Journal of the Franklin Institute 361 (2024) 107249

Contents lists available at ScienceDirect

Journal of the Franklin Institute

journal homepage: www.elsevier.com/locate/fi

Harnessing topological machine learning in Raman spectroscopy: Perspectives for Alzheimer's disease detection via cerebrospinal fluid analysis

Francesco Conti [a,b,*], Martina Banchelli [c], Valentina Bessi [d], Cristina Cecchi [e], Fabrizio Chiti [e], Sara Colantonio [a], Cristiano D'Andrea [c], Marella de Angelis [c], Davide Moroni [a], Benedetta Nacmias [d,f], Maria Antonietta Pascali [a], Sandro Sorbi [d,f], Paolo Matteini [c]

[a] Institute of Information Science and Technologies "A. Faedo", National Research Council, via G. Moruzzi 1, Pisa, 56124, PI, Italy
[b] Department of Mathematics, University of Pisa, Largo B. Pontercorvo 5, Pisa, 56126, PI, Italy
[c] Institute of Applied Physics "N. Carrara", National Research Council, Via Madonna del Piano 10, Sesto Fiorentino, 50019, FI, Italy
[d] Department of Neuroscience, Psychology, Drug Research and Child Health, University of Florence, Viale Pieraccini 6, Firenze, 50139, FI, Italy
[e] Department of Experimental and Clinical Biomedical Sciences, University of Florence, Florence, Viale Morgagni 50, Firenze, 50134, FI, Italy
[f] IRCCS Fondazione Don Carlo Gnocchi, Via di Scandicci 269, Firenze, 50143, FI, Italy

ARTICLE INFO

Keywords:
Raman spectroscopy
Cerebrospinal fluid
Alzheimer's disease
Persistent homology
Topological data analysis
Topological machine learning

ABSTRACT

The cerebrospinal fluid of 21 subjects who received a clinical diagnosis of Alzheimer's disease (AD) as well as of 22 pathological controls has been collected and analysed by Raman spectroscopy (RS). We investigated whether the Raman spectra could be used to distinguish AD from controls, after a preprocessing procedure. We applied machine learning to a set of topological descriptors extracted from the spectra, achieving a high classification accuracy of 86%. Our experimentation indicates that RS and topological analysis may be a reliable and effective combination to confirm or disprove a clinical diagnosis of Alzheimer's disease. The following steps will aim at leveraging the intrinsic interpretability of the topological data analysis to characterize the AD subtypes, e.g. by identifying the bands of the Raman spectrum relevant for AD detection, possibly increasing and/or confirming the knowledge about the precise molecular events and biological pathways behind the Alzheimer's disease.

1. Introduction

Alzheimer's disease (AD) impacts millions of people globally, emerging as a predominant neurodegenerative condition. With the aging population, projections suggest approximately 152 million individuals will grapple with Alzheimer's disease and related dementias by 2050 [1]. Currently, diagnosing AD involves a sequence of neurological assessments, for instance, those based on the National Institute of Aging - Alzheimer's Association criteria. However, a conclusive diagnosis is attainable only posthumously through brain tissue analysis conducted ex vivo. Therefore, enhancing diagnostic precision demands innovative and targeted methodologies that should also meet additional requirements, including low invasivity and cost-effectiveness.

Raman spectroscopy (RS) represents a fast and efficient diagnostic tool [2] that has found applications to several kinds of biological samples, including cellular tissues, cell lines and fluids, providing new insight into the mechanisms of pathogenesis as

* Corresponding author at: Institute of Information Science and Technologies "A. Faedo", National Research Council, via G. Moruzzi 1, Pisa, 56124, PI, Italy.
E-mail address: francesco.conti@phd.unipi.it (F. Conti).

https://doi.org/10.1016/j.jfranklin.2024.107249
Received 8 March 2024; Received in revised form 29 August 2024; Accepted 29 August 2024
Available online 6 September 2024

well as practical methods for assessing disease presence and grade. Given its reliance on the examination of biological samples, the invasiveness of RS closely aligns with the specimen collection process, generally maintaining a low level. Recently, Raman-based techniques demonstrated significant potential in identifying AD by detecting specific biomarkers in body fluids [3]. Given the increasing number of RS studies, a systematic evaluation of the accuracy of RS in the diagnosis of AD has already been performed, showing that RS is an effective and accurate tool for diagnosing AD. However, it still cannot rule out the possibility of misdiagnosis [4]. The detection of cerebrospinal fluid (CSF) biomarkers is one of the diagnostic criteria for AD [5] because CSF is more sensitive than blood or other biofluids in diagnosing AD. Therefore, RS can be used as an effective tool to analyse CSF samples, as shown previously [6,7].

In a parallel line of research, Raman spectroscopy of tissue samples has been coupled with Topological Machine Learning (TML) – an emerging area of artificial intelligence which leverages topological data analysis – to support the grading of bone cancer. In this context, TML methods have demonstrated the feasibility of a topological approach for multi-label classification [8].

In this paper, we propose a novel method supporting the diagnosis of AD based on the analysis of the CSF samples through the processing of Raman spectroscopy data via TML. This contribution refines and extends [9] in which very preliminary findings were reported. In more detail, with respect to our previous work, the methodology has been improved, the dataset of CSF has been increased in size, and further experimentation has been carried out to confirm the robustness and the efficacy of the proposed method, which showed a classification accuracy of 86%, outperforming several other approaches, including neural networks and other topology-based methods. Also, an ablation study has been carried out to establish the highly informative content of topological features in the specific domain of RS.

The manuscript structure is the following: Section 2 aims at providing a concise survey on ML methods for RS analysis and a brief introduction to topological machine learning; Section 3 describes both the data acquisition procedure (from the biological sample to the Raman spectra preprocessing) and the classification pipeline based on TML; Section 4 deals with the experimentation, and Section 5 is devoted to the discussion of results, including a band importance analysis inspired by the principles of explainable artificial intelligence, an ablation study and a comparison with other methods; finally, Section 6 concludes the paper and suggests future research perspectives.

2. Related works

In this section, we survey relevant research endeavours concerning approaches centred on machine learning for RS analysis (Section 2.1) and then provide a brief introduction, with no ambition of exhaustiveness, to the emerging TML field (Section 2.2), describing a topological pipeline suitable for data classification.

2.1. Machine learning in Raman spectroscopy

Raman spectroscopy is based on evaluating the inelastic scattering process in which photons incident on a sample transfer energy to or from molecular vibrational modes. Since each molecule's energy levels differ uniquely, Raman spectra exhibit a chemical specificity that makes them suitable for chemometrics. From one side, different bands of the spectra represent specific molecular movements and rotational states, offering an unprecedented insight into molecular behaviour. Since the involved energies are relatively low, RS is applicable for non-destructive analysis and, when considered in the realm of biological investigations, is compatible with in vivo or in vitro, making it suitable for biopsy or laboratory analysis. On the other side, although there is a chemical coherence in RS, when imaging biological samples, in practice, the sources of information are always multiple, and the most prominent ones can be hidden or obscured by other spurious signals. Therefore, advanced data processing methods and machine learning have been used to achieve fast and robust interpretation of the spectra in various application fields. First attempts have focused on extracting, thanks to machine learning, models for identifying the peaks and characteristic patterns of molecules. This is the case of the work by Haka et al. [10] aiming at distinguishing benign and malignant lesions in the breast by analysing Raman spectra obtained from ex vivo samples of tissue and fitting a linear combination model with nine features representing morphological and chemical properties of the spectra, corresponding for instance to the relative content of fat and collagen. In the following years, the attention has moved to more complex machine learning models for the analysis and classification of spectra. After suitable preprocessing, very often motivated by the necessity of removing spurious components for baseline correction or improved repeatability, the acquired Raman spectra are regarded as a whole as spectral signatures or fingerprints of the imaged samples, without confining the analysis to predetermined peaks or windows in the spectra. In this context, many studies have been conducted primarily in oncology and histopathology, where Raman methodologies aimed to detect even pre-malignant and other stages of cancer progression. The combination of Principal Component Analysis (PCA) for dimensionality reduction and the identification of significant features and Linear Discriminant Analysis (LDA) for classification has been proposed on several occasions, for instance, for addressing oesophagal high-grade dysplasia [11] and lung cancerous cell detection [12], as well as for the diagnosis and grading of chondrogenic tumours [13]. Such a simple yet effective combination has improved over the years, encompassing more complex machine learning classifiers, for instance, those based on Support Vector Machines (SVM) [14] that are capable of successfully tackling also non-linear separable classification tasks. In the last decade, the surprising success of the so-called deep learning among the possible strategies in machine learning in several sectors, notably in computer vision and medical imaging, has led to the introduction of such new methods also for RS (see, e.g. [15]). Besides the direct application of state-of-the-art Convolutional Neural Networks (CNN) to Raman spectra, new models and architectures have been conceived to deal specifically with Raman spectra. Among the very extensive literature on the subject, we mention [16] in which a Multi-feature fusion CNN (MCNN),

consisting of four one-dimensional convolutional layers, one flattening layer and two fully connected layers, is trained to diagnose thyroid dysfunction via serum analysis from fresh blood samples. A different classification framework, named Diverse Spectral Band-based deep Residual Network (DSB-ResNet) is presented in [17] and used to distinguish Tongue Squamous Cell Carcinoma (TSCC) from non-cancerous tissue, with future perspectives for intraoperative usage.

According to the tenets of deep learning, its meta-learning capabilities are expected to enable the models to effectively discern and learn the most appropriate data representations for various tasks, e.g. for RS classification. This relieves the need for extensive data preprocessing or manual crafting of features that characterize classical machine learning methods. However, this advantage is somewhat counterbalanced by the increased requirement for ample data to train deep-learning models without encountering overfitting issues. In the realm of biological applications employing RS, collecting substantial datasets for scientific research remains challenging, whereas – by contrast – for general-purpose problems, the internet serves as a vast repository of data. To address this challenge, data augmentation techniques have frequently been employed to simulate instrumental noises and diverse acquisition setups, thereby augmenting real datasets with synthetic counterparts, leading to significant enhancements in the overall performance of deep learning models [18]. The scarcity of large-scale datasets in RS has also recently been addressed in [19], wherein they introduced a substantial synthetic dataset alongside the validation of various neural network architectures using this dataset.

Finally, besides the works already cited in the introduction, the synergy between Raman spectroscopy and Machine learning is an active research topic in AD, where also non-invasive blood samples are expected to convey diagnostically relevant information, as envisaged in a study on rats [20].

2.2. Topological machine learning

TML is a field that combines methods from algebraic topology and machine learning to analyse complex data. Referring to [21,22] for a more complete and formal treatment, we sketch the basics of TML. Algebraic topology is a mathematical subject that studies the shape and structure of mathematical objects called topological spaces, a broad class to which curves, surfaces and more general spaces such as manifolds and simplicial complexes belong. Algebraic topology provides an arsenal of methods for attaching algebraic invariants to topological spaces providing synthetic and quantitative features to distinguish a topological space from another. In particular, such algebraic invariants include features such as the number of connected components, holes and voids and higher dimensional analogues that can be extracted thanks to homology theory. Such numbers are referred to as the Betti numbers; for instance, the zeroth Betti number β_0 counts the number of connected components in a topological space.

In data science, the elements of a dataset can often be regarded as topological spaces. For instance, a time series can be regarded as a one-dimensional function, and the graph Γ of the curve in the Cartesian plane might be considered as its representation as a topological space. More generally, when a sample is represented by a number of points in Euclidean space (i.e. a point cloud), there are several constructions to transform it into a simplicial complex, e.g. by the well-known Delaunay triangulation [23].

However, when given a dataset, computing the homology of its topological realizations rarely yields sufficiently interesting features. It is thus unsuitable for analysing the samples or accomplishing tasks such as data classification. Indeed, such topological features merely capture the dataset in its entirety but fail to encompass other finer characteristics. To address this issue, a theory named persistence homology has been developed [24]. The basic idea of persistent homology is not only to examine the topological space in its entirety but also to introduce suitable operations to analyse the data by scanning it progressively. In more detail, a filtration is introduced to convert a topological space into a nested sequence of topological spaces. Moving from one space to the subsequent one in the nested series, new points are added to the space, potentially changing the topological features. The choice of filtration is determined by the selection of the lens through which we should observe and study the data, often allowing for multi-scale data analysis. For example, when considering the graph of a curve Γ in the Cartesian plane, we might employ the height filtration; namely, for a specific value h of the height, we examine the points $\Gamma_h = \{(x, y) \in \Gamma \mid y < h\}$ of the graph having an ordinate value lower than h. It is possible to interpret that filtration process as the introduction of a time coordinate: at each time t, we have scanned a topological space X up to a certain level X_t; when time grows, the scanned area grows, i.e. $X_{t_1} \subset X_{t_2}$ for $t_1 < t_2$, since finally all the space has been scanned, i.e. $X_t = X$ for $t \gg 0$. Persistent homology allows us to compute the homology of every space in the filtration and, most importantly, to track the topological features over time. In this way, it gathers information about when a topological feature, such as a hole or a higher dimensional analogue, is born and when it is possibly annihilated. In practice, the list of points (b_p, d_p), representing respectively the birth and death time of each topological feature p (with $d_p \in \mathbb{R} \cup \{\infty\}$), might be collected into a multiset D which we call Persistence Diagram (PD). Depending on the topological realization and the selection of the filtration, PDs convey important and fine information that might enable the discernment of different classes within a given dataset. However, PDs, being multisets, lack a structure manageable by standard statistical analysis and machine learning methods; for instance, it is not even possible to compute expected values. To cope with this issue, methods for transforming PDs into Banach spaces have been proposed, based either on direct embedding (vectorization methods [25]) or in an implicit manner (e.g., in kernel-based methods [26]). In this paper, we restrict our attention to the first class that makes PDs directly manageable by standard machine learning classifiers, such as LDA, SVM, and CNN, by converting them into a set of conventional vectors. In [27], a complete TML pipeline for data classification has been proposed and validated on several benchmark datasets. It consists of a number of steps as schematized in Fig. 1: starting from data represented as topological spaces, a suitable filtration is selected. Based on such a filtration and applying persistent homology, data is transformed into PDs which, thanks to one or more vectorization methods, are translated into vectors. Machine learning is then applied to achieve classification and estimate the accuracy of the overall pipeline. We finally notice that there exist other competing approaches in TML based on introducing trainable topological layers inside deep learning architectures, such as PersLay [28].

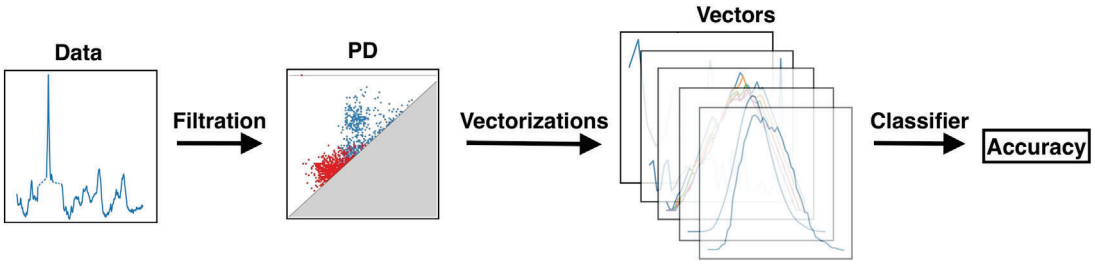

Fig. 1. Scheme of the topological machine learning pipeline. Starting from the data we produce a persistence diagram by means of a filtration, which is then vectorized through various methods and such vectors enter a machine learning algorithm which returns a classification with a certain accuracy.

3. Methods

In this section, we describe the procedure of data acquisition of the Raman spectra (Section 3.1), as well as the preprocessing steps applied to such data (Section 3.2). Next, we formally describe our classification pipeline: the first part, devoted to the extraction of topological features, explaining details about the chosen filtration and vectorization process (in Sections 3.3 and 3.4); and the second part, i.e. the classification built upon standard Machine Learning (ML) classifiers (Section 3.5).

3.1. Study population and Raman spectra acquisition

The study population is made of 43 patients, enrolled in the framework of the Bando Salute 2018 PRAMA project ("Proteomics, RAdiomics & Machine learning-integrated strategy for precision medicine for Alzheimer's"), co-funded by the Tuscany Region, with the approval of the Institutional Ethics Committee of the Careggi University Hospital Area Vasta Centro (ref. number 17918_bio). All of them showed pathological symptoms: 21 subjects have been diagnosed with AD, while the others have been considered as controls (noAD), even if diagnosed with other neurological conditions (e.g., vascular dementia, hydrocephalus and/or multiple sclerosis).

The CSF samples were collected by lumbar puncture, then immediately centrifuged at 200 g for 1 minute, 20° C and stored at −80° C until analysis [29,30]. On the day of analysis, CSF samples were thawed and centrifuged again at 4000 g for 10 minutes at 4° C. The supernatant was separated and further used for the analyses. A 2 μl drop of the sample was deposited onto a gold mirror support (ME1S-M01; Thorlabs, Inc., Newton, NJ), followed by air drying for 30 minutes and acquisition of Raman spectra from the outer ring of the dried drop. A set of 10 Raman spectra has been collected for each drop-casted sample by using a micro-Raman spectrometer (LabRam HR800 Evolution, Horiba, France) in back-scattering configuration, equipped with a laser excitation source tuned at 633 nm (6 mW power, 1 seconds integration time, 10 accumulations) and a Synapse CCD detector. Finally, to prepare the input of the classification pipeline, the preprocessing steps described in the following have been applied to the average of the ten acquisitions of RS.

3.2. Raman spectra preprocessing

The preprocessing applied to the spectra is described in detail in [31]. For self-completion, we report here the complete procedure. The original Raman spectrum contains peaks that correspond to known regions of salts present in cerebrospinal fluid. Such regions are located between $907 - 984$ and $1043 - 1117$ wavenumber and are therefore omitted from the spectra. Subsequentially, an asymmetrically reweighted penalized least squares baseline correction [32] (parameters $l = 1e + 7$ for smoothness and $p = 0.05$ for asymmetry) is applied. Finally, we applied a Savitzky Golay smoothing filter [33] implemented in the `SciPy` Python library with parameters $w = 9$ as window size and $p = 2$ as polynomial order. We refer to Fig. 2 for an example of the resulting spectra through the various preprocessing steps. Moreover, following standard literature in signal analysis [34–37], the Raman spectrum is transformed by means of the Fourier transform, the Welch transform and the Autocorrelation transform. We refer to Fig. 3 for an example of such transformations applied to a preprocessed Raman spectrum and to Fig. 4 for the entirety of the Autocorrelation dataset (left) and the average with standard deviation of the two classes (right).

3.3. From Raman spectra to persistence diagrams

Once the Raman spectrum and its transforms (Fourier, Welch, and Autocorrelation) have been computed, the topological features are extracted. To generate different topological features, we tested the performance of two filtrations. The first filtration, which is standard in persistent homology, is known as the "lower star filtration", and it basically consists of tracking the evolution of each connected component of the sublevel set of a suitable filtration function. When applied to a 1D signal, such a filtration allows to describe critical points in a stable way. The second filtration that has been considered in this work has been applied to digital images and is known as the "dilation filtration" [38]. As a first step, this filtration computes the distance (in the domain) for each point of the spectrum from the nearest point of local maximum. After the application of this distance transform, the lower star filtration is

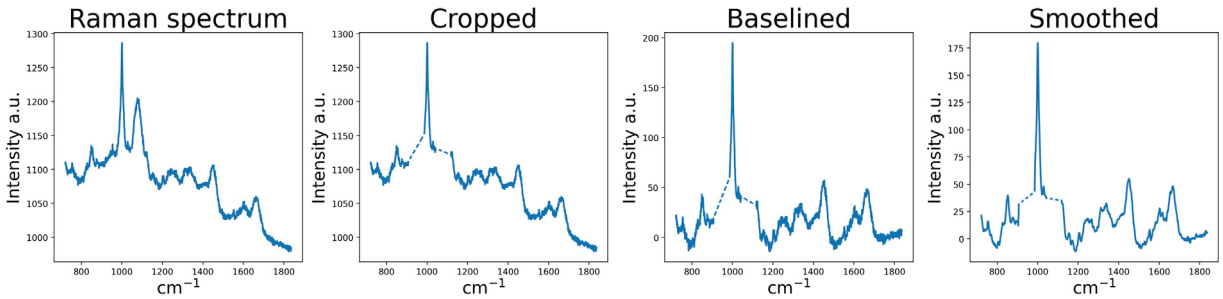

Fig. 2. Preprocessing steps applied to the average of the 10 Raman spectra acquired from each sample. The pieces of graph rendered as dashed represent the omitted regions of the spectrum.

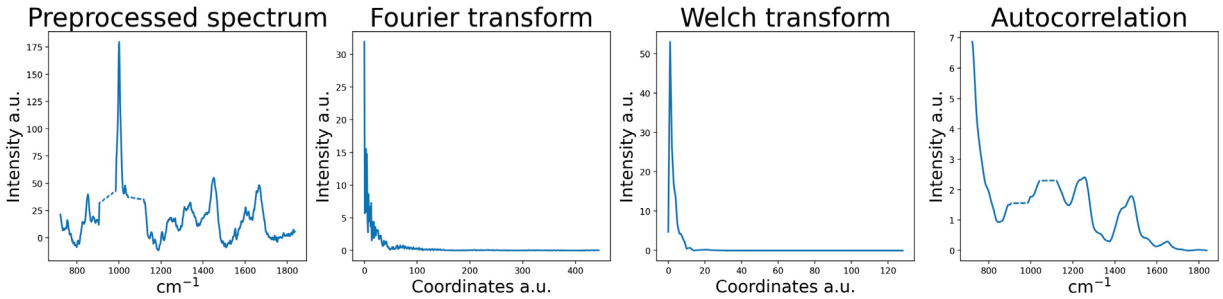

Fig. 3. Transformations applied to a preprocessed Raman spectrum.

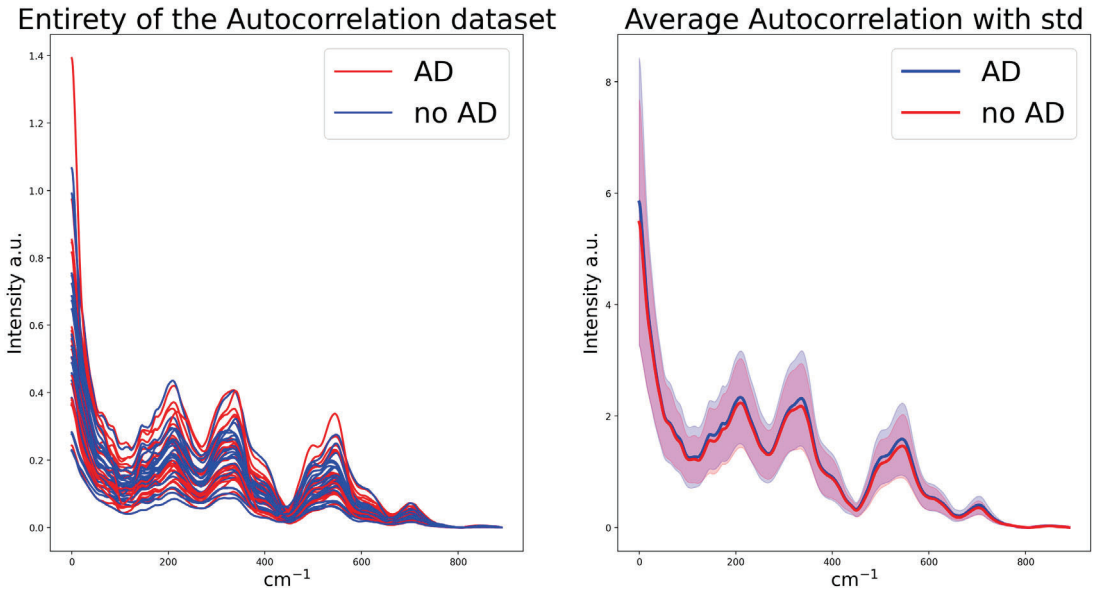

Fig. 4. Entirety of the Autocorrelation dataset (left), and average with respect to the two classes (right). Shaded areas indicate regions within plus or minus one standard deviation (std) from the mean.

performed. In this work, it has been adapted for 1-dimensional signals. We refer to Fig. 5 for an example of this transform applied to a preprocessed Raman spectrum.

As specified in Section 2.2, the choice of the filtration is fundamental to extract features that encode the relevant information of the data shape. Such topological features are computed for each spectrum, and stored into Persistent Diagrams (PDs), resulting in eight datasets (4 types of spectra, 2 filtrations) of 43 persistence diagrams.

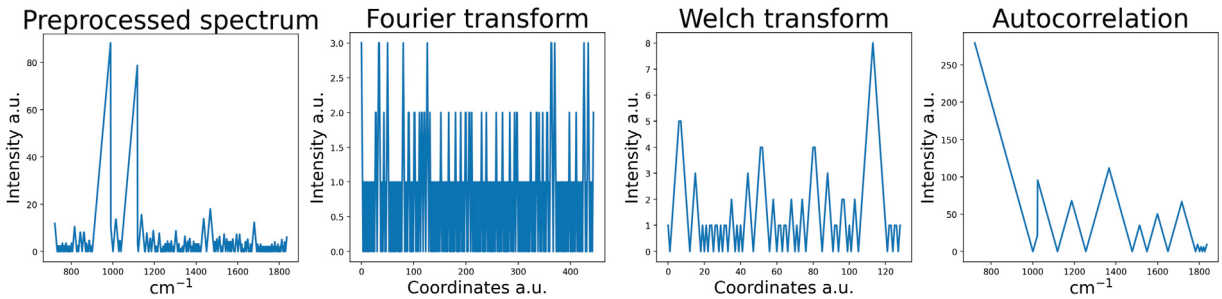

Fig. 5. Dilation filtration of a preprocessed Raman spectrum.

Table 1
Summary of the vectorization methods used in the topological machine learning pipeline, their stability, parameters and reference.

Vectorization method	Stability	Parameters	Reference
Persistence statistics	✗	✗	[42]
Entropy summary	✓	resolution $\in \{50, 100\}$	[43]
Algebraic functions	✗	✗	[44]
Tropical coordinate function	✓	resolution $\in \{50, 100\}$	[45]
Complex polynomial	✓	number of coefficients $\in \{5, 20\}$ polynomial type $\in \{R, T\}$	[46] [47]
Betti curve	✗	resolution $\in \{50, 100\}$	[48]
Lifespan curve	✗	resolution $\in \{50, 100\}$	[39]
Persistence landscapes	✓	number of landscapes $\in \{5, 10\}$ resolution $\in \{50, 100\}$	[40]
Persistence silhouette	✓	weight $\in \{1, 10\}$ resolution $\in \{50, 100\}$	[41]
Persistence image	✓	bandwidth $\in \{0.05, 1\}$ resolution $\in \{50, 100\}$	[49]
Template function	✓	$\delta \in \{5, 25\}$ $\pi \in \{1, 20\}$	[50]
Adaptive template system	✓	number of clusters $\in \{10, 25\}$	[51]
ATOL	✓	number of functions $\in \{2, 4\}$	[52]

3.4. Vectorization methods

As specified in Section 2.2, the persistence diagrams (multisets of points in the plane, with multiplicity) are not suited for entering a machine learning classifier; hence, in literature, considerable effort has been devoted to embedding PDs into a more manageable space, resulting in the definition of a plethora of such embeddings. Moreover, not all these methods are stable with respect to the input. This means that two similar persistence diagrams may yield very different vectors when transformed by certain vectorizations. We refer to Table 1 for a detailed schematic of the employed vectorization methods in our study, their stability, the parameters choice and a reference to their definition. For self-completion, we will briefly describe only three of them, which are the descriptors leading to the best performances in our study: the lifespan curve, the persistence landscape, and the persistence silhouette.

The lifespan curve [39] tracks lifespan information over the filtration, where the lifespan is the difference between death time and birth time of a topological feature. It has been interpreted as the topological persistence because it accounts for the size of topological features. The persistence landscape [40] originated from the idea of converting PDs into a function in an additive fashion. Since the resulting descriptors are functions (living in a Banach space), it is easy to apply statistical tools to it. The persistence landscape counts the number of points in the PD in the upper left quadrant of each point of the domain. The vectorization is obtained by "stacking isosceles triangles" whose bases are the intervals in the barcode. The persistence silhouette, introduced as a variant of persistence landscapes in [41], offers flexibility through the use of a trade-off parameter. These parameters enable a balance between uniformly treating all pairs in the persistence diagram and focusing solely on the most persistent pairs. When such a parameter is small, the persistence silhouette is dominated by the effect of low persistence pairs. Conversely, the persistence silhouette is dominated by the most persistent pair when it is large. We refer to Fig. 6 for a graphical example of a persistence diagram and its various vectorizations by means of the techniques listed in Table 1.

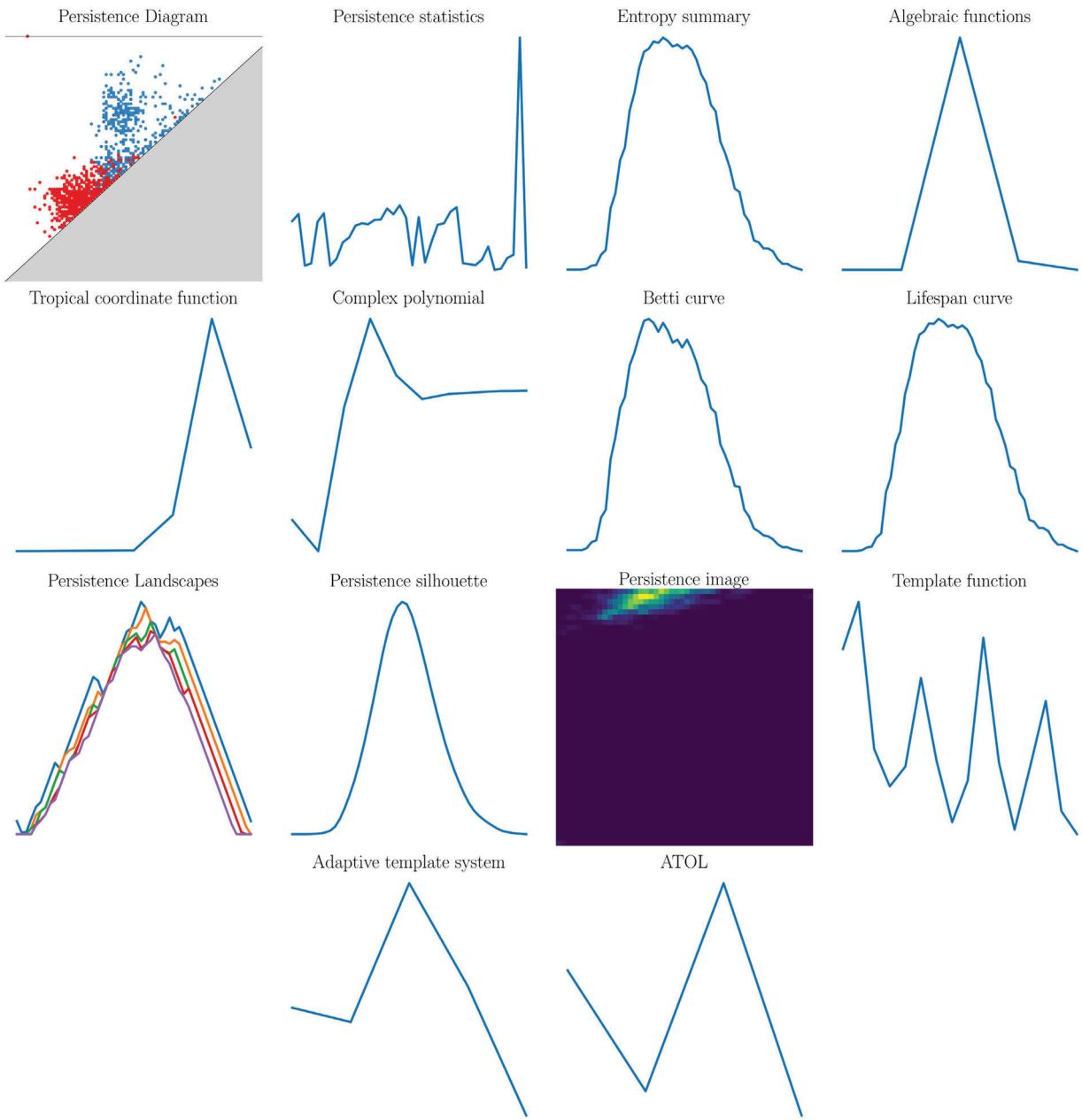

Fig. 6. Examples of a PD and its vectorizations by means of the various techniques used in the topological machine learning pipeline.

3.5. Classifiers

The classification part of the TML pipeline is made of several ML classifiers, which are all trained across all the vectorizations computed by the first part of the pipeline. This study used three different classifiers from standard machine learning literature [53]. The classifiers are the Support Vector classifier (SCV), the Random Forest classifier and the Ridge classifier. All these methods are implemented in Python via the `scikit-learn` library [54].

The training is performed using a leave-one-out cross validation scheme [55]: we train each classifier n times where n is our dataset's size. Each time, only one sample is used as a test set, while the others are used to train the classifiers. Indeed, when the dataset size is small, the leave-one-out is more appropriate as a cross validation scheme since it enables the classifier to learn better representations and returns very precise metrics.

Table 2
TML pipeline results. Accuracy results for the different vectorization methods, transformation of the Raman spectra and filtrations.

Vectorization	Lower star filtration				Dilation filtration			
	RS	Fourier	Welch	Autoc.	RS	Fourier	Welch	Autoc.
Pers. Statistics	0.53	0.58	0.67	0.58	0.60	0.58	0.72	0.58
Entropy Summary	0.58	0.74	0.74	0.60	0.58	0.74	0.44	0.60
Algebraic Functions	0.72	0.49	0.72	0.58	0.70	0.47	0.49	0.58
Tropical Coordinates	0.60	0.67	0.67	0.56	0.60	0.67	0.51	0.56
Betti Curve	0.58	0.72	0.72	0.63	0.60	0.72	0.53	0.56
Lifespan Curve	0.60	0.77	0.77	0.70	0.60	0.77	0.60	0.70
Pers. Landscapes	0.65	0.77	0.77	**0.86**	0.60	0.77	0.65	**0.86**
Pers. Silhouette	0.67	0.81	0.74	0.58	0.58	0.81	0.63	0.58
Pers. Images	0.63	0.65	0.65	0.56	0.63	0.65	0.58	0.56
Template Functions	0.70	0.63	0.70	0.67	0.67	0.63	0.58	0.67
ATS	0.58	0.67	0.67	0.58	0.53	0.70	0.51	0.63
ATOL	0.60	0.63	0.63	0.67	0.63	0.65	0.47	0.67

4. Experimental results

The experimentation aimed at three main objectives: to assess the classification accuracy of the proposed method ; to assess the relevance of the extraction of the topological features in this specific setting; and to carry out a comparison with other topology-based methods.

The dataset used for the experimentation has been acquired and preprocessed as described in Sections 3.1 and 3.2. It is made of 43 Raman spectra, 21 AD vs. 22 noAD samples; both the Raman spectra and their transformed spectra (via Fourier, Welch, and Autocorrelation transform) are considered, separately, in order to assess the impact of the 1D signal processing to the classification performance. With respect to our previous study, the dataset has been increased in size (from 24 to 43 patients), and it is more balanced (48.8% Alzheimer's disease *vs.* 51.2% other pathologies).

With reference to the topological machine learning pipeline presented in Section 3, we recall that we employed two different filtrations, four transformations, a total of 13 vectorizations of the topological descriptors, and three classifiers. We stress the fact that the validation scheme was leave-one-out cross validation, due to the relatively limited amount of data available. The accuracy values of the topological machine learning pipeline for each combination are computed as the mean test accuracy across all the 43 folds of the leave-one-out cross validation scheme, and such values are reported in Table 2. Each cell reports only the accuracy achieved by the best ML classifier.

The accuracy results achieved by both filtrations for the autocorrelation transformation of the persistence landscape vectorization are quite promising. Moreover, the consistently high accuracy for both filtrations and different vectorizations (especially persistence landscapes and silhouettes) further validates the claim that topological features offer a valid representation of the input data, which a machine learning algorithm can exploit. For the sake of completeness, the best performing method is achieved by a ridge classifier applied to the persistence landscapes vectorization (resolution: 25, number of landscapes: 5 or 10). The confusion matrix values for this method are $(19, 3, 3, 18)$ for true negative, false positive, false negative and true positive, respectively.

4.1. Data augmentation

Data augmentation is a key step for any deep learning techniques, such as CNN, which are data-hungry models, both to increase the model performances and to reduce overfitting while training and, in the end, to get a model with a better generalization capability. As it is difficult to increase the size of our dataset, we decided to rely on data augmentation. Various data augmentation techniques, standard for 1D signal analysis, have been performed as in [56–58]. Here we opted for a data augmentation similar to that proposed by Liu et al. [57]. Following a $70 - 30\%$ train-test split, small Gaussian perturbations are applied to convex linear combinations of all spectra belonging to the same class as augmented data, both in the train and in the test dataset. We highlight the fact that the test set is composed of never-seen data. Following this procedure, the resulting augmented dataset is composed of 2043 spectra, of which 1430 of training and the remaining as test. Also, the datasets obtained by applying the Fourier, the Welch and the Autocorrelation transforms are produced and included in the experimentation, as already done for the original dataset.

The TML pipeline for the augmented dataset, using the lower star filtration, achieved a classification accuracy close to those achieved for the original dataset, i.e.: Autocorrelation 79%; Welch 76%; Fourier 85%; None 69%. Complete results are reported in Table 3.

5. Discussion

In [9], the TML pipeline showed promising performances in classifying the AD sample against other pathological samples: the best accuracy value was of (87.5%). Actually, such a high value was considered very preliminary not only due to the small size of the dataset, but also due to the imbalance of the dataset. In fact, such an imbalance leads to a very high baseline accuracy of 73.3% (here, the baseline accuracy is the classification accuracy achieved by the classifier which assigns to any sample the most frequent label).

Table 3
TML pipeline results. Accuracy results for the different vectorization methods and transformation of the Raman spectra, obtained after data augmentation.

Vectorization	RS	Fourier	Welch	Autoc.
Pers. Statistics	0.61	0.79	0.40	0.62
Entropy Summary	0.67	0.78	0.64	**0.79**
Algebraic Functions	0.56	0.50	0.60	0.50
Tropical Coordinates	0.49	0.38	**0.76**	0.51
Betti Curve	0.66	**0.85**	0.29	0.77
Lifespan Curve	0.64	0.79	0.35	0.68
Pers. Landscapes	0.65	0.80	0.50	0.76
Pers. Silhouette	0.64	0.76	0.54	0.69
Pers. Images	0.55	0.79	0.51	0.50
Template Functions	0.50	0.59	0.50	0.56
ATS	0.60	0.50	0.50	0.54
ATOL	**0.69**	0.72	0.55	0.60

Table 4
Ablation study. Accuracy results using state-of-the-art methods for RS classification. It is clear that not taking into account the topological contribution greatly impacts the accuracy of the classification, resulting in a performance drop.

	RS	Fourier	Welch	Autoc.
Machine learning	0.51	0.58	0.47	**0.63**
FNN	0.49	0.42	0.37	0.51
CNN	0.47	0.51	0.44	0.53

In the present experimentation, with a baseline accuracy of about 50% and a larger and well-balanced dataset, the high classification accuracy is confirmed, with a best accuracy value of 86%. Such a value is achieved using the autocorrelation transformed spectra as input of the TML pipeline, which returns persistence landscape and a ridge classifier as the best combination of vectorization and ML classifier. Also, the choice of the Fourier transform coupled with persistence silhouettes, which in our previous work performed best, confirmed very good results achieving a classification accuracy of 81%. Notably, the stability of the proposed method has been tested with respect to the choice of two different filtrations, leading to very similar results, with the exception of the Welch transform, for which the best filtration is the lower star. Also, while data augmentation usually leads to an increase in performances and to a greater generalization capability of almost any neural network models, results reported in Section 4.1 show that the performances achieved by TML do not increase after data augmentation.

5.1. Band importance analysis

In order to identify the specific Raman spectral bands/features which are most relevant for the AD detection, we used an approach similar to RISE [59], a post hoc method used for achieving explainability in artificial intelligence. In a nutshell, given a model trained on images to perform a task, RISE identifies the regions of the image which are relevant for the model prediction and quantifies its relevance, producing an importance map.

In our approach, we investigate recursively the importance of each band of the Raman spectrum by assessing the impact of its removal in the prediction performances. In this way, the estimation of the band importance is carried out empirically by probing the model with masked versions of the input 1D signal and looking at the corresponding outputs.

We repeated our TML experiment using a set of different training datasets, which have been produced by masking each spectrum using a sliding band, varying the band width in $(10, 20, 40, 60, 100)$ and the stride in $(1, 5, 10)$.

In this way, one can appreciate and estimate, if any, the performance drop, and visually correlate such a drop with the band importance by representing in the same chart a sample Raman spectrum and the graph of the performance drop.

Fig. 7 shows that the relevance of the spectral bands with respect to AD diagnosis is more widespread than expected, and it is not due to single peaks. This would imply a reduced capability of providing clues of the role of specific chemical compounds in AD. On the other hand, such a piece of information may be used to verify if the representatives of the topological features with the longest lifespan can truly capture the most important information of the Raman spectrum.

5.2. Ablation study

In order to assess the relevance of the topological contribution to the classification of RS, we performed an ablation study: we applied both standard machine learning methods and two different neural networks to the four transformations of the Raman spectrum. In more detail, the machine learning methods are the same applied in the topological machine learning pipeline, with the difference that the input vectors are not the topological features coming from persistent homology, rather directly the Raman spectrum or its transformations. The first neural network applied is a fully connected neural network (FNN) with two hidden layers of 250 and 100 neurons, respectively, and ReLU activation function. This results in approximately 250,000 training parameters.

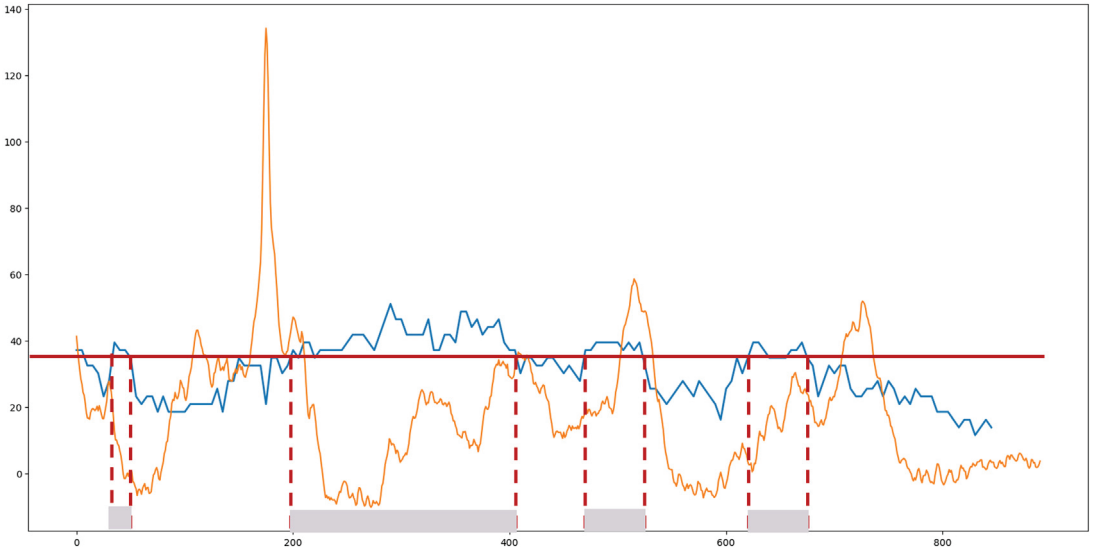

Fig. 7. The performance drop (expressed in percentage, blue), computed using a bandwidth of 40 units and a stride of 5, and one sample Raman spectrum (orange) are represented in the same graph. Using a threshold of 37% in performance drop, observe that there is no visual correlation between relevant bands (marked in grey) and the most visible peaks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Comparison results. Mean test accuracy of the leave-one-out cross validation from other classification pipelines built using PersLay [28] to combine a topological layer with one (1D or 2D) CNN and using PersLay as a feature extractor combined with ML classifiers.

	RS	Fourier	Welch	Autoc.
Pers. Image + 2D Conv.	0.49	0.49	0.49	0.49
Pers. Landscape + 1D Conv.	0.51	0.46	0.46	0.53
Pers. Entropy + 1D Conv.	0.37	0.51	0.51	**0.56**
Pers. Image + ML	0.44	0.56	0.23	0.35
Pers. Landscape + ML	0.63	0.61	**0.83**	0.53
Pers. Entropy + ML	0.65	0.58	0.35	0.56

The second neural network is a 1D convolutional neural network (CNN) with two hidden layers, the first one a convolutional layer with 8 kernels, each with kernel size of 5, and the second one a fully connected linear layer with 100 neurons. Again, the activation function is the ReLU. The CNN has approximately 100,000 parameters. Due to the limited size of the dataset, the two neural network models here reported have been trained and tested on the augmented datasets of Section 4.1. We report the accuracy results for these three competitors in Table 4. It is clear that, in our experiments, the features extracted from persistent homology are fundamental in achieving higher classification accuracy, since all methods that do not employ such features perform notably worse. Moreover, neural networks underperform, probably due to the scarcity of data.

5.3. Comparison with other methods

In order to better validate our findings, we compare the proposed method with two state-of-the-art approaches: the first is PersLay [28], while the latter is presented in [57] as a unified solution for the recognition of Raman spectrum using convolutional neural network.

PersLay is a general and versatile framework for learning vectorizations of persistence diagrams, which has been applied to graph classification with excellent results. In more detail, PersLay has been used to build another classification pipeline, able to combine vectorizations with convolutional layers (1D or 2D). Such a pipeline used as input the concatenation of the PDs of the two filtrations, for each processing of the Raman spectrum (preprocessed Raman spectrum, Fourier, Welch, and Autocorrelation). In order to better compare with our TML pipeline, only the best-performing vectorizations have been used. Moreover, PersLay has been used as a topological feature extractor, and a ML classification step (using Support Vector, Random Forest and AdaBoost classifiers, with hyper-parameter optimization) has been applied, mimicking the TML pipeline structure, following the same validation scheme (see Table 5).

In both experiments (PD vectorization + Convolutional layer and PD vectorization + ML classifier), the accuracy values are worse than those achieved by the proposed TML pipeline. Notably, the best result using PersLay is obtained by mimicking the TML

pipeline, and it achieves an accuracy of 83%, using persistence landscapes. All the other tests achieve very low accuracy, being lower than 50% for most of them.

Even using a state-of-the-art deep learning approach based on CNN, performances still remain around 50%. In more detail, the Liu's model [57], which is a model specialized for RS analysis, has been borrowed and trained in three different ways:

1. The CNN is trained from scratch using our augmented data. (Epochs: 20; Optimizer: Adam; learning rate: 3e−4). The accuracy is 0.49.
2. The CNN has been modified by reducing the size of the last fully connected layer, and then trained using our augmented data. (Epochs: 20; Optimizer: Adam; learning rate: 3e−4) The accuracy is 0.49.
3. Finally, we tested a transfer learning approach: the original CNN is trained on the data used in the reference paper [57]; then all layers are freezed but the last one and the additional fully connected one used to perform the final binary classification. The last two fully connected layers are finally trained using our augmented data. (Epochs: 20 Optimizer: Adam; learning rate 3e−4) The accuracy is 0.51.

6. Conclusions and future work

This paper introduces a TML pipeline for detecting Alzheimer's disease by analysing Raman spectra acquired from CSF samples. After reviewing recent studies on the interplay between Raman spectroscopy and machine learning and summarizing the basic ideas of TML, we have detailed the study design, including sample preparation and Raman spectra acquisitions. Next, the two main key ingredients of topological feature extraction are discussed, namely the considered filtrations and the vectorization methods, as well as the machine learning classifiers, utilized thereof for feature classification.

The results reported in Section 4 make it clear that the classification accuracy of the TML pipeline, assessed through a leave-one-out cross validation scheme, is stable and high. Also, as reported in the previous section, the high relevance of the topological information is established through an ablation study, and a comparison with another classification method using the topological features showed that the TML pipeline outperforms other methods (both using and ignoring the topological features) when applied to the analysis and classification of RS. Moreover, we want to emphasize that the only methods that achieve good results in terms of accuracy are those that exploit topological descriptors: our TML pipeline with 86% and PersLay with 83%. The importance study reported above is far from being conclusive in disclosing the explanations that can be recovered using topological methods and we hope to improve it further in the next future. Such a study would be crucial for increasing and/or confirming the knowledge about the precise molecular events and biological pathways behind the AD.

From a clinical perspective, the low occurrence of false positives (3 out of 43) and false negatives (3 out of 43) is highly promising for the development of a reliable support for AD diagnosis, deployed in a real scenario. Furthermore, since the proposed solution for CSF classification does not need the choice or set of any parameters, we are convinced that it may evolve in a dependable and automatic support in AD diagnosis, to be integrated in a commercial platform of RS.

Ethical statement

The study population is made of 43 patients, enrolled in the framework of the Bando Salute 2018 PRAMA project (Proteomics, RAdiomics & Machine learning-integrated strategy for precision medicine for Alzheime's'), co-funded by the Tuscany Region, with the approval of the Institutional Ethics Committee of the Careggi University Hospital Area Vasta Centro (ref. number 17918_bio).

CRediT authorship contribution statement

Francesco Conti: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Martina Banchelli:** Conceptualization, Resources. **Valentina Bessi:** Conceptualization, Resources. **Cristina Cecchi:** Conceptualization, Investigation, Resources. **Fabrizio Chiti:** Conceptualization, Data curation, Funding acquisition. **Sara Colantonio:** Conceptualization, Funding acquisition, Supervision. **Cristiano D'Andrea:** Conceptualization, Resources. **Marella de Angelis:** Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Davide Moroni:** Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Benedetta Nacmias:** Conceptualization, Resources. **Maria Antonietta Pascali:** Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Sandro Sorbi:** Conceptualization, Resources. **Paolo Matteini:** Conceptualization, Data curation, Funding acquisition, Methodology, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

Sara Colantonio is an author of the submitted paper and she is an Associate Editor for the Journal of the Franklin Institute and will not be involved in the editorial review or the decision to publish this article.

All the other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset used in this work is not public due to privacy restriction, as stated in the PRAMA PROJECT protocol (reference number 17918_bio) approved by the Institutional Ethics Committee of the Careggi University Hospital Area Vasta Centro. Informed consent was obtained from all subjects involved in the study.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Grammarly in order to improve the quality of English. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgements

This research was partially funded in the framework of the Bando Salute 2018 PRAMA project co-funded by the Tuscany Region, Italy. M.B., C.D., M.D.A., and P.M. acknowledge Andrea Donati for his expert technical assistance.

References

[1] Alzheimer's Disease International, McGill University, World Alzheimer Report 2021, 2021, avalable on line: https://www.alzint.org/resource/world-alzheimer-report-2021/; Last retrieved 2024.
[2] K. Eberhardt, C. Stiebing, C. Matthäus, M. Schmitt, J. Popp, Advantages and limitations of raman spectroscopy for molecular diagnostics: an update, Expert Rev. Molecular Diagnost 15 (6) (2015) 773–787, http://dx.doi.org/10.1586/14737159.2015.1036744.
[3] P. Polykretis, M. Banchelli, C. D'Andrea, M. de Angelis, P. Matteini, Raman spectroscopy techniques for the investigation and diagnosis of alzheimer's disease, FBS 14 (3) (2022) 22, http://dx.doi.org/10.31083/j.fbs1403022.
[4] Y. Xu, X. Pan, H. Li, Q. Cao, F. Xu, J. Zhang, Accuracy of raman spectroscopy in the diagnosis of alzheimer's disease, Front. Psychiatry 14 (2023) http://dx.doi.org/10.3389/fpsyt.2023.1112615.
[5] K. Blennow, H. Zetterberg, Biomarkers for alzheimer's disease: current status and prospects for the future, J. Int. Med 284 (6) (2018) 643–663, http://dx.doi.org/10.1111/joim.12816.
[6] E. Ryzhikova, N.M. Ralbovsky, V. Sikirzhytski, O. Kazakov, L. Halamkova, J. Quinn, E.A. Zimmerman, I.K. Lednev, Raman spectroscopy and machine learning for biomedical applications: Alzheimer's disease diagnosis based on the analysis of cerebrospinal fluid, Spectrochim. Acta A 248 (2021) 119188, http://dx.doi.org/10.1016/j.saa.2020.119188.
[7] C.-C. Huang, C. Isidoro, Raman spectrometric detection methods for early and non-invasive diagnosis of alzheimer's disease, J. Alzheimer's Dis 57 (2017) 1145–1156, http://dx.doi.org/10.3233/JAD-161238.
[8] F. Conti, M. D'Acunto, C. Caudai, S. Colantonio, R. Gaeta, D. Moroni, M.A. Pascali, Raman spectroscopy and topological machine learning for cancer grading, Sci. Rep. 13 (1) (2023) 7282, http://dx.doi.org/10.1038/s41598-023-34457-5.
[9] F. Conti, M. Banchelli, V. Bessi, C. Cecchi, F. Chiti, S. Colantonio, C. D'Andrea, M. de Angelis, D. Moroni, B. Nacmias, et al., Alzheimer disease detection from raman spectroscopy of the cerebrospinal fluid via topological machine learning, Eng. Proceed 51 (1) (2023) 14.
[10] A.S. Haka, K.E. Shafer-Peltier, M. Fitzmaurice, J. Crowe, R.R. Dasari, M.S. Feld, Diagnosing breast cancer by using raman spectroscopy, Proc. Natl. Acad. Sci. 102 (35) (2005) 12371–12376.
[11] J. Hutchings, C. Kendall, N. Shepherd, H. Barr, N. Stone, Evaluation of linear discriminant analysis for automated raman histological mapping of esophageal high-grade dysplasia, J. Biomed. Opt. 15 (6) (2010) 066015.
[12] Y. Oshima, H. Shinzawa, T. Takenaka, C. Furihata, H. Sato, Discrimination analysis of human lung cancer cells associated with histological type and malignancy using raman spectroscopy, J. Biomed. Opt. 15 (1) (2010) 017009–017009.
[13] M. D'Acunto, R. Gaeta, R. Capanna, A. Franchi, Contribution of raman spectroscopy to diagnosis and grading of chondrogenic tumors, Sci. Rep. 10 (1) (2020) 2155.
[14] L. Zhang, C. Li, D. Peng, X. Yi, S. He, F. Liu, X. Zheng, W.E. Huang, L. Zhao, X. Huang, Raman spectroscopy and machine learning for the classification of breast cancers, Spectrochim. Acta A: Molecular and Biomolecular Spectroscopy 264 (2022) 120300.
[15] R. Luo, J. Popp, T. Bocklitz, Deep learning for raman spectroscopy: a review, Analytica 3 (3) (2022) 287–301.
[16] H. Chen, C. Chen, H. Wang, C. Chen, Z. Guo, D. Tong, H. Li, H. Li, R. Si, H. Lai, et al., Serum raman spectroscopy combined with a multi-feature fusion convolutional neural network diagnosing thyroid dysfunction, Optik 216 (2020) 164961.
[17] J. Ding, M. Yu, L. Zhu, T. Zhang, J. Xia, G. Sun, Diverse spectral band-based deep residual network for tongue squamous cell carcinoma classification using fiber optic raman spectroscopy, Photodiagnosis Photodyn. Therapy 32 (2020) 102048.
[18] N. Blake, R. Gaifulina, L.D. Griffin, I.M. Bell, G.M. Thomas, Machine learning of raman spectroscopy data for classifying cancers: a review of the recent literature, Diagnostics 12 (6) (2022) 1491.
[19] J. Schuetzke, N.J. Szymanski, M. Reischl, Validating neural networks for spectroscopic classification on a universal synthetic dataset, npj Comput. Mater 9 (1) (2023) 100.
[20] N.M. Ralbovsky, G.S. Fitzgerald, E.C. McNay, I.K. Lednev, Towards development of a novel screening method for identifying alzheimer's disease risk: Raman spectroscopy of blood serum and machine learning, Spectrochim. Acta A: Molecular and Biomolecular Spectroscopy 254 (2021) 119603.
[21] F. Hensel, M. Moor, B. Rieck, A survey of topological machine learning methods, Frontiers Artificial Intelligence Appl. 4 (2021) 681108.
[22] G. Carlsson, M. Vejdemo-Johansson, Topological Data Analysis with Applications, Cambridge University Press, 2021.
[23] H. Edelsbrunner, A Short Course in Computational Geometry and Topology, No. Mathematical Methods, Springer, 2014.
[24] H. Edelsbrunner, J. Harer, et al., Persistent homology-a survey, Contemp. Math. 453 (26) (2008) 257–282.
[25] D. Ali, A. Asaad, M.-J. Jimenez, V. Nanda, E. Paluzo-Hidalgo, M. Soriano-Trigueros, A survey of vectorization methods in topological data analysis, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[26] G. Kusano, K. Fukumizu, Y. Hiraoka, Kernel method for persistence diagrams via kernel embedding and weight factor, J. Mach. Learn. Res. 18 (189) (2018) 1–41.
[27] F. Conti, D. Moroni, M.A. Pascali, A topological machine learning pipeline for classification, Mathematics 10 (17) (2022) http://dx.doi.org/10.3390/math10173086.
[28] M. Carrière, F. Chazal, Y. Ike, T. Lacombe, M. Royer, Y. Umeda, Perslay: A neural network layer for persistence diagrams and new graph topological signatures, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2786–2796.
[29] R.S. Tashjian, H.V. Vinters, W.H. Yong, Biobanking of cerebrospinal fluid, Biobanking: Methods Protocols (2019) 107–114.

[30] H. Vanderstichele, M. Bibl, S. Engelborghs, N. Le Bastard, P. Lewczuk, J.L. Molinuevo, L. Parnetti, A. Perret-Liaudet, L.M. Shaw, C. Teunissen, D. Wouters, K. Blennow, Standardization of preanalytical aspects of cerebrospinal fluid biomarker testing for alzheimer's disease diagnosis, Alzheimer's Dement 8 (1) (2012) 65–73, http://dx.doi.org/10.1016/j.jalz.2011.07.004.
[31] O. Ryabchykov, S. Guo, T. Bocklitz, Analyzing raman spectroscopic data, Phys. Sci. Rev 4 (2) (2019) 20170043, http://dx.doi.org/10.1515/psr-2017-0043, [cited 2024-02-13].
[32] P.H. Eilers, H.F. Boelens, Baseline correction with asymmetric least squares smoothing, Leiden Univ. Med. Centre Rep 1 (1) (2005) 5.
[33] W.H. Press, S.A. Teukolsky, Savitzky-golay smoothing filters, Comput. Phys. 4 (6) (1990) 669–672.
[34] R.J. Marks, Handbook of Fourier Analysis & Its Applications, Oxford University Press, 2009.
[35] P.K. Rahi, R. Mehra, et al., Analysis of power spectrum estimation using welch method for various window techniques, Int. J. Emerg. Technol. Eng 2 (6) (2014) 106–109.
[36] L. Rabiner, On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoust. Speech Signal Process. 25 (1) (1977) 24–33.
[37] D.A. Warde, S.M. Torres, The autocorrelation spectral density for doppler-weather-radar signal analysis, IEEE Trans. Geosci. Remote Sens. 52 (1) (2013) 508–518.
[38] A. Garin, G. Tauzin, A topological reading lesson: Classification of mnist using tda, in: 2019 18th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2019, pp. 1551–1556.
[39] Y.-M. Chung, A. Lawson, Persistence curves: A canonical framework for summarizing persistence diagrams, Adv. Comput. Math. 48 (1) (2022) 6.
[40] P. Bubenik, et al., Statistical topological data analysis using persistence landscapes., J. Mach. Learn. Res. 16 (1) (2015) 77–102.
[41] F. Chazal, B.T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, Stochastic convergence of persistence landscapes and silhouettes, in: Proceedings of the Thirtieth Annual Symposium on Computational Geometry, 2014, pp. 474–483.
[42] C.S. Pun, S.X. Lee, K. Xia, Persistent-homology-based machine learning: A survey and a comparative study, Artif. Intell. Rev. 55 (7) (2022) 5169–5213, http://dx.doi.org/10.1007/s10462-022-10146-z.
[43] N. Atienza, R. Gonzalez-Díaz, M. Soriano-Trigueros, On the stability of persistent entropy and new summary functions for topological data analysis, Pattern Recognit. 107 (2020) 107509, http://dx.doi.org/10.1016/j.patcog.2020.107509, https://www.sciencedirect.com/science/article/pii/S0031320320303125.
[44] A.B. Adcock, E. Carlsson, G.E. Carlsson, The ring of algebraic functions on persistence bar codes, in: ArXiv: Rings and Algebras, 2013, https://api.semanticscholar.org/CorpusID:2964961.
[45] S. Kališnik, Tropical coordinates on the space of persistence barcodes, Found. Comput. Math. 19 (1) (2019) 101–129.
[46] M. Ferri, C. Landi, Representing size functions by complex polynomials, Proc. Math. Met. in Pattern Recognit 9 (1999) 16–19.
[47] B. Di Fabio, M. Ferri, Comparing persistence diagrams through complex vectors, in: Image Analysis and Processing—ICIAP 2015: 18th International Conference, Genoa, Italy, September 7-11, 2015, in: Proceedings, Part I 18, Springer, 2015, pp. 294–305.
[48] Y. Umeda, Time series classification via topological data analysis, Inf. Media Technol 12 (2017) 228–239.
[49] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, L. Ziegelmeier, Persistence images: A stable vector representation of persistent homology, J. Mach. Learn. Res. 18 (2017).
[50] J.A. Perea, E. Munch, F.A. Khasawneh, Approximating continuous functions on persistence diagrams using template functions, Found. Comput. Math. 23 (4) (2023) 1215–1272.
[51] L. Polanco, J.A. Perea, Adaptive template systems: Data-driven feature selection for learning with persistence diagrams, in: 2019 18th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2019, pp. 1115–1121.
[52] M. Royer, F. Chazal, C. Levrard, Y. Umeda, Y. Ike, Atol: measure vectorization for automatic topologically-oriented learning, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1000–1008.
[53] Z.-H. Zhou, Machine Learning, Springer Nature, 2021.
[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[55] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning, Vol. 2, Springer, 2009.
[56] P. Manganelli Conforti, M. D'Acunto, P. Russo, Deep learning for chondrogenic tumor classification through wavelet transform of Raman spectra, Sensors 22 (19) (2022) http://dx.doi.org/10.3390/s22197492.
[57] J. Liu, M. Osadchy, L. Ashton, M. Foster, C.J. Solomon, S.J. Gibson, Deep convolutional neural networks for Raman spectrum recognition: a unified solution, Analyst 142 (2017) 4067–4074, http://dx.doi.org/10.1039/C7AN01371J.
[58] D. Ma, L. Shang, J. Tang, Y. Bao, J. Fu, J. Yin, Classifying breast cancer tissue by Raman spectroscopy with one-dimensional convolutional neural network, Spectrochim. Acta A 256 (2021) 119732, http://dx.doi.org/10.1016/j.saa.2021.119732.
[59] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, in: British Machine Vision Conference 2018, BMVC 2018, 2019.