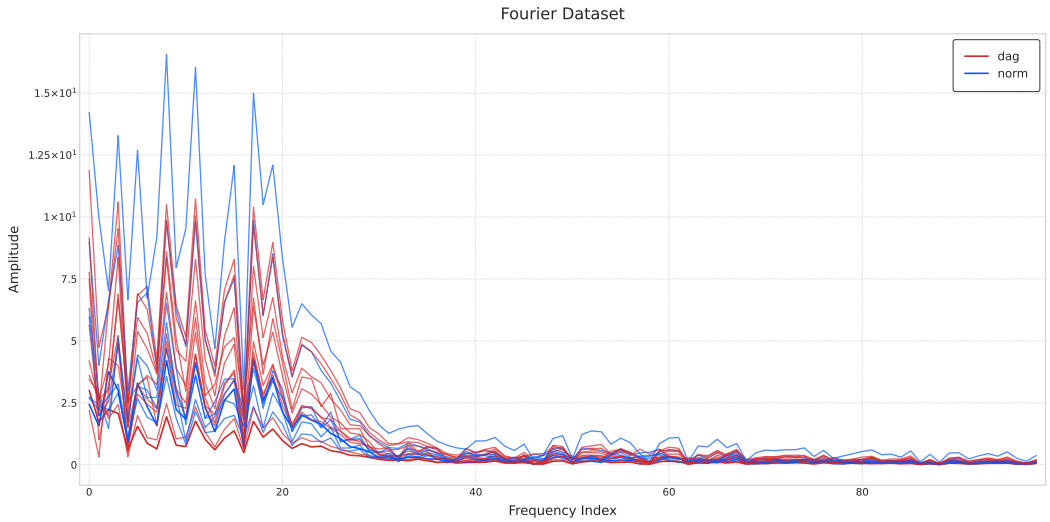

Figure 1. Fourier-transformed dataset showing the frequency components of the processed signals.

· Histopathology (EUS-FNA): Despite being the gold standard, EUS-guided FNA histopathology is inherently invasive, carrying risks of pancreatitis (1–2%) and bleeding, while exhibiting a 10–20% non-diagnostic rate due to insufficient cellularity, sampling error, or obscuring blood—with accuracy further compromised in early-stage lesions (<1 cm) or fibrotic pancreatitis [4].

Raman spectroscopy (RS) has emerged as a rapid, label-free, and minimally invasive diagnostic tool capable of detecting molecular alterations in tissues and biofluids with high specificity [5]. More importantly, in oncology, RS may enable discrimination of malignant from benign conditions based on unique biomolecular signatures [6,7] or even provide support for cancer grading when coupled with topological machine learning (TML) [8]. Recently, RS and AI have been proposed as a diagnostic tool in the case of pancreatic cancer [9].

We propose to combine RS-based vibrational fingerprinting of pancreatic biopsies with TML analysis to improve dag/pc classification. Our preliminary results show that topological data analysis combined with machine learning may serve as a valuable computational approach for this diagnostic task. Further validation in larger cohorts is warranted to standardize protocols and evaluate clinical translatability.

2. Materials and Methods

The study involved an ensemble of 14 μm-thick Formalin-Fixed and Paraffin-Embedded (FFPE) sections of pancreatic solid biopsies corresponding to Normal Pancreas (4 biopsies), Chronic Pancreatitis (3 biopsies), Ductal Adenocarcinoma of grade 2 (6 biopsies), and Ductal Adenocarcinoma of grade 3 (4 biopsies). For further details about the fabrication process, the optical system, and the RS acquisition procedure we refer to [9]. The result of the acquisition phase was a dataset of 2592 spectra, acquired in the spectral interval between 400 and 1800 cm^{-1}, with a spectral resolution of ~ 2 cm^{-1}. The spectra were preprocessed according to a standard procedure, and all spectra from the same biopsy with the same histotype were averaged, producing a dataset of 17 spectra.

Spectra were preprocessed by applying a baseline correction and smoothing as in [10], and a Fourier transform. Figure 1 displays the dataset after FT.

The two grades of ductal adenocarcinoma previously collected have been merged into a single category 'dag'. We highlight that the dataset is imbalanced, with 10 'dag' patients, 4 control 'norm' patients, and 3 chronic pancreatitis 'pc' patients. Subsequently, a binary classification task with 'dag' *vs* 'no dag' division has been performed, in order to discern a

cancer or non-cancer scenario. For this reason, in both settings, the baseline of the naive classifier that classifies according to the most frequent class is 59%. The Fourier-transformed dataset is processed using TML (see [11] for more information). The pipeline applies a lower-star filtration to extract the Persistence Diagrams (PDs). Since the data consists of 1D spectra, the only non-trivial homology group is H_0. The PDs are vectorized using standard techniques from the literature (e.g. [12]). These vectors are then fed into one of the following machine learning (ML) classifiers: a Support Vector Classifier (SVC), Random Forest Classifier (RFC), or Ridge Classifier. The pipeline is evaluated using Leave-One-Out cross-validation.

3. Results

To compare our method with others, we used ML to classify the average Raman spectra and a CNN to classify the same dataset, but augmented in the same fashion as [13]. For the 3-class classification, ML achieved an accuracy value of 0.58 (baseline accuracy 0.59) while the CNN achieved 0.47 (baseline accuracy 0.33). For the binary classification, ML (SVM) achieved an accuracy value of 0.76 (baseline accuracy 0.59), while the CNN achieved 0.65 (baseline acc. 0.5). Experiments showed that TML approach got the highest accuracy: 0.82 for the 3-class classification and 0.88 for the binary one (both using PL descriptors + a Ridge classifier). We report the confusion matrices for the binary classification in Table 1 and for the 3-class classification in Table 2.

For the 3-class problem, the baseline accuracy is 0.59, matching the performance of the traditional ML model. The CNN-based method, adapted from [13], is trained and tested on the augmented dataset; thus, the baseline accuracy is 0.33 for the 3-class problem and 0.5 for the binary problem. Despite the slightly different setting, our TML method outperforms a state-of-the-art CNN specifically designed for spectral analysis, suggesting that topological features capture more effectively discriminative patterns in spectra.

In binary classification, standard ML attains 0.76, while the CNN achieves 0.65, again underperforming compared to TML (0.88). This aligns with prior observations where conventional preprocessing or deep learning methods struggled with low SNRs.

3.1. CNN Implementation Details

We evaluated the CNN under a transfer learning regime: all but the last layer of a pre-trained CNN [13] were frozen, and the model was fine-tuned on our augmented data (epochs: 20; optimizer: Adam; learning rate: 3×10^{-4}). After splitting the data into 70% training and 30% testing, augmentation was performed via convex combinations of spectra from the same class, ensuring no data leakage between training and testing sets.

Table 1. Binary confusion matrix

Predicted:	DAG	No DAG
True DAG	9	1
True No DAG	1	6

Table 2. 3-class confusion matrix

Predicted:	DAG	PC	NORM
True DAG	9	0	1
True PC	0	3	0
True NORM	1	1	2

4. Conclusions

The results strongly suggest that Raman spectroscopy (RS) combined with topological analysis may offer an effective approach to discriminate dag from chronic pancreatitis or normal tissue in the 3-class setting, and dag from non-dag cases in the binary classification task. Notably, our TML frameworkachieves robust performance while requiring minimal parameter tuning or preprocessing steps. This makes the methodology particularly suitable for potential integration into automated diagnostic pipelines, where consistency and ease of use are critical for clinical adoption. If confirmed in larger studies, this RS-TML pipeline

could evolve into a decision-support tool for the diagnosis of pancreatic cancer, potentially reducing reliance on invasive procedures.

Author Contributions: Conceptualization, FC, GL, MDA, DM, MAP RG; methodology, FC, DM, MAP, GL, MDA; software, FC, MAP, DM; sample acquisition (surgical), NF, LM; data preparation (glass slides), RG, LEP, AC; data acquisition (RAMAN), GL, MDA; writing—original draft preparation, FC, GL; writing—review and editing, FC, GL, DM, MAP; supervision, DM, MDA. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: NanoBioTLab CNR-IBF and the joint laboratory BIOICT Lab are warmly acknowledged.

Institutional Review Board Statement: This observational study was conducted in accordance with the principles of the Declaration of Helsinki and ethical approval was obtained from CEAVNO (Comitato Etico Regionale per la Sperimentazione Clinica della Toscana - sezione AREA VASTA NORD OVEST) Committee on date 16 February 2023, Protocol Code: PANOMIC; Clinic Study: "Personalized medicine of pancreatic cancer using genomics and avatars".

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

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

References

1. Rahib, L.; Smith, B.D.; Aizenberg, R.; Rosenzweig, A.B.; Fleshman, J.M.; Matrisian, L.M. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer research* **2014**, *74*, 2913–2921.
2. Tirkes, T.; Shah, Z.K.; Takahashi, N.; Grajo, J.R.; Chang, S.T.; Venkatesh, S.K.; Conwell, D.L.; Fogel, E.L.; Park, W.; Topazian, M.; et al. Reporting standards for chronic pancreatitis by using CT, MRI, and MR cholangiopancreatography: the consortium for the study of chronic pancreatitis, diabetes, and pancreatic cancer. *Radiology* **2019**, *290*, 207–215.
3. E Poruk, K.; Z Gay, D.; Brown, K.; D Mulvihill, J.; M Boucher, K.; L Scaife, C.; A Firpo, M.; J Mulvihill, S. The clinical utility of CA 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. *Current molecular medicine* **2013**, *13*, 340–351.
4. Puli, S.R.; Bechtold, M.L.; Buxbaum, J.L.; Eloubeidi, M.A. How good is endoscopic ultrasound–guided fine-needle aspiration in diagnosing the correct etiology for a solid pancreatic mass?: a meta-analysis and systematic review. *Pancreas* **2013**, *42*, 20–26.
5. 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.
6. Auner, G.W.; Koya, S.K.; Huang, C.; Broadbent, B.; Trexler, M.; Auner, Z.; Elias, A.; Mehne, K.C.; Brusatori, M.A. Applications of Raman spectroscopy in cancer diagnosis. *Cancer and Metastasis Reviews* **2018**, *37*, 691–717.
7. Cui, S.; Zhang, S.; Yue, S. Raman spectroscopy and imaging for cancer diagnosis. *Journal of healthcare engineering* **2018**, *2018*, 8619342.
8. 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.
9. 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.
10. 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.
11. Conti, F.; Moroni, D.; Pascali, M.A. A topological machine learning pipeline for classification. *Mathematics* **2022**, *10*, 3086.
12. Bubenik, P.; et al. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **2015**, *16*, 77–102.
13. Liu, J.; Osadchy, M.; Ashton, L.; Foster, M.; Solomon, C.J.; Gibson, S.J. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst* **2017**, *142*, 4067–4074.