scientific reports

Check for updates

OPEN

Raman spectroscopy based diagnosis of pancreatic ductal adenocarcinoma

Gianmarco Lazzini[1], Raffele Gaeta[2], Luca Emanuele Pollina[2], Annalisa Comandatore[3,4], Niccolò Furbetta[4], Luca Morelli[4] & Mario D'Acunto[1]✉

Pancreatic ductal adenocarcinoma is currently the 12th most frequent form of cancer worldwide, characterized by a very low 5-year survival rate. Although several therapeutic approaches have been proposed to treat this form of pancreatic cancer, surgical resection is still commonly recognized as the most effective technique to slow down the disease progression and maximize the 5-year survival rate. Analogously, one critical issue is the ability of current diagnostic methodologies to distinguish between irregular growth of the tumor mass and surrounding inflammatory tissues. In this pilot study, we apply Raman spectroscopy, supported by a series of machine learning techniques, to distinguish among healthy, pancreatitis and ductal adenocarcinoma tissues, respectively, for a total of 15 cases. Raman spectroscopy is a label-free, non-destructive spectral technique exploiting Raman scattering. In turn, by applying a combination of principal component analysis and random forest classifier on the Raman spectral dataset, we achieved a maximum accuracy of up to ∼96%. Our findings clearly indicate that Raman spectroscopy could become a powerful spectral technique to support pathologists in improving pancreatic cancer diagnosis.

Keywords Raman spectroscopy, Pancreatic ductal adenocarcinoma, Gaussian Naive-Bayes, Random forest classifier, SPectral SELection

Abbreviations

DA	Ductal adenocarcinoma
IFSA	Intraoperatice forzen-section analysis
IUS	Intraoperatice ultrasonography
INIR	Intraoperatice near infrared
IG	Indocyanine green
RS	Raman spectroscopy
ML	Machine learning
CRM	Confocal Raman microscopy
LDA	Linear discriminant analysis
GNB	Gaussian Naive-Bayes
RFC	Random forest classifier
N	Normal pancreas
P	Pancreatitis
PCA	Principal component analysis
H&E	Hematoxylin & Eosin
N_s	Number of Raman spectra
N_p	Number of points belonging to a single Raman spectrum
SPSEL	SPectral SELection

According to the estimations provided by GLOBOCAN 2018, pancreatic ductal adenocarcinoma (DA) turns out to be relatively uncommon in comparison to other types of tumors, being the 12th most frequent form of cancer worldwide[1]. However, this class of tumors is universally known for its low 5-year survival rate, reaching averaged

[1]CNR-IBF, Istituto di Biofisica Consiglio Nazionale delle Ricerche, via Moruzzi 1, 56124 Pisa, Italy. [2]Second Division of Surgical Pathology, University Hospital of Pisa, Pisa, Italy. [3]General Surgery Unit, Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy. [4]Department of Surgery, Amsterdam UMC, Location Vrije Universiteit, Amsterdam, The Netherlands. ✉email: mario.dacunto@ibf.cnr.it

values of 4.2% in 2011[2]. The projections in terms of the incidence and years of life lost correlated to pancreatic diseases, such as pancreatitis and DA, are set to increase in the next years unless adequate strategies for screening and early-stage therapies are introduced[3].

In the last years several therapeutic approaches, including surgery[4], radiotherapy[5], chemotherapy[6], targeted therapy[7], immunotherapy[8], microbial therapy[9] etc. have been experimented to treat DA. However, up to now, surgical resection is still considered the most effective technique to significantly slow down the disease progression and to improve the patient's quality of life[10]. In particular, several studies[11–13], highlighted how the achievement of a complete tumor removal leads to a significant increase in the 5-year survival rate. This objective is made challenging by the frequent presence of metastases and/or small portions, attributable to the irregular growth of the tumor mass and difficult to distinguish from the surrounding inflammatory tissues[14]. For these reasons, scientists devoted their efforts to developing fast and powerful adjuvant techniques capable of helping the surgeon to distinguish the malignant and the normal tissues during the tumor resection. Up to now, intraoperative frozen-section analysis (IFSA) is universally considered the most important method for intraoperative pancreatic tumor margin assessment[15]. In this procedure, the excised tumor mass is frozen and its section is subjected to a histopathological analysis, to assess the possible presence of tumor cells. IFSA is a well-established procedure in surgical pathology; however, it is subject to numerous challenges and issues, including the duration of the sequence of steps between tumor excision and the final histopathological response. Moreover, the quality of frozen tissue sections is considerably inferior to that of sections from formalin-fixed and paraffin-embedded specimens. Additionally, frozen tissue may undergo artifact-induced changes that significantly complicate the intraoperative diagnosis. In turn, the absence of standardized protocols for the histopathological analysis of the frozen sections leads to low levels of sensitivity, reaching only $\sim 33\%$[16]. Another low-cost and safe investigated approach is represented by intraoperative ultrasonography (IUS), which distinguishes the tumor mass based on the reflectance of ultrasounds. Despite its adaptability either for open surgery or laparoscopy, IUS requires training and expertise for the interpretation of the resulting images. In addition, the reduced field of view hinders the capability of detecting masses within deep layers of tissue[17]. This is not the case with intraoperative near infrared (INIR) imaging which has emerged as a promising approach for intraoperative tumor margin assessment. This technique exploits hypothetical differences in the fluorescence emission of the tumor with respect to the surrounding tissues[14]. A major drawback of such an approach regards the need for fluorescent probes, such as indocyanine green (IG), necessary to enhance the optical contrast. Such molecules are often characterized by insufficient quantum yield, photobleaching, high affinity with plasma proteins, and a tendency to form aggregates in aqueous solution[18].

In the last decade, Raman spectroscopy (RS) has been evidenced to be a promising route for cancer diagnosis, with particular reference to the intraoperative tumor margin detection[19–21]. RS is based on the measurement of the so-called Raman effect[22], i. e. the inelastic light scattering produced by a molecular group. This phenomenon induces a change in the wavelength of the scattered photons, strongly dependent on the chemical properties of the irradiated molecular group. For this reason, the Raman light scattered by a complex sample represents a sort of "fingerprint" of the molecular composition of the sample itself. This feature makes RS particularly suitable for cancer diagnosis, especially for *in vivo* applications, since it doesn't require time-consuming pre-treatment operations, such as freezing and/or staining. In addition, the large amount of molecular information obtainable through RS could potentially allow to identify tumor non-uniformities within the same tumor mass, corresponding to different progression stages. This perspective could have important consequences in terms of the scheduling of personalized therapeutic plans. On the other hand, the high molecular sensitivity characterizing RS often makes the qualitative interpretation of Raman-based experimental data challenging. In particular, the Raman analysis of biological tissues leads to frequent cases of strong similarities between Raman spectra of tissues corresponding to different diseases of the same type of organ. On the other hand, Raman spectra collected at different portions of the same mass of tissue may show completely different properties. A possible solution to overcome these limits is represented by the adoption of machine learning (ML) algorithms for the manipulation and interpretation of Raman-based data, to bring out the desired information. This route was investigated in previous works[23–28], putting the spotlight on the ability of RS to detect cancer. In the case of intraoperative cancer detection, ML offers the additional advantage of not requiring expert people to interpret the experimental signal, with important consequences in terms of user-friendliness and reduced diagnostic time. Today, together with the problem of maximizing diagnostic accuracy, the low diagnostic speed represents the most relevant obstacle towards the implementation of RS and ML in engineered devices for the *in situ* cancer detection. In this sense, the data acquisition on a relatively wide spectra range represents a limiting factor, since it imposes stringent choices regarding optical apparatus, ML classifier, and, consequently, the computing hardware.

Several works in literature have already proven the effectiveness of RS and ML in diagnosing pancreatic cancer. One of the first investigations in this sense was conducted by Pandya et al.[29], which employed RS with near-infrared radiation and conventional ML models (Principal Component Analysis and Discriminant Analysis) to distinguish ductal adenocarcinoma from COLO 357 and L3.6pl cells grown on murine models. The results, revealing averaged sensitivities of $\sim$93% and averaged specificities of $\sim$92%, opened the route for further investigations in this sense.

Li et al.[30] developed convolutional neural network-based tools for the intraoperative Raman analysis of fresh tissues grown in mice to differentiate between CFPAC-1 cell line ductal adenocarcinoma and normal pancreas. This research demonstrated excellent classification accuracies, especially when such classifiers were fed with 2D maps obtained from the first principal component derived from the dot product between a single spectrum and its transpose, with the values exceeding 97%. However, the use of 2D maps imposes constraints in terms of the choice of the hardware for the accomplishment of such calculations, due to the increased dataset dimensionality. This issue could represent a potential limit in view of future applications in real contexts.

RS and ML in pancreatic cancer research showed encouraging performances also for more refined diagnostic tasks, e.g. the problem of distinguishing tumor repopulating cells and parental control cells. In this sense Mandrell et al.[31] adopted Surface-Enhanced RS combined with classical ML algorithms to discriminate between MIA PaCa-2 human pancreatic cancer cell cultures corresponding to tumor-repopulating cells and parental control cells. This study showed good results, with classification accuracies of 98%, obtained with k-Nearest Neighbor and Support Vector Machine classifiers. The investigation is undoubtedly of relevant importance since it demonstrates the ability to distinguish between two classes that, in principle, are characterized by strong chemical affinities. The major drawbacks are represented by the small size of the dataset and the large dispersion associated with the classification accuracies, requiring further investigation to clarify the real potential of the approach.

The studies mentioned above all respond to the need to diagnose pancreatic cancer from tumor cells or tissues. However, another interesting route is represented by the non-invasive cancer diagnosis from blood-seum-derived Extracellular Vesicles. Uthamacumaran et al.[32] Conducted an exploratory study aimed at diagnosing several forms of tumors, including pancreatic cancer, from isolated extracellular vesicles. To this aim, Raman spectroscopy, probed with visible light, was coupled with several machine learning models, i.e. support vector machine and decision tree-based algorithms. The results are extremely encouraging, with classification accuracies larger than 90%. Despite the enormous potential of this research line in view of the realization of new and fast screening protocols, further efforts have to be devoted to finding easy-in-use and fast procedures to effectively isolate extracellular vesicles from blood.

In this paper, we focused our attention on the problem of diagnosing DA from human pancreatic tissues through an approach based on coupling Confocal Raman microscopy (CRM) to ML. At the same time, we analyzed the issue to find strategies to minimize the computational cost and, consequently, the computational time. Specifically, we verified three ML models, i.e. Linear discriminant analysis (LDA), Gaussian Naive-Bayes (GNB), and random forest classifier (RFC). In addition, intending to increase the diagnostic speed, we exploited two procedures aimed at reducing the dimensionality of the dataset employed to feed the ML models: a method based on principal component analysis (PCA), and a conceptually simpler approach, based on the extraction of spectral components from the related spectra. In the following, we will refer to this technique as SPectral SELection (SPSEL). The approach allowed us to distinguish normal pancreas (N), pancreatitis (P), and DA with a maximum accuracy of $\sim 96\%$ for PCA+RFC, demonstrating its potential in view of future implementations in devices for the *in vivo* diagnosis. On the other hand, SPSEL allowed the detection of specific sub-bands, corresponding to molecules recognized as biomarkers for cancer detection. Among these characteristic molecules, we mention methionine, i.e. an indicator of DNA hypomethylation, a frequent phenomenon that accompanies many types of cancer.

Methods

Samples

The study was conducted on surgical specimens obtained from the General Surgery Department and analyzed at the Unit of Pathology 2 of the Azienda Ospedaliero-Universitaria Pisana (Pisa, Italy). In Table 1 we resumed the general characteristics of the patients involved in the study, i.e. sex, age, and histotype. The specimen, once excised, was fixed in 4% buffered formalin and sampled according to routine protocols for histopathological examination. Subsequently, the tissue was processed and embedded in paraffin (formalin-fixed paraffin-embedded, FFPE). Although this fixation procedure likely influences the tissue chemistry, using such samples offers the advantage of employing retrospective samples, with the possibility of maximizing the size of the statistical ensemble available, with important consequences in terms of diagnostic reliability. This issue is particularly significant for tumors, such as pancreatic cancer, which are relatively rare among the population. The treated tissues were then

Patient index	Sex	Age	Histotype (s)
1	M	74	Normal
2	M	79	Normal, chronic pancreatitis
3	M	90	Normal
4	F	75	Normal, chronic pancreatitis
5	F	74	Chronic pancreatitis
6	M	76	Ductal adenocarcinoma (Grade II)
7	F	74	Ductal adenocarcinoma (Grade II)
8	F	75	Ductal adenocarcinoma (Grade II)
9	F	77	Ductal adenocarcinoma (Grade II)
10	F	79	Ductal adenocarcinoma (Grade II)
11	M	56	Ductal adenocarcinoma (Grade II)
12	M	64	Ductal adenocarcinoma (Grade III)
13	M	58	Ductal adenocarcinoma (Grade III)
14	M	79	Ductal adenocarcinoma (Grade III)
15	F	61	Ductal adenocarcinoma (Grade III)

Table 1. Table resuming the general characteristics of the patients' cohort.

sectioned to obtain two slices of 5 μm and 20 μm thickness, respectively, for each tissue sample. We deposited the aforementioned tissue sections on conventional glass slides employed in optical microscopy. In the framework of RS, despite in some cases such type of substrate could be a source of an undesired intense backscattering contribution to the measured signal, it offers the advantage of being cost-effective and easy to implement, making the whole approach more promising in view of the implementation in engineered devices for common use. However, for our purposes, we chose samples resulting in a negligible contribution of the substrate to the Raman signal, occurring as a characteristic wide and triangular peak around $\sim 1000\,\mathrm{cm}^{-1}$. After placing the sections on the glass slides, they were washed in two xylene baths for 10 minutes to remove residual formalin. The 5 μm section was subsequently subjected to a staining protocol using Hematoxylin and Eosin (H&E) and covered with a coverslip, as typically used in optical microscopy. Two pathologists utilized this stained section for histopathological analysis to determine the corresponding histotype. In Fig. 1 we reported illustrative optical images of the aforementioned stained histologic sections.

As we will explain further, this procedure enabled us to assign a "label" to each Raman spectrum, allowing for supervised ML. H&E dramatically promote light absorption. This effect could lead to an intense fluorescence signal, tending to cover the Raman component of interest. Furthermore, light absorption is generally followed by non-radiative relaxation phenomena, determining sample heating with consequent tissue degradation. To suppress such phenomena, we acquired the Raman signal from the 20 μm unstained tissues section. Particular care was taken to avoid depleting or compromising the material for potential future investigations.

In total, the study involved 15 biopsies, 2 of Normal pancreas (N), 2 containing either the histotypes N or chronic pancreatitis (P), 1 with P, 6 with ductal adenocarcinoma (DA) of grade II, and 4 with DA of grade III. As we will see later, we focused our attention on the problem of distinguishing N, P, and DA regardless of the grade of DA. Therefore, we grouped all the DA cases in a unique class. In Fig. 1 we reported examples of white light optical images, obtained in reflection mode, of the aforementioned stained sections.

Raman spectroscopy

To collect the Raman-based data, we adopted an Horiba™ Xplora Plus Confocal Raman Microscope, featuring a 532 nm diode laser. For all the measurements, we set the laser power to 8 mW. The optical system of the microscope included: a $50\times$ objective of numerical aperture 0.80, resulting in a nominal laser spot of diameter of approximately $1\,\mu\mathrm{m}$; an aperture (pinhole) of diameter $100\,\mu\mathrm{m}$ leading to an axial resolution of $\sim 2\,\mu\mathrm{m}$; a 1200 gr/mm diffraction grating, leading to Raman spectra with a resolution of $\sim 2\ \mathrm{cm}^{-1}$. To suppress the signal

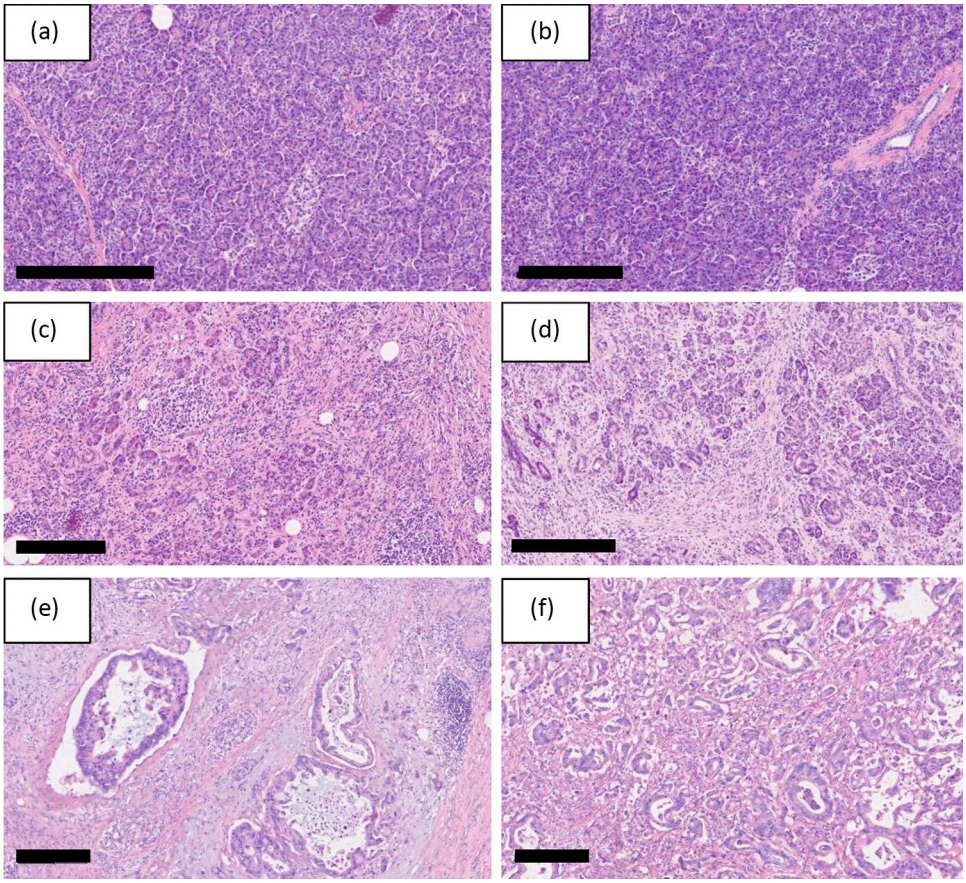

Fig. 1. Optical images of H&E stained tissue sections corresponding to pancreatic ducts. (**a**) and (**b**): normal pancreas (N); (**c**) and (**d**): pancreatitis (P); (**e**) and (**f**): ductal adenocarcinoma (DA). marker size: 300 μm.

noise and, at the same time, to preserve the physicochemical integrity of the samples, we collected the Raman spectra with an acquisition time of 0.2 s and 35 accumulations for each spectrum. Despite such tissues showing an intense Raman signal between 2800 and 3400 cm^{-1}, the high dispersion associated with such Raman band pushed us to restrict the spectral interval to 400 and 1800 cm^{-1} as the most relevant for our purposes. For our analysis, we focused our attention on the portions of tissue located around the pancreatic ducts by selecting areas, or stripes, of width $\sim 100\,\mu\mathrm{m}$ adjacent to the boundary of the ducts (see Fig. 2).

By qualitatively observing the laser light reflected by several portions of the samples' surface, we noted a highly deformed spot of effective diameter $\sim 60\,\mu\mathrm{m}$. This feature is likely attributable to the high level of roughness of the tissue surface, leading to multiple light reflections. To avoid superposition between adjacent laser spots, we collected the Raman spectra in points distributed in square grids of pixel resolution determined by the aforementioned "effective" laser spot diameter.

Data pre-processing

As we will clarify in the following, we dedicated our attention to a supervised learning problem aimed at distinguishing three classes: N, P, and DA. To this aim, we built the dataset adopted for the following ML analysis by collecting a number of Raman spectra between 150 and 170 for each sample, depending on the size of the available tissue. Since the number of samples for each class was not the same, possible artifacts related to the class imbalance could affect the diagnostic accuracy. For this reason, we applied a Synthetic Minority Oversampling Technique (SMOTE), to balance the number of spectra belonging to each class[33]. The result of this operation was a dataset of $N_s = 3420$ spectra, 1140 per class. To suppress the contribution of sample fluorescence, we applied an Improved Multi-Polynomial Fitting algorithm (polynomial order: 3, number of iterations: 200) for the baseline subtraction[34]. To remove the high-frequency noise, we applied a Savitzky–Gloay filter (window:17 points; polynomial order: 3). Finally, we subtracted the minimum signal from each spectrum and normalized the result by dividing it by the numerical integral, calculated through a trapezoidal rule. This last operation allowed to minimize the signal dispersion attributable to topographical inhomogeneities (cracks and/or voids), due to the sample preparation. We used Python3.10 scripts to perform all the pre-processing operations.

Machine learning

A single numerical Raman spectrum can be represented as a vector of components, or features, $\mathrm{x} \equiv \{x_i\}$, with $i = 1, \ldots, N_p$. As anticipated in the previous subsections, the ML algorithms tested in this work belong to the so-called supervised learning. Therefore, the Raman spectrum x^j, where $j = 1, \ldots, N_s$, can be associated with a label y_j assuming one of the values N, P, or DA. In this investigation, we assessed the classification performances of the following algorithms:

- *Linear discriminant analysis (LDA)* this algorithm is one of the most popular supervised classifiers[35]. LDA is based on the resolution of the so-called Fisher's criterion, based on the simultaneous maximization of the inter-classes variance and minimization of the intra-classes variance. This problem corresponds to the determination of the eigenvectors of the matrix $W^{-1}B$, where W represents the pooled within-group covariance matrix and B is the between-group covariance matrix[36];
- *Gaussian Naive-Bayes (GNB)* this classifier belongs to a category of ML models based on Bayes' theorem[37], with the assumptions of normal and conditionally independent features x_i. GNB reduces to the search for the

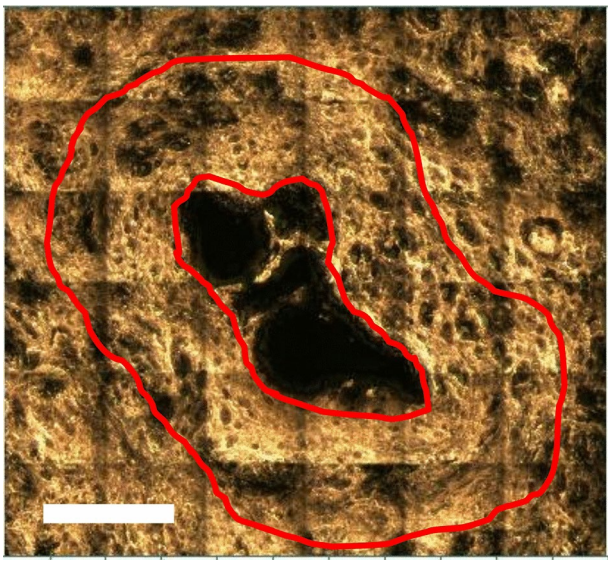

Fig. 2. Reflectance optical image of an unstained tissue pancreatic section, showing a pancreatic duct. We obtained this image by stitching multiple images, obtained with a magnification of 50×. We enclosed within a red line the portion of tissue of interest for Raman acquisition. Marker size: 100 µm.

value of the label variable *y* maximizing the quantity $P(y)\prod_{i=1}^{N_p} P(x_i|y)$, where *P*(*y*) represents the *a priori* probability associated to the aleatory variable *y* and $P(x_i|y)$ is the probability of the event x_i conditioned to the event *y*;

- *Random forest classifier (RFC)* RFC belongs to the family of classifiers named "bagging". In this case, a trained classifier can be viewed as an ensemble, or forest, of weak classifiers, or decisional trees. Each tree is grown from a dataset obtained by randomly extracting N_s from the N_s available spectra, without the constraint of the absence of repetitions. A single node of the tree is then built by randomly selecting $N_p^{\frac{1}{2}}$ features from the N_p available. Then, the "best" feature among the $N_p^{\frac{1}{2}}$ selected is chosen on the basis of the ability to separate the spectra corresponding to the different values of *y*. In this work, we quantified this ability in terms of the Gini's index[38]. In addition, each tree was grown up to get a training accuracy of 100%. This operation was repeated to obtain a forest of 200 trees, chosen as a good compromise between the need to maximize the classification performance with small overfitting and to minimize the out-of-bag error[38]. The prediction of the whole forest corresponds to the majority of the predictions of the trees belonging to it.

We assessed the classification performances with a 5-fold cross-validation, in terms of accuracy, recall, and precision[39]. To quantify the global classification performances, we averaged these parameters over the values of the label *y* and over the folds. In the following, we will refer to the resulting averaged quantities as *A*, *R*, and *PR*, respectively.

We carried out all the ML procedures with Python3.10 scripts, through an Intel Xeon CPU with 2 virtual CPUs and 13GB of RAM.

Dimensionality reduction

As explained in the Introduction section, one of the most challenging issues towards the realization of engineered probes for the *in situ* cancer detection is represented by the time needed to express a diagnosis. In principle, this parameter is influenced by several factors. Among them, the amount of data employed to feed the ML models is supposed to play a major role. In reference to the notation adopted in the previous subsection, the "size" of the dataset found at the input of the ML routine can be quantified in terms of the number N_s of Raman spectra and of the number N_p of points belonging to a single spectrum. In the specific case of N_p, in this work, we tested two strategies for the minimization of this quantity:

- *Principal component analysis (PCA)* This widespread technique aims to find the coefficients of a linear combination of the features $\{x_i\}$, maximizing the variance associated with the dataset[40]. This problem is equivalent to the search for the eigenvectors of the covariance matrix Σ. The related eigenvalues represent the variance associated with the respective eigenvectors. From a geometrical point of view, this algorithm consists of projecting the vectors $\{\mathrm{x}^j\}$ with $j = 1, ..., N_s$ corresponding to the Raman spectra, on the directions represented by the eigenvectors of Σ. The basic idea is to neglect some of the resulting coordinates on the basis of the variance associated with the relative eigenvectors;
- We compared PCA to a conceptually simpler approach, inspired by the work of Araujo et al.[41] and based on selecting sub-bands from the spectral range. In the following, this technique will be referred to as SPectral SELection (SPSEL). In SPSEL, the whole spectral range was divided into five sub-intervals or bins. Then, we tested the ML models mentioned above by selecting the aforementioned bins according to all the possible combinations, including the case where we included all the spectral range in the input. We labeled all the 31 possible combinations with an index $n = 0, ..., 30$. We decided to divide the spectral interval into five bins according to the following criteria: a too-large number of bins may lead to a large number of combinations of bins associated with comparable classification accuracies. In principle, this issue could complicate the individuation of the combinations of bins leading to the best performances. On the other hand, choosing a too-small number of bins undoubtedly decreases the resolution related to the search of the aforementioned bands.

Results

Qualitative interpretation of the Raman spectra

In Fig. 3 we present the averaged Raman spectra associated with normal pancreas (N), pancreatitis (P), and ductal adenocarcinoma (DA). Table 2 provides the wavenumbers corresponding to the most relevant Raman peaks and their molecular interpretation. For a correct classification of the aforementioned classes, a key observation is the significant overlap of the averaged spectra occurring at values of the wavenumber larger than $\sim 1250\ \mathrm{cm}^{-1}$, which makes it challenging to distinguish between the classes in this region. On the contrary, under this wavenumber threshold, there are more pronunced differences in the intensity of the Raman signal associated with the classes of interest, especially between DA and the benign pancreatic conditions (N and P). This observation suggests that lower wavenumber regions may hold more discriminative power for differentiating DA from the other two conditions, warranting further investigation to identify characteristic Raman features useful for accurate classification. An increased Raman signal of DA in comparison to N and P has been found for the peaks at 424 (symmetric stretching vibration of ν_2 PO_4^{3-}), 476 (polysaccharides, amylose, amylopectin), 953 (choline) and 1238 cm^{-1} (amide III). A possible interpretation for these spectral trends can be associated to the traditional picture of cancer proliferation, characterized by the massive production of cellular material. Analogously, the presence of an increased Raman signal at 953 cm^{-1} for DA turns out to be in line with the idea of an increased content in the choline metabolite, characterizing several types of cancer[42]. In turn, DA showed a decreased Raman signal at 630 (glycerol or cysteine), 695 (ν (S–S) *trans* (methionine)), 760 and 803 cm^{-1} (tryptophan). The attribution of the peak at 695 cm^{-1} to methionine seems to be in line with a correlation between low levels of methionine and DNA hypomethylation, a frequent phenomenon in cancerous tissues[43].

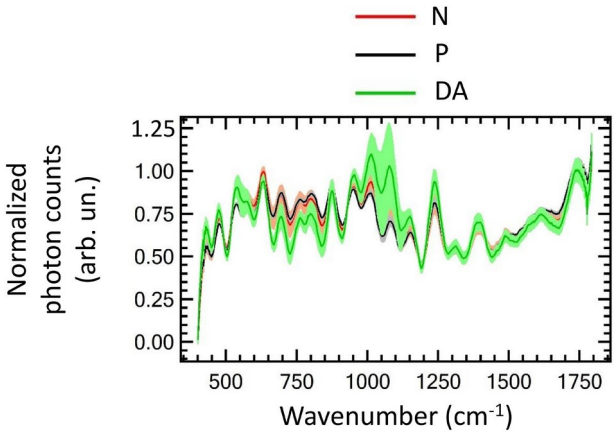

Fig. 3. Averaged Raman spectra associated with normal pancreas (N), pancreatitis (P), and ductal adenocarcinoma (DA). The shaded areas correspond to the standard deviation.

Wavenumber (cm^{-1})	Raman band assignation and references
424	Symmetric stretching vibration of ν_2 PO_4^{3-} [44]
476	Polysaccharides, amylose, amylopectin
540	PO_4^{3-} out-of-plane bending[45]
574	Tryptophan, cytosine, guanine[44]
630	Glycerol[44,46], cysteine[47]
695	ν S–S *trans* (methionine)[48]
760	Tryptophan (ring breathing)[49,50]
803	Tryptophan[51]
875	Choline[44]
953	Choline[52]
1012	Tryptophan[53], C–O stretching (ribose)[54]
1078	Phospholipids[51,55]
1150	P–O (nucleic acids, carbohydrates)[56]
1238	Amide III[56]
1312	CH_3CH_2 twisting mode of collagen/lipids[44]
1392	C–N stretching[44]
1487	RNA[57], collagen[44]
1617	ν(C=C) (tryptophan)[44]
1741	C=O stretching (lipids)[58,59]

Table 2. List showing the wavenumbers associated with the Raman peaks observed in the averaged spectra, their assignation and correspondent references.

The difference observed between DA and the other benign classes (N and P) applies also to the signal dispersion, especially in the Raman bands between 980 and 1180 cm^{-1}. This spectral range includes peaks attributable to lipids, nucleic acids, and carbohydrates (see Table 2), presumably belonging to the cells situated at the boundary of the pancreatic ducts. As shown in Fig. 1, DA tissues are characterized by a ore pronounced roughness around pancreatic ducts. This roughness, if compared with the rest of DA tissues, is a potential source of molecular inhomogeneities and consequent large signal dispersion.

Classification performances with PCA

Figure 4 shows the performances of the LDA (red squares), GNB (black circles), and RFC (green triangles), trained on a dataset obtained by applying PCA to the pre-processed Raman spectra. In particular, we reported the three parameters *A* (Fig. 4a), *R* (Fig. 4b), and *PR* (Fig. 4c) as a function of the number of principal components N_{PCs} employed to feed the ML models, within the interval $1 < N_{PCs} < 100$. In turn, the duration of the whole ML routine (training and test phase) (Fig. 4d) was accounted to provide information about the computational time. The results highlight that GNB led to the worst classification performances, with a classification accuracy reaching a maximum value of $(78.7 \pm 1.9)\%$ for $N_{PCs} = 5$. A possible reason for this outcome can be found in the invalidity of the implicit assumption of normally distributed features. Furthermore, the overfitting associated with GNB is sensitively more important than for LDA and RFC, as suggested by the width of the

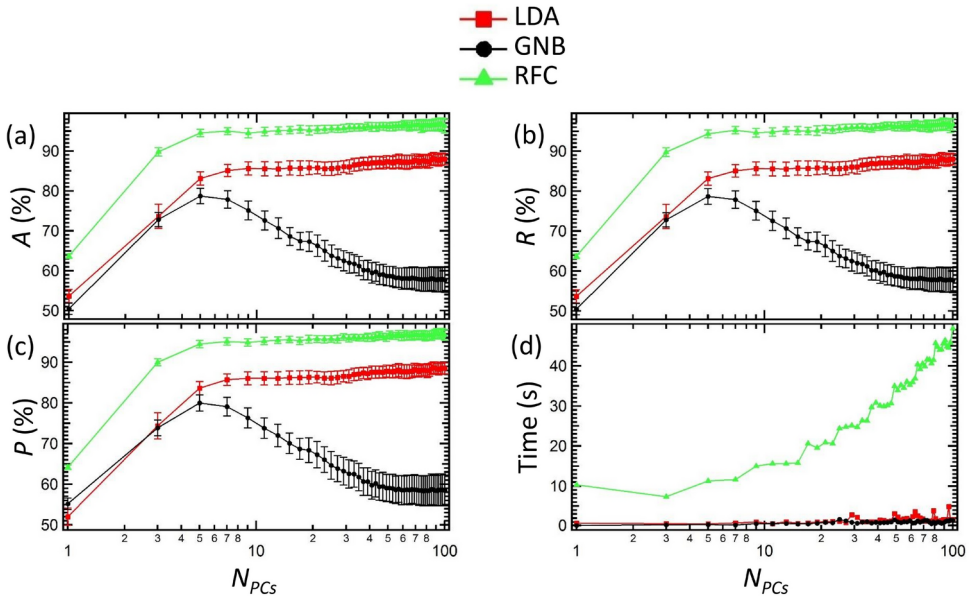

Fig. 4. Performances of LDA (red squares), GNB (black circles), and RFC (green triangles) as a function of the number of principal components N_{PCs} employed as an input of the ML models considered. All the quantities were averaged over the values of the label and over the folds of a 5-fold cross-validation. (**a**): accuracy *A*; (**b**): recall *R*; (**c**): precision *PR*; (**d**) time needed for the training and the test phase. Error bars: standard deviations.

error bars associated with *A*, *R*, and *PR*. We believe that this feature can be directly attributed to the implicit assumption of GNB of statistically independent features. On the other hand, features corresponding to adjacent spectral components can be considered strongly correlated to each other, due to the continuous character of the Raman spectra. Such correlation is particularly evident in Raman spectra of biological samples, characterized by a relatively small peak aspect ratio. On the other hand, RFC showed the best performances: compatibly with the error bars, the maximum classification accuracy reached the maximum value of $(95.0 \pm 1.8)\%$ for $N_{PCs} = 7$. It should be emphasized that the small width of the error bars of RFC in comparison to GNB suggests a smaller overfitting, resulting in increased reliability. Finally, LDA led to values of the classification performances intermediate to RFC and GNB. In particular, for $N_{PCs} = 7$, i.e. the configuration leading to the best performances for RFC, LDA reached an accuracy of $(85.1 \pm 3.0)\%$.

In terms of classification performances, the results presented above substantially highlighted that RFC and LDA represent the most accurate and reliable classifiers if compared to GNB. The difference in the classification accuracy between RFC and LDA can be likely ascribed to the higher complexity of RFC. However, a difference between RFC and LDA is evidenced by the computational time, as shown in Fig. 4d. For instance, for $N_{PCs} = 7$, i.e. the condition leading to the maximum value of *A* for RFC, the overall computation needed of a time of ~ 12.0 s for RFC and of ~ 0.3 s for LDA, with a difference of about one order of magnitude. This marked difference is due to the implicit complexity of RFC compared to LDA. this feature potentially influences the choice of the most promising ML model for future technological implementations in engineered devices.

In Fig. 5a–e we reported the absolute value of the loadings associated with the first five PCs[40], as a function of the wavenumber. The shape of loadings associated with the first PC, which is supposed to carry the dominant contribution to the total variance, appears not to recall the shape of the Raman signal. This outcome is probably attributable to a residual autofluorescence contribution, which however didn't affect significantly the classification performances. On the other hand, the other loadings, especially the loadings associated with the second and the fourth PCs show peaks of width compatible with Raman peaks. As evident by the study of the scree plot (see Fig. 5f), the first five PCs roughly contain the dominant contribution to the system variance.

Classification performances with SPSEL

In analogy with PCA, we assessed the performances of the two ML models, previously described, by adopting the technique SPSEL for dimensionality reduction. First, the procedure is applied on the whole spectral range acquired, by employing the combinations of sub-intervals, or bins, resumed in the binary diagram reported in Fig. 6a. As explained in the Methods section, on the horizontal axis of Fig. 6a we reported an index *n*, with $n = 0, ..., 30$, labeling each combination, while on the vertical axis, we reported the wavenumber. The bins adopted as input of the ML models are indicated in red whereas in black the excluded ones. In Fig. 6b we reported the number N_p of points belonging to a single spectrum as a function of *n*. Finally, the resulting behavior of *A*, *R*, *PR*, and computational time related to the training and test phase as a function of *n* are shown in Fig. 6c–f, respectively.

Figure 6 confirmed some of the features observed with PCA: RFC maximized the classification performances, allowing to obtain a maximum accuracy of ∼96% for $n = 17$, while GNB represented the worst classifier, either in terms of the averaged values of the classification parameters or in terms of the corresponding standard

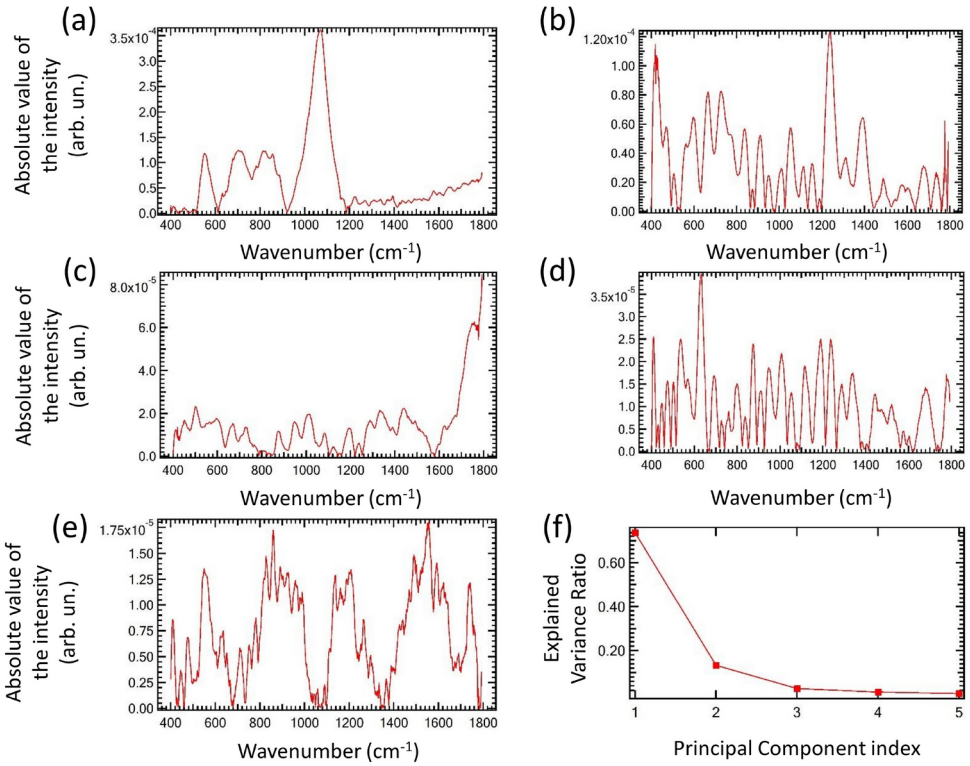

Fig. 5. Absolute value of the loadings associated with the first five PCs as a function of the wavenumber: (**a**) 1st PC; (**b**) 2nd PC; (**c**) 3rd PC; (**d**) 4th PC; (**e**) 5th PC. (**f**): Scree plot associated with the first five PCs.

deviation. In addition, the computational time needed to accomplish the training and the test phases of ML showed significantly larger values for RFC in comparison to LDA and GNB. However, by carefully analyzing the observed trends, it was possible to retrieve important information about the presence of specific biomarkers for cancer diagnosis. To this aim, since RFC showed almost no dependence on *n* and GNB was characterized by high dispersion of the classification performances, we focused our attention on the behavior of LDA for increasing values of *n*. In particular, for $0 \leq n < 4$, i. e. the combinations corresponding to the choice of a single bin as the input of the ML models, *A*, *R*, and *PR* showed a decrease, coinciding with a displacement of the spectral components employed as input of ML towards large wavenumbers. This outcome was an indication that the most promising biomarkers for the differentiation correspond to spectral components located at small wavenumbers. This trend repeated for $5 \leq n < 14$, i.e. for the combinations of two bins, and for $15 \leq n < 24$, i.e. for the combinations of three bins, with a constant and continuous decrease of the classification performances. This last behavior can be attributed to the fact that, by definition, the larger N_p the greater the similarities between the combinations of bins corresponding to the same value of N_p. Given such results, $n = 15$ can be identified as the most promising combination of bins for searching the spectral interval maximizing the classification performances, and, in the same time, minimizing N_p. This threshold, corresponding to the wavenumbers between 400 and 1230 cm^{-1}, supported the indications provided by the loading associated with the first principal component (Fig. 5a), according to which this spectral interval carried out the majority of the information.

In the attempt to find cancer biomarkers at a higher spectral resolution, we repeated the aforementioned SPSEL procedure to the spectral range between 400 and 1230 cm^{-1}, corresponding to the combination $n = 15$ specified above. In Fig. 7 we reported the result of this calculation. In this case, in comparison to SPSEL performed on the whole spectral range, the behavior of *A* (Fig. 7c), *R* (Fig. 7d), and *PR* (Fig. 7e) for RFC and GNB as a function of the combination index *m*, revealed a similar trend. Furthermore, the oscillatory behavior of the classification performances for increasing values of the combination index observed for LDA in the whole spectra appeared to be more marked, assuming a sawteeth-like shape. In particular, by comparing the values of *A*, *R*, and *PR* at the local maxima of LDA, we identified $m = 5$ as the combination of bins that, within the error bars, results in maximized classification performances and minimized N_p. This choice corresponds to the wavenumbers between 400 and 707 cm^{-1}. In reference to the interpretation of the Raman peaks given in Table 2, this spectral interval included peaks attributable to chemical compounds, i.e. phosphates, polysaccharides, tryptophan, DNA bases, etc., whose interpretation is related to the consolidated experience that associates the progression of a tumor malignancy with uncontrolled cellular proliferation. On the other hand, the peak observed at 695 cm^{-1}, attributed to methionine, could find an interpretation of the phenomenon of DNA hypomethylation common in several types of tumors, including pancreatic cancer. This last interpretation seems to be in accordance with the decreased Raman signal of DA at 695 cm^{-1} in comparison to N and P.

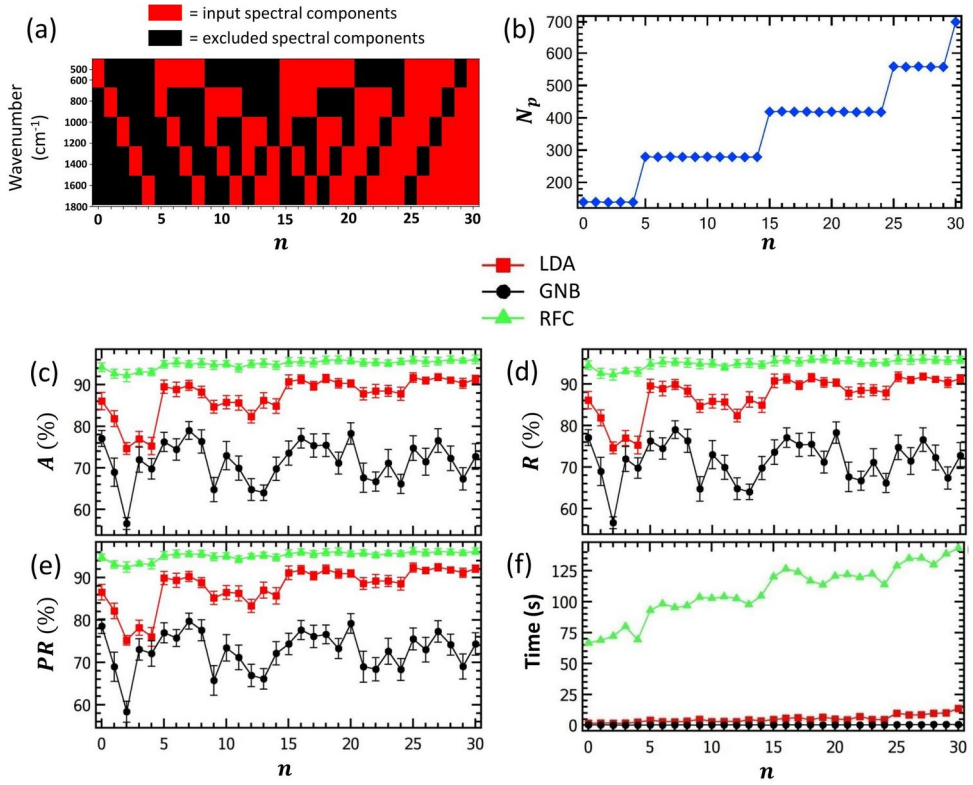

Fig. 6. (**a**): explanatory diagram illustrating the process of SPSEL. In this case, we divided the whole spectral range, 400 and 1800 cm^{-1}, represented on the vertical axis, into five bins. The horizontal axis shows the integer index n, employed to label each combination of bins. We highlighted in red the bins employed to feed the ML models and in black the bins excluded; (**b**): number N_p of points belonging to a single spectrum as a function of n; (**c**): accuracy A, averaged over N, P and DA and over the folds of the 5-fold cross-validation, as a function of n; (**d**): recall R, averaged over N, P, and DA and over the folds of the 5-fold cross-validation, as a function of n; (**e**): precision PR, averaged over N, P, and DA and over the folds of the 5-fold cross-validation, as a function of n; (**f**): duration of all the ML phases (training and test), averaged over N, P, and DA and over the folds of the 5-fold cross-validation, as a function of n. Error bars: standard deviation.

Conclusions

In this study, we explored an innovative approach to diagnosing pancreatic ductal adenocarcinoma (DA) by integrating Confocal Raman microscopy (CRM) with machine learning (ML) techniques. Our goal was not only to achieve high diagnostic accuracy but also to tackle the challenge of lowering computational cost and time, crucial elements for developing future real-time diagnostic tools for clinical use. To achieve this goal, we assessed three ML models-linear discriminant analysis (LDA), Gaussian Naive-Bayes (GNB), and random forest classifier (RFC)-in combination with dimensionality reduction techniques: Principal Component Analysis (PCA) and SPectral SELection (SPSEL). By considering the classification performances, both the approaches for the dimensionality reduction led to analogous results: RFC appeared to be the best classifier, with classification accuracies of 95–96% either fo PCA or SPSEL. On the other hand, GNB represented the worst classifier, with a maximum accuracy less than 80%. PCA turned out to be more efficient than SPSEL as a dimensionality reduction technique. In fact, PCA allowed to reach the maximum classification accuracy by feeding RFC with only 7 Principal Components, with a resulting computational time of ~ 12.0 s. On the other hand, with SPSEL 150 spectral components, corresponding to the combination $n = 17$ were needed to reach an RFC accuracy comparable with PCA. This configuration resulted in a computational time of ~ 16 s i.e. slightly larger than the computational time associated with PCA. Despite this last aspect, SPSEL potentially offers the strength of detecting Raman sub-bands corresponding to maximum classification accuracy. By definition, such relevant bands can be considered representative of biomarkers for the classes under examination. In this case, SPSEL enabled us to pinpoint peaks within the spectral interval between 400 and 707 cm^{-1}. This band comprises several molecules such as methionine, a molecule connected to DNA hypomethylation, frequently observed in different cancers. In addition, contrary to PCA, SPSEL operates directly on the spectral components, providing precious indications about the spectral sub-intervals relevant to the diagnostic process. In this sense, the perspective of realizing fast and cost-effective Raman setups operating on small spectral intervals represents a crucial landmark towards the employment of such a technology in everyday life.

The approach developed in this study effectively addresses a challenging problem in histology, i.e. the problem of distinguishing DA from P. However, further efforts have to be devoted to refining the capabilities of such a

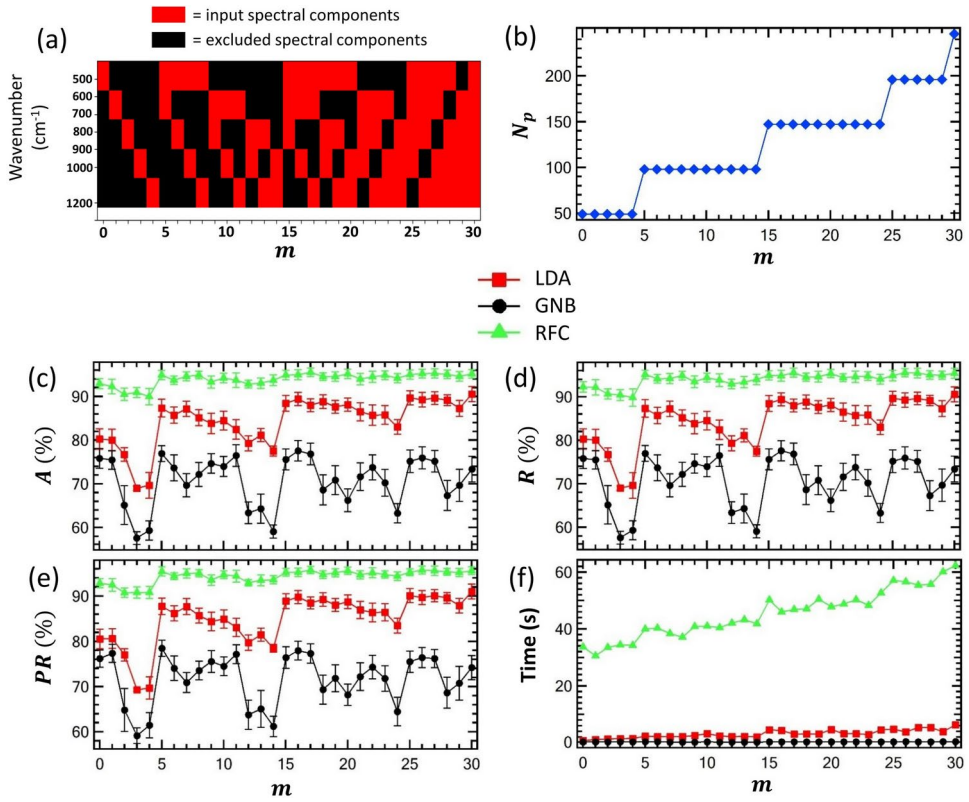

Fig. 7. (**a**): In this case, SPSEL was focused on the spectral range between 400 and 1230 cm^{-1} subdivided into five bins. The horizontal axis shows the integer index *m*, employed to label each combination of bins. We highlighted in red the bins employed to feed the ML models and in black the bins excluded; (**b**): number N_p of points belonging to a single spectrum as a function of *m*; (**c**): accuracy *A*, averaged over N, P and DA and over the folds of the 5-fold cross-validation, as a function of *m*; (**d**): recall *R*, averaged over N, P, and DA and over the folds of the 5-fold cross-validation, as a function of *m*; (**e**): precision *PR*, averaged over N, P, and DA and over the folds of the 5-fold cross-validation, as a function of *m*; (**f**): duration of all the ML phases (training and test), averaged over N, P, and DA and over the folds of the 5-fold cross-validation, as a function of *m*. Error bars: standard deviation.

technique. In this sense, the first and most straightforward step is represented by further increasing the number of classes of interest, e.g. by trying to distinguish the different grades of DA. Such additional results could be of crucial importance for the obtainment of automatic and fast diagnostic tools, designed to maximize the number of biopsies processed *ex vivo* per unit of time. Possible challenges could be represented by the high similarities between the different grades of DA, combined with the high spatial non-uniformity characterizing DA tissues. In this sense, a ML model employing a single spectrum to obtain a prediction could not be sufficient to provide a reliable diagnosis, since the spectrum could not capture sufficient tissue properties. Possible routes could be represented by using multiple spectra as a single example to feed the ML models or by acquiring the spectra at lower magnification. In this last case, the larger laser spot could allow probing optical properties within larger areas. However, the collection of light from a larger area could average out optical properties relevant to the discrimination.

The potential of this diagnostic tool could be of crucial importance in frozen-section analysis (IFSA), which is currently considered the preferred method for assessing pancreatic tumor margins during surgery. The ability to provide rapid, accurate analyses of tumor margins during surgery could improve surgical decision-making, allowing for immediate adjustments in tumor resection based on real-time feedback. This innovative approach could significantly reduce the time between tissue excision and diagnostic results, leading to more efficient surgeries and potentially better patient outcomes. However, an interesting perspective could be represented by the realization of engineered tools for intraoperative diagnosis, e.g. for the detection of tumor metastases. In this case, major challenges are represented by the need to use near-infrared radiation, which offers the advantage of penetrating deeply into the tissues, with the capability of potentially detecting invisible tumor metastatic margins. Despite this strength, near-infrared radiation weakly interacts with the tissues, with a consequent weak Raman signal.

In general, the integration of CRM and ML, especially with dimensionality reduction strategies like PCA and SPSEL, shows potential for the advancing of in vivo diagnostic devices. Although our study focused on ex vivo samples, the next step is to improve these methods for real-time clinical applications, guaranteeing that diagnostic tools are not only accurate but also fast and efficient enough for practical use. Further research should

investigate the trade-off between speed of computation and accuracy of diagnosis accuracy, particularly in the case of engineered probes for immediate in situ detection of neoplastic tissues.

Data availibility

The request for dataset, both raw and processed data, generated during the present investigation can be agreed and made directly to the corresponding author.

Received: 28 November 2024; Accepted: 9 April 2025
Published online: 17 April 2025

References

1. Hawksworth, G. et al. Pancreatic cancer trends in Europe: Epidemiology and risk factors. *Med. Stud.* **35**, 164–171 (2019).
2. Bengtsson, A., Andersson, R. & Ansari, D. The actual 5-year survivors of pancreatic ductal adenocarcinoma based on real-world data. *Sci. Rep.* **10**, 16425 (2020).
3. Cho, J. & Petrov, M. S. Pancreatitis, pancreatic cancer, and their metabolic sequelae: Projected burden to 2050. *Clin. Transl. Gastroenterol.* **11**, e00251 (2020).
4. Schneider, M., Hackert, T., Strobel, O. & Büchler, M. Technical advances in surgery for pancreatic cancer. *Br. J. Surg.* **108**, 777–785 (2021).
5. Bouchart, C. et al. Novel strategies using modern radiotherapy to improve pancreatic cancer outcomes: Toward a new standard?. *Ther. Adv. Med. Oncol.* **12**, 1758835920936093 (2020).
6. Okusaka, T. & Furuse, J. Recent advances in chemotherapy for pancreatic cancer: Evidence from Japan and recommendations in guidelines. *J. Gastroenterol.* **55**, 369–382 (2020).
7. Jiang, B. et al. Stroma-targeting therapy in pancreatic cancer: One coin with two sides?. *Front. Oncol.* **10**, 576399 (2020).
8. Ye, X., Yu, Y., Zheng, X. & Ma, H. Clinical immunotherapy in pancreatic cancer. *Cancer Immunol. Immunother.* **73**, 64 (2024).
9. Guo, X., Wang, P., Li, Y., Chang, Y. & Wang, X. Microbiomes in pancreatic cancer can be an accomplice or a weapon. *Crit. Rev. Oncol. Hematol.* **194**, 104262 (2024).
10. Zhao, Z. & Liu, W. Pancreatic cancer: A review of risk factors, diagnosis, and treatment. *Technol. Cancer Res. Treat.* **19**, 1533033820962117 (2020).
11. Kim, K. S., Kwon, J., Kim, K. & Chie, E. K. Impact of resection margin distance on survival of pancreatic cancer: A systematic review and meta-analysis. *Cancer Res. Treat. Off. J. Korean Cancer Assoc.* **49**, 824 (2017).
12. Kaltenmeier, C. et al. Impact of resection margin status in patients with pancreatic cancer: A national cohort study. *J. Gastrointest. Surg.* **25**, 2307–2316 (2021).
13. Einama, T. et al. Prognosis of pancreatic cancer based on resectability: A single center experience. *Cancers* **15**, 1101 (2023).
14. Newton, A. D. et al. Intraoperative near-infrared imaging can identify neoplasms and aid in real-time margin assessment during pancreatic resection. *Ann. Surg.* **270**, 12–20 (2019).
15. Zhang, B. et al. Revision of pancreatic neck margins based on intraoperative frozen section analysis is associated with improved survival in patients undergoing pancreatectomy for ductal adenocarcinoma. *Ann. Surg.* **274**, e134–e142 (2021).
16. Nelson, D. W., Blanchard, T. H., Causey, M. W., Homann, J. F. & Brown, T. A. Examining the accuracy and clinical usefulness of intraoperative frozen section analysis in the management of pancreatic lesions. *Am. J. Surg.* **205**, 613–617 (2013).
17. Šteňo, A. et al. Current limitations of intraoperative ultrasound in brain tumor surgery. *Front. Oncol.* **11**, 659048 (2021).
18. Egloff-Juras, C., Bezdetnaya, L., Dolivet, G. & Lassalle, H.-P. NIR fluorescence-guided tumor surgery: New strategies for the use of indocyanine green. *Int. J. Nanomed.* **14**, 7823–7838 (2019).
19. Aaboubout, Y. et al. Intraoperative assessment of resection margins by Raman spectroscopy to guide oral cancer surgery. *Analyst* **148**, 4116–4126 (2023).
20. Kouri, M. A. et al. Raman spectroscopy: A personalized decision-making tool on clinicians' hands for in situ cancer diagnosis and surgery guidance. *Cancers* **14**, 1144 (2022).
21. Krafft, C., Popp, J., Bronsert, P. & Miernik, A. Raman spectroscopic imaging of human bladder resectates towards intraoperative cancer assessment. *Cancers* **15**, 2162 (2023).
22. Raman, C. V. & Krishnan, K. S. A new type of secondary radiation. *Nature* **121**, 501–502 (1928).
23. Conforti, P. M., Lazzini, G., Russo, P. & D'Acunto, M. Raman spectroscopy and ai applications in cancer grading. An overview. *IEEE Access* **12**, 54816–54852 (2024).
24. Conti, F. et al. Raman spectroscopy and topological machine learning for cancer grading. *Sci. Rep.* **13**, 7282 (2023).
25. Lazzini, G. & D'Acunto, M. Grading of melanoma tissues by Raman microspectroscopy. *Eng. Proc.* **51**, 10 (2023).
26. Lazzini, G. & D'Acunto, M. Chondrogenic cancer grading by combining machine and deep learning with Raman spectra of histopathological tissues. *Appl. Sci. (2076–3417)* **14**, 10555 (2024).
27. Hanna, K. et al. Raman spectroscopy: Current applications in breast cancer diagnosis, challenges and future prospects. *Br. J. Cancer* **126**, 1125–1139 (2022).
28. Liu, K., Zhao, Q., Li, B. & Zhao, X. Raman spectroscopy: A novel technology for gastric cancer diagnosis. *Front. Bioeng. Biotechnol.* **10**, 856591 (2022).
29. Pandya, A. K. et al. Evaluation of pancreatic cancer with Raman spectroscopy in a mouse model. *Pancreas* **36**, e1–e8 (2008).
30. Li, Z. et al. Detection of pancreatic cancer by convolutional-neural-network-assisted spontaneous Raman spectroscopy with critical feature visualization. *Neural Netw.* **144**, 455–464 (2021).
31. Mandrell, C. T. et al. Machine learning approach to Raman spectrum analysis of MIA PaCa-2 pancreatic cancer tumor repopulating cells for classification and feature analysis. *Life* **10**, 181 (2020).
32. Uthamacumaran, A. et al. Machine learning characterization of cancer patients-derived extracellular vesicles using vibrational spectroscopies: Results from a pilot study. *Appl. Intell.* **52**, 12737–12753 (2022).
33. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
34. Zhao, J., Lui, H., McLean, D. I. & Zeng, H. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *Appl. Spectrosc.* **61**, 1225–1232 (2007).
35. Xanthopoulos, P. *et al.* Linear discriminant analysis. *In Robust data mining* 27–33 (2013).
36. Tharwat, A., Gaber, T., Ibrahim, A. & Hassanien, A. E. Linear discriminant analysis: A detailed tutorial. *AI Commun.* **30**, 169–190 (2017).
37. Reddy, E. M. K., Gurrala, A., Hasitha, V. B. & Kumar, K. V. R. Introduction to naive bayes and a review on its subtypes with applications. *In Bayesian reasoning and gaussian processes for machine learning applications* 1–14 (2022).
38. Genuer, R., Poggi, J.-M., Genuer, R. & Poggi, J.-M. *Random forests* (Springer, 2020).
39. Foody, G. M. Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PLoS One* **18**, e0291908 (2023).

40. Jolliffe, I. T. & Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
41. Araújo, D. C. et al. Finding reduced Raman spectroscopy fingerprint of skin samples for melanoma diagnosis through machine learning. *Artif. Intell. Med.* **120**, 102161 (2021).
42. Glunde, K., Jacobs, M. A. & Bhujwalla, Z. M. Choline metabolism in cancer: Implications for diagnosis and therapy. *Expert Rev. Mol. Diagn.* **6**, 821–829 (2006).
43. Lowenfels, A. B. & Maisonneuve, P. Methionine intake and pancreatic cancer risk: Digesting the evidence. *Gastroenterology* **132**, 441–443 (2007).
44. Movasaghi, Z., Rehman, S. & Rehman, I. U. Raman spectroscopy of biological tissues. *Appl. Spectrosc. Rev.* **42**, 493–541 (2007).
45. Rygula, A. et al. Raman spectroscopy of proteins: A review. *J. Raman Spectrosc.* **44**, 1061–1076 (2013).
46. Medipally, D. K. et al. Vibrational spectroscopy of liquid biopsies for prostate cancer diagnosis. *Ther. Adv. Med. Oncol.* **12**, 1758835920918499 (2020).
47. Naseer, K., Saleem, M. & Qazi, J. Optical diagnosis of typhoid infection in human blood sera using Raman spectroscopy. *Spectrosc. Lett.* **53**, 249–255 (2020).
48. Cao, X. et al. Hollow au nanoflower substrates for identification and discrimination of the differentiation of bone marrow mesenchymal stem cells by surface-enhanced Raman spectroscopy. *J. Mater. Chem. B* **5**, 5983–5995 (2017).
49. Ning, T. et al. Raman spectroscopy based pathological analysis and discrimination of formalin fixed paraffin embedded breast cancer tissue. *Vib. Spectrosc.* **115**, 103260 (2021).
50. Mondol, A. S. et al. High-content screening Raman spectroscopy (HCS-RS) of panitumumab-exposed colorectal cancer cells. *Analyst* **144**, 6098–6107 (2019).
51. Shamina, L. A. et al. Raman and autofluorescence analysis of human body fluids from patients with malignant tumors. *J. Biomed. Photonics Eng.* **3**, 020308 (2017).
52. Elumalai, B., Prakasarao, A., Ganesan, B., Dornadula, K. & Ganesan, S. Raman spectroscopic characterization of urine of normal and oral cancer subjects. *J. Raman Spectrosc.* **46**, 84–93 (2015).
53. Contorno, S., Darienzo, R. E. & Tannenbaum, R. Evaluation of aromatic amino acids as potential biomarkers in breast cancer by Raman spectroscopy analysis. *Sci. Rep.* **11**, 1698 (2021).
54. Maitra, I. et al. Raman spectral discrimination in human liquid biopsies of oesophageal transformation to adenocarcinoma. *J. Biophotonics* **13**, e201960132 (2020).
55. Amber, A., Nawaz, H., Bhatti, H. N. & Mushtaq, Z. Surface-enhanced Raman spectroscopy for the characterization of different anatomical subtypes of oral cavity cancer. *Photodiagn. Photodyn. Ther.* **42**, 103607 (2023).
56. Teske, C. et al. Label-free differentiation of human pancreatic cancer, pancreatitis, and normal pancreatic tissue by molecular spectroscopy. *J. Biomed. Opt.* **27**, 075001–075001 (2022).
57. Oo, S.-L. et al. Highly sensitive and cost-effective portable sensor for early gastric carcinoma diagnosis. *Sensors* **21**, 2639 (2021).
58. Chen, H. et al. Rapid and sensitive detection of esophageal cancer by FTIR spectroscopy of serum and plasma. *Photodiagn. Photodyn. Ther.* **40**, 103177 (2022).
59. Banerjee, A. et al. Metabolomics profiling of pituitary adenomas by Raman spectroscopy, attenuated total reflection-Fourier transform infrared spectroscopy, and mass spectrometry of serum samples. *Anal. Chem.* **94**, 11898–11907 (2022).

Acknowledgements

The authors wish to thank Cost Action CA21116 "Identification of biological markers for prevention and translational medicine in pancreatic cancer (TRANSPAN)". M.D. and G.L. thank TELEMO project Bando Ricerca Salute 2018. THE - Tuscany Health Ecosystem grant ECS_00000017. NanoBioTLab CNR-IBF and the joint laboratory BIOICT Lab (http://www.pi.ibf.cnr.it/) are warmly acknowledged by G.L. and M.D.

Author contributions

M.D. conceived the investigation, N.F. and L.M. surgically removed the samples from the patients, R.G., L.E.P., and A.C. prepared glass slides with cancer tissues to be analyzed, G.L. and M.D. conducted the Raman acquisition, G.L. used the ML algorithms, G.L., R.G., L.E.P., and M.D. analyzed the results. All authors reviewed the manuscript.

Declarations

Comepting interests

The authors declare no competing interests.

Ethical approval

Owing to the observational nature of this study, 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". This study was conducted in accordance with the principles of the Declaration of Helsinki. All patients provided informed consent authorizing the anonymous scientific use of their collected data.

Additional information

Correspondence and requests for materials should be addressed to M.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025