Analytical and Bioanalytical Chemistry (2023) 415:7281–7295
https://doi.org/10.1007/s00216-023-04997-w

RESEARCH PAPER

Combined analytical approach empowers precise spectroscopic interpretation of subcellular components of pancreatic cancer cells

Krzysztof Szymoński[1,2] · Katarzyna Skirlińska-Nosek[3,4] · Ewelina Lipiec[3] · Kamila Sofińska[3] · Michał Czaja[3,4] · Natalia Wilkosz[3,5] · Matylda Krupa[1] · Filip Wanat[1] · Magdalena Ulatowska-Białas[1,2] · Dariusz Adamek[1]

Received: 4 September 2023 / Revised: 27 September 2023 / Accepted: 9 October 2023 / Published online: 31 October 2023

Abstract

The lack of specific and sensitive early diagnostic options for pancreatic cancer (PC) results in patients being largely diagnosed with late-stage disease, thus inoperable and burdened with high mortality. Molecular spectroscopic methodologies, such as Raman or infrared spectroscopies, show promise in becoming a leader in screening for early-stage cancer diseases, including PC. However, should such technology be introduced, the identification of differentiating spectral features between various cancer types is required. This would not be possible without the precise extraction of spectra without the contamination by necrosis, inflammation, desmoplasia, or extracellular fluids such as mucous that surround tumor cells. Moreover, an efficient methodology for their interpretation has not been well defined. In this study, we compared different methods of spectral analysis to find the best for investigating the biomolecular composition of PC cells cytoplasm and nuclei separately. Sixteen PC tissue samples of main PC subtypes (ductal adenocarcinoma, intraductal papillary mucinous carcinoma, and ampulla of Vater carcinoma) were collected with Raman hyperspectral mapping, resulting in 191,355 Raman spectra and analyzed with comparative methodologies, specifically, hierarchical cluster analysis, non-negative matrix factorization, T-distributed stochastic neighbor embedding, principal components analysis (PCA), and convolutional neural networks (CNN). As a result, we propose an innovative approach to spectra classification by CNN, combined with PCA for molecular characterization. The CNN-based spectra classification achieved over 98% successful validation rate. Subsequent analyses of spectral features revealed differences among PC subtypes and between the cytoplasm and nuclei of their cells. Our study establishes an optimal methodology for cancer tissue spectral data classification and interpretation that allows precise and cognitive studies of cancer cells and their subcellular components, without mixing the results with cancer-surrounding tissue. As a proof of concept, we describe findings that add to the spectroscopic understanding of PC.

Keywords Pancreatic cancer · Ampullary cancer · Molecular spectroscopy · Convolutional neural networks · Raman imaging · Spectral analysis

✉ Krzysztof Szymoński
krzysztof.szymonski@uj.edu.pl

[1] Department of Pathomorphology, Medical College, Jagiellonian University, Kraków, Poland

[2] Department of Pathomorphology, University Hospital, Kraków, Poland

[3] Faculty of Physics, Astronomy and Applied Computer Science, M. Smoluchowski Institute of Physics, Jagiellonian University, Kraków, Poland

[4] Doctoral School of Exact and Natural Sciences, Jagiellonian University, Kraków, Poland

[5] AGH University of Krakow, Faculty of Physics and Applied Computer Science, Kraków, Poland

Abbreviations

PC	Pancreatic cancer
RHM	Raman hyperspectral mapping
VS	Vibrational spectroscopy
RS	Raman spectroscopy
NMF	Non-negative matrix factorization
tSNE	T-distributed stochastic neighbor embedding
PCA	Principal component analysis
HCA	Hierarchical cluster analysis
CNN	Convolutional neural network
NN	Neural network
IPMN	Intraductal papillary mucinous neoplasm
PanIN	Pancreatic intraepithelial neoplasia

cPDAC	Conventional pancreatic ductal adenocarcinoma
AVAC	Adenocarcinoma of the ampulla of Vater
IPMC	Carcinoma derived from IPMN
FTIR	Fourier transform infrared spectroscopy
ATR-FTIR	Attenuated total reflection Fourier transform infrared spectroscopy
SERS	Surface-enhanced Raman spectroscopy
Trp	Tryptophan
Tyr	Tyrosine
Phe	Phenylalanine
PC-1, PC-2, PC-3	Principal components 1, 2, 3
RASSF1A	Ras association domain family 1A
PI3K	Phosphoinositide 3-kinase
FAK	Focal adhesion kinase
HDR	Homology-directed repair pathway
DSB	Double-strand breaks
BRCA1	Breast cancer type 1 susceptibility protein
NF-κB	Nuclear factor kappa-light chain enhancer of activated B cells
IGFBP2	Insulin-like growth factor-binding protein 2

Introduction

The rising understanding of pathomechanisms of pancreatic cancer (PC) initiation and evolution that we have witnessed in recent decades [1] has not improved the very low PC 5-year survival rates, which remain below 10% [2]. Due to the lack of specific and sensitive early diagnostic options, patients are largely diagnosed with late-stage disease [1]. The PC tumors' molecular and morphological heterogeneity is also responsible for the PC being chemoresistant to available treatment options [1]. These are the main reasons for the drastically poor prognosis of PC patients [1]. New methods of studying the molecular composition of PC are required to develop efficient early detection technologies, as well as to extend the knowledge of the mechanisms of chemoresistance [3] and counteract them. Despite the urgent need for efficient and universal malignancy screening technologies, currently, none would fulfill the criteria of good specificity and sensitivity for PC. However, multiple serum-based biomarkers have been proposed without satisfactory results [4–6]. Some authors reported better usefulness in assessing interleukin-6 (IL-6) serum levels in differentiating PC patients from chronic or acute pancreatitis [7–9], or recently, leukemia inhibitory factor (LIF) was reported to be a promising serum biomarker of pancreatic malignancy [10]. Nevertheless, all these are only singular protein markers, which entails unsatisfactory diagnostic specificity and sensitivity [11]. Moreover, the single-biomarker methods cannot generalize to multiple varieties of cancer.

Serum liquid biopsy samples are considered ideal for cancer screening [6, 11]. Nevertheless, conventional liquid biopsy biomarkers (LBMs) comprising of circulating tumor nucleic acids (i.e., ctDNA, ctRNA), circulating tumor cells (CTCs), or extracellular vesicles (EVs) have not been introduced into medical practice because of limitations of the molecular/genetic testing techniques, such as described in [6].

A promising twist in the liquid biopsy analysis might be brought by the implementation of methods of vibrational spectroscopy (VS) such as surface-enhanced Raman spectroscopy (SERS) or attenuated total reflection Fourier-transformed infrared spectroscopy (ATR-FTIR) [3, 12]. VS was confirmed to be an excellent tool for the characterization of malignant tissue's chemical structure and composition [13–16]. Due to the fingerprint-like character of resulting data acquired from VS, all information about the studied sample is ready for interpretation, making VS a universal technique of molecular characterization. Nevertheless, only a few studies utilized VS for the differentiation of multiple malignancies [17, 18], with most papers comparing only malignant *vs.* benign control [19]. Further exploration is needed to reveal the full potential of molecular spectroscopy in cancer screening. The only way for promising VS to be introduced as a diagnostic technology is by detailed characterization of the spectral results obtained for various cancer types.

Following this, in our study, we aimed to deepen the knowledge of VS landscapes of PC tumors by using an innovative and comprehensive methodology, including combined Raman hyperspectral mapping (RHM), conventional multivariate data analysis, and deep networking techniques. This approach allowed for the cognitive recognition of spectral markers of PC. We separately measured and analyzed cellular nuclei, the cytoplasm, and the tumors' stroma compartment of main groups of PC, specifically the conventional pancreatic ductal adenocarcinoma (cPDAC), intraductal papillary mucinous carcinoma (IPMC), and ampulla of Vater adenocarcinoma (AVAC). Often these tumors are indistinguishable using standard histopathological evaluation techniques (i.e., the tumor's epicenter location, morphology, and immunohistochemistry) [1, 20, 21]. Although current clinical management protocols recommend treating these tumors similarly, significant differences in cancer differentiation level, the occurrence of perineural and venous invasion, and lymph node involvement were reported [22]. A large study on cPDAC vs AVAC (476 vs 232 cases) resulted in showing significant differences in patients' survival, specifically 15.6 vs 41 months, for cPDAC and AVAC, respectively [23]. Notably, in this study by Reid et al., the authors describe

no impact of tumor size and lymph node metastasis on the patients' survival [23].

Methods of VS complement each other and are usually used in different ways considering the type of samples that are to be studied. For example, as stated by some authors [3, 24, 25], tissues are best measured by the RHM technique, whereas blood serum probing involves highly sensitive SERS [26–30] and ATR-FTIR [31–34], which do not provide spatial resolution; however, they are effective in the investigation into bulk samples such as blood serum, thus ideal for early diagnostic of malignancies, such as PC. We believe that for the successful implementation of VS serum-based diagnostic technologies, firstly, an understanding of the spectroscopic characteristics of the tumors is required. Techniques of hyperspectral imaging (such as RHM) allow precise selection of cancer areas to further analyze them. The learned knowledge might be subsequently translated into serum-based liquid biopsy ATR-FTIR or SERS measurements.

Because of the nature of the Raman effect, VS is very sensitive to distortion factors. Fluorescence, thermal noise, and the measuring equipment quality might have a major impact on the results; thus, typically spectra preprocessing is required. Usually, it involves cosmic ray removal, baseline correction, and smoothing. During these operations, some seemingly irrelevant data might be lost if an improper preprocessing model was applied [35]. Conversely, CNNs are surprisingly efficient in classifying raw, unpreprocessed data [36], although they generalize better if unmeaningful information is removed. The successful use of neural networks, such as CNNs in spectroscopic data evaluation and classification, was shown in multiple studies [24, 36, 37]. Briefly, the CNN is trained by allowing it to analyze spectra from the so-called training dataset, from which CNN identifies characteristic features. The main advantage of CNN among other neural network types is its ability to self-extract discriminating features (automatic features extractor) [38]. Although manual feature extraction with proper setup may be as good as automatic in some classification scenarios [39], in cancer diagnostics, automation and high throughput of the process are crucially important [3].

RHM enables high-resolution imaging of tissue samples without the need for special labeling and with comparable costs to other standard techniques, such as magnetic resonance [40]. RHM is a molecular spectroscopy technique that utilizes multiple RS measurements of adjacent parts of the studied sample, followed by plotting the combined results as a tissue map image. The use of Raman spectroscopy (RS) to investigate cancer tissue samples is not new [14, 15]; however, only a few studies utilized RHM for the cognitive selection of areas of analysis [24, 25]. Contrary to RS random blind spot measurements [41] or rare grid mapping [42], this approach prevents mixing the results with areas of necrosis, inflammation, fibrosis, or colloid [3]. Without this precision, the interpretation of the molecular contents of PC cells is impossible, let alone the cellular cytoplasm and nuclei separately.

Here, as a first part of the study, we conducted comparative studies between spectral data analysis methodologies to reveal the fittest for defining spectroscopic landscapes of cytoplasm and nuclei of PC subgroups separately. Specifically, we compared hierarchical cluster analysis (HCA), non-negative matrix factorization (NMF), T-distributed stochastic neighbor embedding (tSNE), principal components analysis (PCA), and convolutional neural networks (CNN). Each has advantages and limitations, which we describe briefly in the *Supplementary section – Methods of multivariate data analysis used in the study*. In conclusion, we propose using a combined approach (CNN + PCA) that allows for automatic and high-throughput spectra classification and subsequent comprehensive characterization of the smallest spectral differences among them allowing molecular interpretation.

Methods

Tissue slide preparation

Raman imaging of 16 PC tissue slides from 15 patients was conducted. Specifically, 6 AVAC, 5 cPDAC, and 5 IPMC were included. The tissue samples were collected from patients with a diagnosis of PC who underwent pancreatoduodenectomy (Whipple or Traverso) or distal pancreatectomy, with the exclusion of benign pancreatic neoplasm or neuroendocrine neoplasm cases. The details of patients included in the study are summarized in Supplementary Table S1. In this study, we used methods of tissue slide preparation already described by others [24]. Briefly, tissue samples were selected from the Cracow University Hospital's Pathomorphology Department's archive, normally stored as conventional formalin-fixed paraffin-embedded (FFPE) blocks after the diagnostic process. The initial sample selection was performed by two independent experienced pancreatic pathologists, by assessing the standard hematoxylin–eosin-stained glass slides (H&E). A routine light microscope (Olympus BX53 Microscope, RRID:SCR_022568) was used for this stage. During the selection process, a detailed reevaluation of the tumor type was conducted, and initial diagnoses were confirmed. Subsequently, before Raman measurements, for each selected case, a single 2.5-µm-thick tissue section was sliced with a Microm® HM355S Automatic Microtome and mounted onto a CaF_2 window (Raman Grade Calcium Fluoride substrates – CRYSTRAN LTD, England). Then, on unstained CaF_2 slides, areas of cancer were marked by pathologists. A complete paraffin removal

procedure was conducted involving a 12-h xylene bath and graded ethanol rehydration.

Raman measurements

After PC tissue slide preparation, Raman measurements were executed using the already described procedure [24, 25]. Briefly, RS was conducted with a Horiba LabRam spectrometer equipped with a green (532 nm) laser and electron-multiplying charge-coupled device (EM-CCD) camera cooled to − 70 °C. During the measurements, a × 60 water immersion objective lens (Nikon) was used to allow measurements of tissue sections immersed in a physiological saline solution. Spectra were acquired in the fingerprint spectral region (1900–600 cm^{-1}) with a spectral resolution of 2 cm^{-1}. The RHM maps included 6724 to 13,284 spectra for a single slide in this study. The exposure time for each pixel was 6 s. The pixel size (step size) was 1 µm or smaller depending on the size of the preselected area of cancer (see "Tissue slide preparation"), which varied from 80 × 80 µm to 140 × 140 µm.

CNN dataset annotation and data augmentation

Spectral data for the training of the CNN were selected from obtained RHM maps, by direct comparison with the unstained optical microscopy tissue slide images, normally collected before RS measurements. It was performed by the same pathologists, who preselected the areas of cancer (see "Tissue slide preparation"). Additional comparisons with H&E-stained slides were also performed, which clarified cellular components (such as the cancer cells' nuclei) better. However, direct transmission of areas of nuclei and cytoplasm from H&E images was not possible, because the H&E slides were sliced from FFPE blocks before the tissues for RHM measurements, thus differing slightly from the investigated tissues. The LassoSelector widget from the MatPlotLib library (MatPlotLib, RRID:SCR_008624) of Python (IPython, RRID:SCR_001658) was used to annotate the data into 7 separate classes, including nuclear and cytoplasmic areas of each of the PC types (AVAC, IPMC, and cPDAC). Additionally, the stroma/empty class was annotated. Only a small part of all spectra in the RHM map was annotated for the training dataset (42,098 spectra from 191,355 — approximately 22%), leaving the rest for the CNN model validation with new data, testing its ability to generalize. The process of training dataset annotation is depicted in Supplementary Figure S1. Before feeding the data into CNN, it was preprocessed, which involved sequentially applying a baseline correction (3rd polynomial order), smoothing with the Savitzky–Golay algorithm (third-order, 15 smoothing points), and trimming in the spectral range characteristic for biological molecules (1800–650 cm^{-1}). Subsequently, to prevent overfitting with the smoothed spectra, we applied data augmentation, by adding random noise to each spectrum. After the augmentation, the CNN training dataset included 126,294 spectra.

Additionally, to prevent the impact of different Raman shifts on the prediction results, spectral intensity values were packed with the Raman shift values into a single integer (int64) value.

After achieving 92% validation accuracy in the CNN training, to increase the model performance, we used another neural network dense classifier. This model required another training on different data. For each spectrum in the training dataset of the initial CNN model, we counted spectral ratios of (i) DNA methylation, (ii) β-sheet proteins, and (iii) random coil proteins. Some of these ratios were described to be specific for PC subtypes in VS spectra [24]. Specifically, for each spectrum, the relation between Raman bands characteristic of DNA methylation ($\delta(CH_2, CH_3)$, 1420–1360 cm^{-1} to $\nu_s(PO_2^-)$, 1150–1050 cm^{-1}), the β-sheet secondary structure of proteins (the β-sheet amide III, 1228–1218 cm^{-1} to total amide I, 1750–1514 cm^{-1}), and proteins' random coil secondary structure (the random coil amide I, 1640–1664 cm^{-1} to total amide I, 1750–1514 cm^{-1}) were established. The counted ratios were combined with the spectral prediction result (performed by the initial CNN model) and fed into the shallow dense classifier.

CNN architectures, training, and validation

In this study, we used two neural network models. First, for the spectra classification, a CNN was used, while for the final, combined spectral + ratio classification we utilized a shallow dense classifier. Below are the details of each of them.

The CNN was trained in predicting seven classes, representing separately the nucleus and the cytoplasm of AVAC, IPMC, and cPDAC, and additionally the stroma/empty space class. We used a custom-designed CNN architecture with 13 1D convolutional layers for feature identification and 3 fully connected layers for classification. A "sequential" base model was used. For each layer, a "glorot uniform" initializing mode was used. The "Adam" optimizer and "categorical crossentropy" loss function was applied. The CNN training involved 40 epochs with a batch size equal to 105.

The second dense classifier involved 3 fully connected layers. The batch size equaled 9 and the training took 17 epochs.

The total number of spectra used for CNN training was 126,294 with the 2D NumPy (NumPy, RRID:SCR_008633) array shape presented as (126,294, 320) and included spectra for each of 7 classes. Then, class arrays were "one-hot encoded," and the training and testing dataset split was

conducted with a 70/30 ratio. The initial CNN training time was 40 min.

Each 2D NumPy array of training dataset for the final classifier included values for (i) summed intensities of DNA bands, (ii) summed intensities of methylated DNA bands, (iii) DNA methylation ratio value, (iv) summed intensities of total proteins bands, (v) summed intensities of beta-sheet proteins bands, (vi) beta-sheet protein ratio, (vii) summed intensities of random coil protein bands, (viii) random coil protein ratio, and (ix) CNN prediction class for the spectrum. The shape of this NumPy array was (126,294, 9).

The programming of the CNN and dense classifier was performed in Python version 3.10.5 (IPython, RRID:SCR_001658) with TensorFlow (RRID:SCR_016345) and Keras application programming interfaces (API). The proposed CNN and dense classifier architecture details are summarized in Supplementary Figure S2.

After the CNN and final classifier were trained with satisfying performance, additional RHM spectral data was obtained from new PC patient tissues. This RHM map included 13,924 spectra not "seen" by the CNN during the training phase. The performance of the prediction on this tissue sample was recognized.

Visualizing the CNN classification results

An efficient way of visualizing the spectral data prediction is to present the CNN classification results by plotting it as a tissue map image (prediction map) [24]. Every spectrum obtained with RHM for each tissue sample was fed into the CNN, followed by the final "ratios" classifier, and marked as one of the 7 classes. The predicted class values (classes 0–6 standing for stroma/empty, AVAC nucleus, AVAC cytoplasm, cPDAC nucleus, cPDAC cytoplasm, IPMC nucleus, and IPMC cytoplasm) created an array, which combined with the x and y coordinates of the original RHM map enabled the plotting of the prediction map. Each image pixel expressed a CNN-predicted class with a different color. The plotting was performed in Python (IPython, RRID:SCR_001658) with the MatPlotLib library (MatPlotLib, RRID:SCR_008624). The total number of spectra used for generating the prediction maps was 191,355, of which 149,257 were new spectral data not trained on by the CNN (approximately 78%).

Multivariate data analysis methodology

Data analysis was conducted in the MATLAB (RRID:SCR_001622) environment from MathWorks (Natick, USA) (HCA, NMF) and in Python (IPython, RRID:SCR_001658) with Sklearn library (Sklearn, RRID:SCR_019053) (PCA, tSNE). Before the analyses, spectral data were preprocessed similarly to the spectra used for training the CNN (see "CNN dataset annotation and data augmentation"). The spectral range used for data analyses was trimmed to 1800–650 cm^{-1}, which is desirable for biological structure examinations of samples obtained from FFPE tissue blocks. Certainly, in the higher spectral ranges (3,100–2,800 cm^{-1}), one could expect the C-H stretching motions from methyl and methylene functional groups mainly from lipids; however, due to the required process of deparaffinization, the lipids in the samples were washed out (see "Tissue slide preparation").

Results

CNN classifies Raman spectra of PC types with high accuracy and generalizes well on new data

To retrieve the subcellular components of AVAC, cPDAC, and IPMC, we trained a custom CNN model (see "Methods" for a detailed description of the methodology). All spectra from RHM maps obtained from PC tissue sections were fed into pre-trained CNN in prediction mode. CNN classified the spectra into 7 classes (see "Visualizing the CNN classification results"). The validation accuracy of the prediction process reached slightly above 97%. The exemplary results of this part of our study are depicted in Fig. 1 as CNN prediction maps plotted for each type of PC tumor (AVAC, cPDAC, and IPMC). In that figure, the comparison of obtained map images to unstained light microscope tissue images, taken in the corresponding spot to the Raman measurements, proper cellular and subcellular tissue elements (the nuclei or cytoplasm) can be distinguished (classified adequately by the CNN). To the left of each figure 1st row panel, the H&E-stained tissue image is presented; however, these sections were sliced from the FFPE tissue blocks before the sections used for Raman measurements (see "Tissue slide preparation"); thus, they might be slightly different regarding the placement of tissue and cellular components. Nevertheless, the H&E image highlights the cells, nuclei, and cytoplasm of cells, showing a clear resemblance between unstained sections and CNN-predicted map images.

After achieving good CNN model validation performance on spectra obtained from the same PC tissues as these used for the training dataset (note that only approximately 22% of spectra were annotated, followed by training–validating dataset split in 70:30 ratio — see "CNN architectures, training, and validation"), to show the eligibility of our CNN model to generalize, we validated it on a new PC sample. To accomplish this, we sliced another tissue slide from the FFPE block obtained from a new PC case and conducted RHM measurements. None of the spectra from that tissue sample was used for CNN training, making the whole RHM map a validation field. The results of CNN prediction

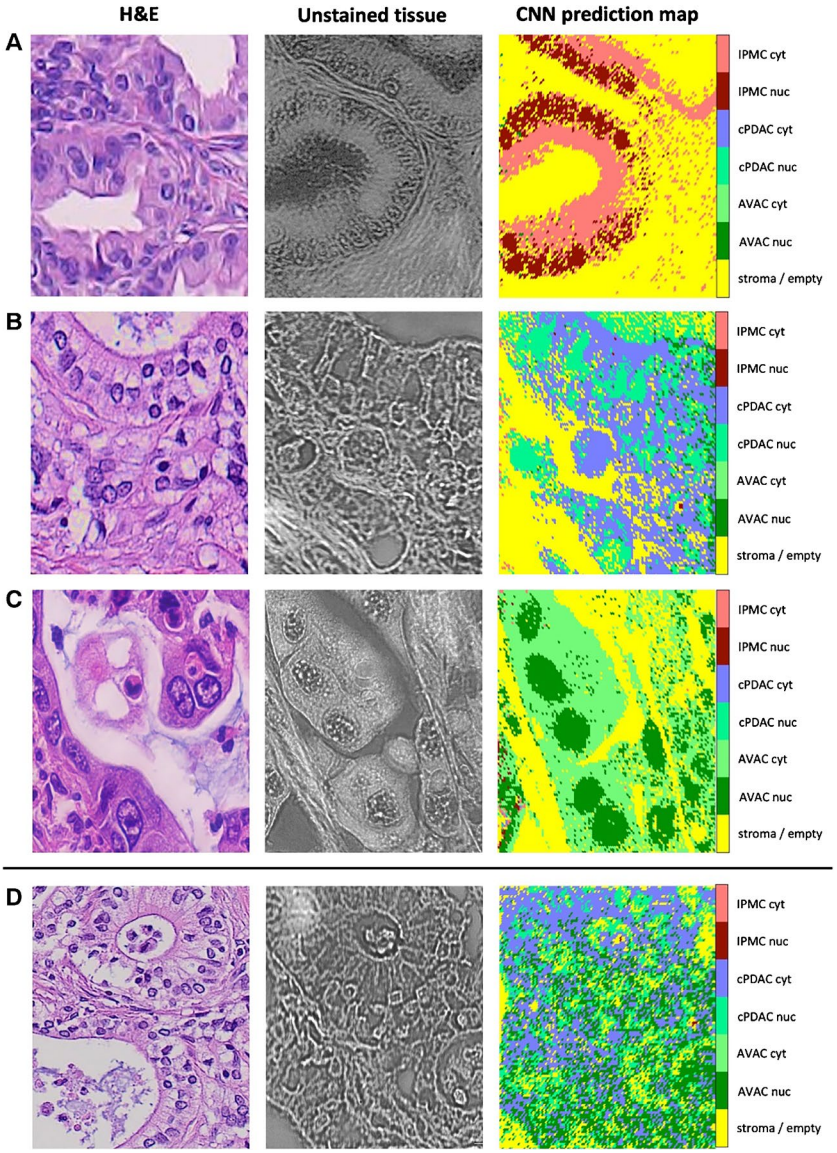

Fig. 1 Tissue map images of PC tumors generated by the CNN. The samples of IPMC (**A**), cPDAC (**B**), and AVAC (**C**) tissues were classified by the CNN into 7 classes. Some of the spectra (about 22%) were initially annotated for the CNN training dataset. An additional RHM map was obtained from the cPDAC tissue section and fed into the pre-trained CNN (without retraining it with the new data) to check the generalization efficacy of the CNN model (**D**). Note that the H&E slides were sliced from the FFPE blocks before the slides used for the Raman measurements; thus, differences in tissue details might be observed (H&E, hematoxylin–eosin stain, original magnification 40×)

plotting are presented in Fig. 1D and show the proper classification of a very complex tissue section of cPDAC.

Comparison studies of different spectral analysis methodologies

In Supplementary Figure S3, we present exemplary results of different analytical approaches to spectral data. The CNN (Supplementary Figure S3A) identifies spectra of interest (originating from such components as cellular nuclei and cytoplasm) with great accuracy (over 97%). Conversely, with HCA (Supplementary Figure S3B), the distinction of nuclear areas is less exposed. On the other hand, NMF reveals the distribution of separate molecular components, such as nucleic acids, proteins, or water (Supplementary Figures S3C, S3D, and S3E).

The results of PCA and tSNE were similar in our datasets (Supplementary Figure S4); however, tSNE would not provide insight into the spectral characteristics (such as the loadings plots for PCA) of PC subtypes' cellular nuclei and cytoplasm, which was the main goal of our study — to recognize and describe the spectral landscapes of AVAC, cPDAC, and IPMC. As a result, we propose a combined approach, utilizing CNN for spectral classification of subcellular components and then PCA on the extracted spectra to identify the molecular details of each CNN-predicted class. This allows for an automatic, high-throughput, yet detailed characterization of PC tissue samples, on a subcellular scale. Moreover, this methodology might be adjusted in the analysis of any cancerous tissues.

After the successful CNN-based classification, spectra were analyzed by PCA concerning the distinction between

those used for the CNN training dataset and the ones that the CNN model was validated on. In Supplementary Figure S3, the exemplary results considering spectra from the cytoplasm of AVAC cells are presented. As expected, such explorations revealed no significant spectral differences (Supplementary Figure S5A); however, the generalization ability of our CNN model is well illustrated with the extended plotting of the prediction spectra in the PCA scores (Supplementary Figure S5B).

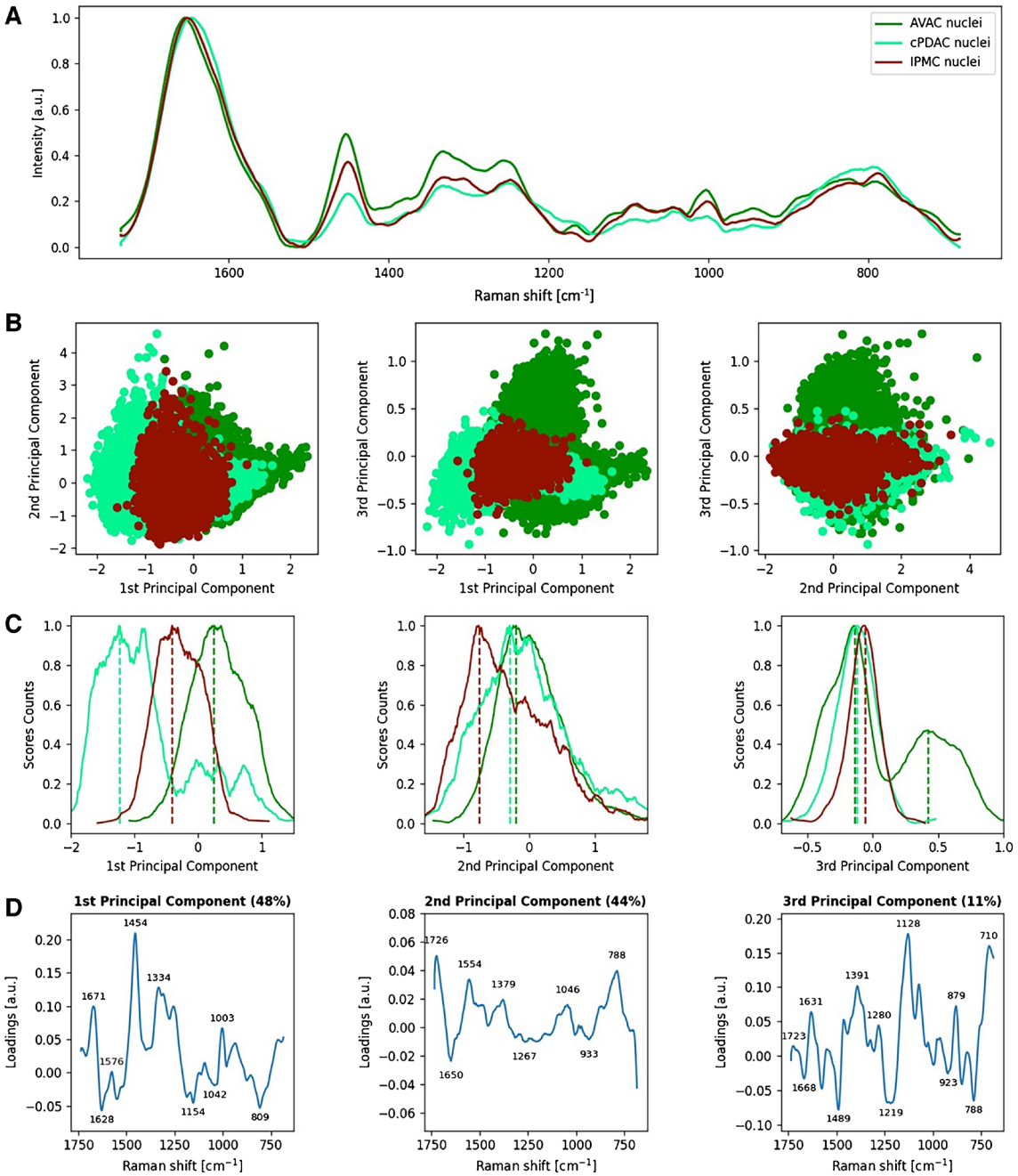

Fig. 2 Molecular characteristics of the nuclei of PC tumors. Raman spectra from nuclei of all cells obtained from AVAC, cPDAC, and IPMC samples are presented as **A** linearly plotted mean spectra, **B** PCA 2D scores plots, **C** the PCA scores count plot showing the separation of their centroids, and **D** corresponding loadings plots with significant peaks marked

Variations in global DNA methylation and contents of β-sheet-rich intranuclear proteins among AVAC, cPDAC, and IPMC revealed with the PCA of Raman spectra CNN classified as PC nuclei

The results of PCA analysis performed on spectra extracted from nuclear areas of all studied PC types' RHM maps are depicted in Fig. 2, as 2D score plots of spectra clustering (Fig. 2B) and the corresponding loading plots showing peaks responsible for spectra separation (Fig. 2D). For the sake of clarity, to show the centroid positions of PCA scores and their shifts along the principal components, the counts of the scores were plotted (Fig. 2C).

In Fig. 2, the loading plot, which explains separation along the principal component 1 (PC-1), is dominated by the protein bands such as the phenylalanine (Phe) ring breathing mode at 1004 cm^{-1}; the amide I and amide III in the spectral ranges of 1700–1600 cm^{-1} and 1340–1230 cm^{-1}, respectively; and the methyl and methylene bending motions at 1454 cm^{-1}. All PC groups are clearly separated along PC-1, showing variable content of β-sheet secondary structure in nuclear proteins, the highest in cPDAC and the lowest in AVAC (amide I peak at 1628 cm^{-1}). Spectra of cPDAC and IPMC nuclei are located on the negative side of PC-1, whereas almost all spectra acquired from AVAC nuclei are located peripherally at the positive side of PC-1. The corresponding maxima of PC-1 loadings are related to protein bands including the CH_2 and CH_3 bending, the amide bands, and the Phe ring breathing indicating a relatively high content of overall proteins in AVAC nuclei; however, combined with the above negative side of PC-1 interpretation, these proteins are less rich in β-sheet secondary structure than proteins in AVAV and cPDAC.

The principal component 2 (PC-2) explains 44% of the total variance within the considered dataset. The spectra acquired from IPMC nuclei are located at the negative side of PC-2, in contrast to those from cPDAC and AVAC nuclei, which are shifted towards the positive values of PC-2. The PC-2 loadings are negatively correlated with the already mentioned protein bands from the Phe, the amide III, and the CH_2 and CH_3 bending, indicating similarly to PC-1, the relatively high protein content in IPMC nuclei in comparison to cPDAC and AVAC. Moreover, a strong minimum of PC-2 is observed in the amide I spectral range at 1650 cm^{-1}, showing the relatively high content of turns and unstructured coils secondary structures in nuclear proteins of IPMC.

The principal component 3 (PC-3), which explains 11% of the PCA total variance within the dataset, exhibits maxima related to the DNA backbone. These include 879 cm^{-1}, 1070 cm^{-1}, and 1128 cm^{-1}, as well as from the methyl and methylene motions at 1489 cm^{-1}, 1280 cm^{-1}, and 1391 cm^{-1}, indicating high methylation of DNA and/or histones. These bands are characteristic of spectra shifted towards positive values of PC-3, precisely a part of the spectra acquired from AVACs' nuclei, revealing the local character of high methylation.

To summarize, PCA performed on spectra from the nuclei of PC cells differentiated AVAC, IPMC, and cPDAC. Although all PC types were rich in nuclear proteins, we found variable content of protein secondary structures among them. Specifically, β-sheet-rich proteins prevail in cPDAC, whereas turns and unstructured coils proteins are characteristic of IPMC nuclei. The high protein content in nuclei volume is considered a hallmark of the ongoing process of DNA repair [44]. Moreover, locally, in AVAC, a high DNA and/or histone methylation level was found; however, it was variable among PC types.

The variable proteins' secondary structure composition differentiates the cytoplasmic region of PC types determined by PCA of CNN-classified Raman spectra

Spectra acquired from the cytoplasm of IPMC, AVAC, and cPDAC exhibit grouping in the 3D space of PCA (Fig. 3B); however, a more discrete level of separation is observed, compared to the PCA of nuclear spectra. The separation along PC-1 (42% of the PCA total variance) is driven mainly by the minima related to protein bands including CH_2, CH_3 bending, the amide III, and the Phe ring breathing (Fig. 3D). Figure 3 B and C highlight these bands to be characteristic of IPMC, in contrast to cPDAC spectra, and thus, indicate a high content of cytoplasmatic proteins and peptides in cancerous cells of this type of PC tumors. Spectra of IPMC cells' cytoplasm are shifted towards negative values of PC-1, and the characteristic minimum of PC-1 loading at 1662 cm^{-1} indicates a high contribution of turns and unstructured coils secondary structure in the proteins of IPMC cytoplasm.

There is only a slight separation of spectra acquired from the three investigated types of cancer along PC-2; however, the scores of IPMC are split into two groups, from which one is significantly shifted towards positive values of PC-2. The corresponding loading plot is dominated by maxima from proteins, including amide II at 1540 cm^{-1}, the C–C, C-N, and C-O stretching modes in the spectral range from 1160 to 1050 cm^{-1}, and C-H bending in tyrosine (Tyr) at 1170 cm^{-1} [45–48], suggesting local high content of Tyr-rich proteins in the cytoplasm of IPMC tumors.

PC-3 explains 8% of the PCA total variance and separates cytoplasm spectra of cPDAC (shifted towards lower PC-3 values) from AVAC and IPMC (PC-3 right-shifted). The PC-3 loadings are positively correlated with bands from the methyl and methylene motions from proteins and peptides at 1448 cm^{-1}, 1299 cm^{-1}, and 1378 cm^{-1}, suggesting

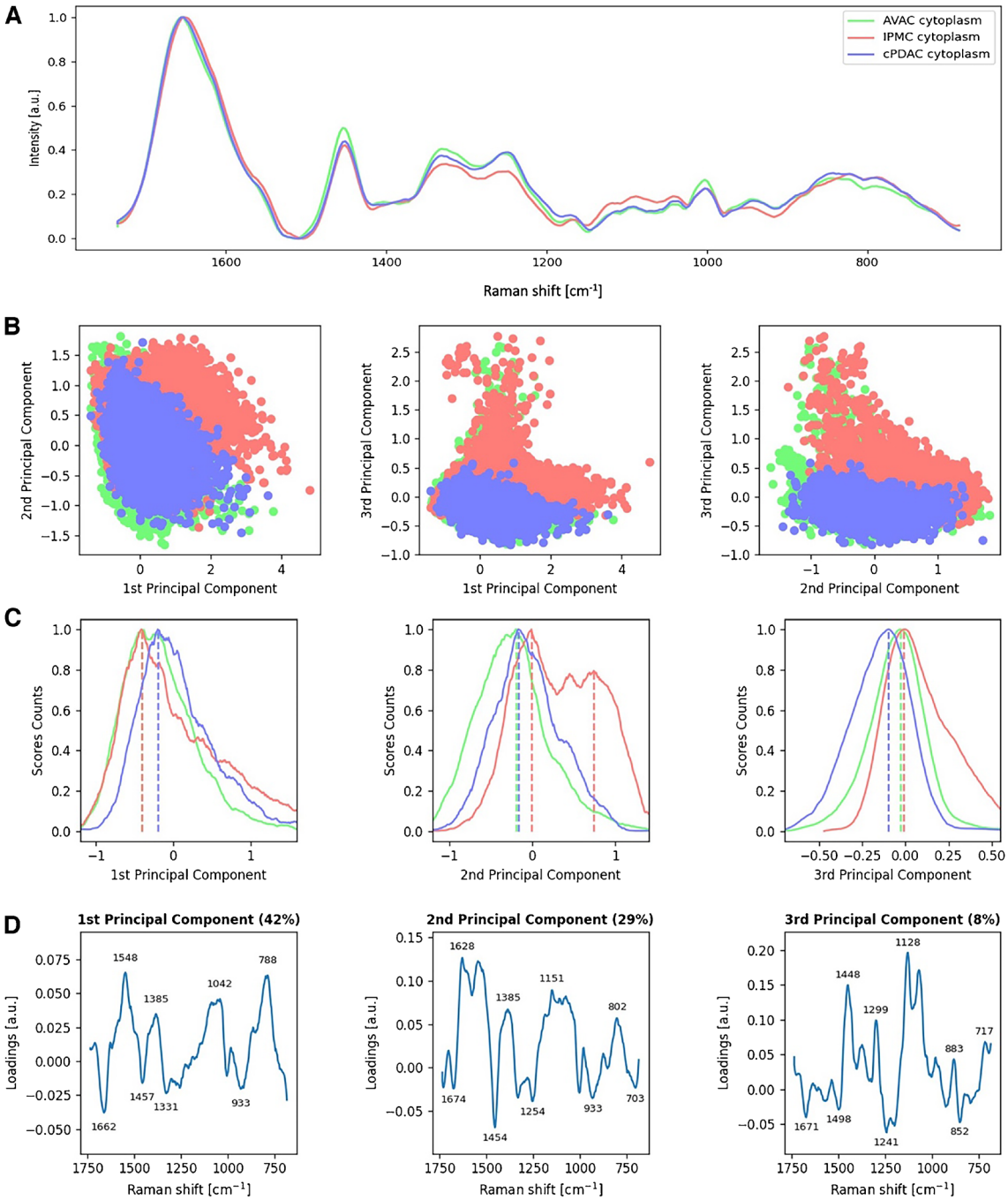

Fig. 3 Molecular characteristics of the cytoplasm of PC tumors. Raman spectra from the cytoplasm of all cells obtained from AVAC, cPDAC, and IPMC samples are presented as **A** linearly plotted mean spectra, **B** PCA 2D scores plots, **C** the PCA scores count plot showing the separation of their centroids, and **D** corresponding loadings plots with significant peaks marked

high content of methylated amino acids in the cytoplasm of AVAC and IPMC. Moreover, a strong positive correlation is also visible at 1128 cm^{-1}, which was found to be a spectroscopic marker of reduced cytochrome C in mitochondria [49]. The negative correlation of PC-3 loadings, which is characteristic of spectra acquired from cytoplasmic areas of cPDAC, is observed for the band from the C-C_6H_5 stretching in Phe and tryptophan (Trp) at 1208 cm^{-1} indicating the presence of Phe- and Trp-rich proteins in cPDAC cells cytoplasm.

To summarize, the exploration of cytoplasmic parts of PC cells revealed a high content of turns and unstructured

coil proteins in the IPMC, and what is more high content of Tyr-rich proteins was observed locally in the PMC cytoplasm. AVAC and some IPMCs were filled with methylated proteins, and a high content of reduced cytochrome C was detected. On the other hand, cPDAC cytoplasm included proteins rich in Phe and Trp.

Discussion

It was shown that CNN efficiently differentiates between the main types of PC [24]. Currently, we deepen the insight into the PC's molecular composition by extracting subcellular regions of cancerous tissues, allowing the distinction not only between AVAC, cPDAC, and IPMC but the nuclei and cytoplasm of their cells. This approach revealed new possibilities for investigating molecular characteristics of PC cells very precisely with considerable certainty of results interpretation. VS methods are great tools for molecular studies of cancerous tissues; however, the standard approach of random blind spot measurements [41] or rare grid mapping [42] might be prone to mixing the results with tumor necrosis and fibrosis, areas of inflammation, or benign pancreatic tissue. This leads to false results or misinterpretation. The VS methodology's success depends on the cognitive selection of areas of measurement. The CNN-driven selection of target spectra is a very efficient, unsupervised, automatic, and high-throughput approach for errorless investigations into the nature of PC. The distinction between nuclear and cytoplasmic regions of cancer cells allows for separate interpretations of each, and thus, adds to the knowledge of the molecular nature of PC.

Various models of neural networks including CNNs were introduced to solve classification problems in medical science. Widely reported algorithms, which were successful in image recognition in pathology [50] or radiology [51–53], proteomics classification [54], or spectral data recognition for cancer diagnostics [19, 55–59], focus largely on classification only. Although CNNs are generally known to be efficient in predicting (true problem solvers), the mechanisms of that verdict are somehow hidden, making them unsuitable for studying the nature of the problem. Although CNNs are great at reading the slightest differences in spectral analysis, due to the "fingerprint" character of VS results, in the path of CNN prediction that information is lost, and only the classification result remains. The findings of our study show differently. We go beyond CNN-based classification, and we propose using CNN for spectral data biomolecular interpretation. Moreover, our approach is automatic, time-efficient, and might be unsupervised. We combined a very accurate CNN classifier with PCA which reduced the dimensionality of CNN-predicted spectra revealing significant differences in spectral bands characteristic of cellular nuclei and cytoplasm of main PC tumors, specifically AVAC, cPDAC, and IPMC. Both methods (CNN and PCA) utilize automatic feature extraction; however, using PCA only would not recognize the spectra of each class (in subcellular scale) with such specificity and accuracy as the CNN. Moreover, CNN has a great feature of generalization, meaning that it handles relatively well data that it has not been trained on before. On the other hand, CNN only would not highlight spectral variabilities between classes, making it useless in terms of molecular interpretation. Both methods supplement each other and together become a universal tool for molecular studies of tissue samples by VS (such as RHM).

Here, in an automatic manner, we identified cellular nuclei and cytoplasm regions of AVAC's, cPDAC's, and IPMC's cells with a custom-designed CNN model. The annotation process for the CNN training required only less than 22% of the spectra to be marked, leaving the rest for the CNN generalization. First, we trained the CNN model by feeding it with raw spectral data. The achieved training and validation accuracy was very good (approximately 97%); however, that model did not perform well in predicting all seven classes when validated on new samples (new RHM map). To accentuate the meaningful information, spectra preprocessing were required. We removed the noise, baseline-corrected, and limited the range of analysis to that of biological samples (1800–650 cm^{-1}). Not surprisingly, the CNN model trained well but relatively early it started to overfit (the validation accuracy would not improve above 85%). The data augmentation was performed, by adding random noise to the preprocessed spectra. Trained on this new dataset, our CNN model reached 92% in a successful validation. To further elevate the correct prediction score, we applied another neural network (NN) model — a shallow dense classifier. Some specific spectroscopic landmarks were recently established for PC subtypes, including global DNA methylation ratio or proteins secondary structure ratios [24]. These findings were included in generating the extended training dataset for the new NN, which involved combining the initial spectral CNN-based prediction class with calculated DNA methylation, proteins' β-sheet, and proteins' random coil secondary structure ratios. By using this, combined automatic (by CNN) and manual (by dense classifier) feature extractor approach, the final classification reached over 98%, and more importantly, it performed efficiently with the new dataset (new PC sample).

For the CNN analysis of data, it is often beneficial to take advantage of the CNN's ability to extract spatial features. Although this type of analysis (called CNN-based image recognition) was largely applicated in medical sciences in radiology [60] or pathology [61], not many studies evaluated such an approach to spectral data analysis. When

using RHM, it might be possible; however, we intentionally resigned from image recognition analysis, due to a couple of reasons. First of all, in our study, we aimed to recognize the molecular characteristics of PC tissues, for further direct translation of the results into other VS methodologies eligible for serum-based liquid biopsy testing. Among methods recognized as the most promising candidates in the diagnostics of malignancies including PC, the ATR-FTIR is leading the board [3, 59]. In this particular VS method, as a result of the measurement, a single infrared spectrum is generated. Thus, any spatial recognition, in that case, is not an option. Another reason why we resigned from studying spatial features in RHM map images was that it is hardly possible with good accuracy. To annotate the data for image recognition training of the CNN, one has to precisely select areas of a certain class and select them entirely. Although we could use for comparison the unstained tissue light microscope pictures, normally taken before the Raman measurements, with unstained and deparaffinized tissue, precise recognition of all cellular components (i.e., the nuclei and the cytoplasm) is impossible, especially when dealing with such complicated samples as PC tissues. The results of imprecise image recognition training might be a falsification. Last, but not least, the doubtful benefit of utilizing image recognition in our CNN model for the study results would be recognizable. Our approach of only spectral data CNN-based analysis was successful in 98% of spectra, the results of our studies can be directly translated into the development of PC diagnostic technology based on ATR-FTIR serum measurements, and we showed that the training of the CNN model with RHM data can be relatively easy, time-efficient, automatic, and have high throughput, thus showing the path for its implementation into studying of malignancies other than PC.

Subsequently, after successful classification, we proceeded to the characterization of each of the predicted classes, which included regions of nuclei and cytoplasm of AVAC, cPDAC, and IPMC. The automatic extraction of predicted spectra and the PCA of various pairwise and triple-wise combinations of classes were conducted, uncovering the differentiating spectral bands among them.

Analyses of both nuclear and cytoplasmic cellular regions of PC cells differentiated these main PC groups and allowed biochemical interpretations characteristic of each of them. Specifically, all PC samples were rich in nuclear proteins. This was an expected finding since a high nuclear protein content is considered a hallmark of the ongoing process of DNA repair [44]. Similarly, as expected among nuclear proteins, we found variabilities in their dominating secondary structure. Recently, the highest general content of β-sheet-rich proteins was found in the samples of AVAC [24]; however, in the current study, we found the nuclei of cPDAC to be dominated by β-sheet-rich proteins and IPMCs' by turns and unstructured coils. To explain these discrepancies, it is worth noting that the NMF protein compound used for the analysis by the authors of the aforementioned study [24] represented mainly cytoplasmic and extracellular matrix proteins, while we investigated spectra selected from the nuclei of the PC cells, thus not contaminated with other cellular and extracellular regions of the studied tissue samples.

Both secondary structures we found (β-sheet, turns, and random coils) are domains of various proteins involved in carcinogenesis. In cellular nuclei, histones represent a substantial protein content, involved in DNA packaging or gene expression regulation [62]. Nevertheless, histones are primarily known for their α-helical structure. On the other hand, it was found that the β-sheet secondary structure promoted abnormal protein aggregation [63], a hallmark of cancer initiation and progression [64]. A great example of such protein is p53, which creates β-sheet aggregates in cellular nuclei of multiple cancers, including PC [65]. Another insulin-like growth factor-binding protein 2 (IGFBP2) was found to be associated with worse PC patients' prognosis, by inducing the nuclear translocation and phosphorylation of the p65 subunit of nuclear factor kappa-light-chain enhancer of activated B cells (NF-κB) [66]. The latter (NF-κB) is a nuclear protein complex containing subunits rich in turns and random coils that are important for interacting with other proteins involved in transcriptional regulation [67]. Another random coil-rich protein is breast cancer type 1 susceptibility protein (BRCA1), which is critically involved in the homology-directed repair pathway (HDR) of double-strand breaks (DSB) of DNA [68] and was found to play an important role in PC tumorigenesis [69].

Aside from the protein structural variations, in the studies of nuclear spectra, we found high methylation rates of DNA and histones, most prevailing in AVAC and IPMC samples, which is in line with recent findings of DNA methylation among PC types [24]. DNA methylation and histone methylation are major players in epigenetic modification important in PC growth and progression [70]. The importance of a deeper understanding of these alterations is well presented by DNA methylation profiling-based classification of the central nervous system tumors [43]. The knowledge of epigenetic variabilities among PC types shows promise for the development of new targeted therapies [71].

The analyses of PC cell cytoplasmic regions revealed a high Tyr content in IPMC samples, whereas the proteins of cPDACs cytoplasm were abundant in Phe and Trp. Examples of Tyr-rich cytoplasmic proteins involved in PC development and progression are Src kinase and focal adhesion kinase (FAK). Both interact with each other and combined with other proteins' deregulation; they drive tumor–stroma crosstalk and promote PC survival, adhesion, migration, and invasion [72]. Another Tyr-rich protein, STAT3, is constitutively activated in PC by the phosphorylation of Tyr705, leading to PC tumor progression at multiple stages of tumorigenesis,

starting with the Pdx1 transcription factor-driven initiation of acinar-to-ductal metaplasia [73] (which is believed to be a precursor lesion of PC [1]). Yet, another Tyr-rich protein is phosphoinositide 3-kinase (PI3K), an important part of the PI3K/AKT/mTOR signaling pathway, which plays a key role in regulating PC cell growth and survival, and importantly might be targeted with PI3K inhibitors [74]. Additionally, the cytoplasmic proteins of some of the studied IPMC samples contained methyl and methylene-rich proteins. Hypermethylation of cytoplasmic proteins, such as the Ras association domain family 1A (RASSF1A) promoter, was found in 64% of PCs in the study by Dammann et al. [75]. This epigenetic event downregulates RASSF1A and lowers its action, such as stimulation of mitotic arrest, DNA repair, and apoptosis, thus inhibiting tumor suppressor activity [75].

An important finding was made in AVAC and IPMC samples, in which we detected a high content of reduced mitochondrial cytochrome C. The tyrosine phosphorylation of Tyr48 in mitochondrial cytochrome C puts it in the redox state, thus preventing cancer cell apoptosis [76] and driving cancer progression [77].

Conclusions

The goal of our study was to extend the knowledge about the molecular nature of pancreatic tumors and find the best way of gaining it. These cancers are burdened with very high mortality, despite nearly daily reports adding to their understanding, multiple ongoing new therapeutic trials, or introduction of early diagnostic attempts. In serum-based liquid biopsy testing, as some authors [3], we perceive a solution to elevating the survival rates of PC patients; however, the conventional LBMs have not been introduced into clinical practice, because of the complexity of the methodology, and their not satisfying specificity, although comparable with other single-marker techniques [6]. Blood serum testing by VS combined with neural networking (including CNN) was established to be efficient in differentiating patients with and without malignancies [19, 55–59]. Nevertheless, for successful diagnostic technology, the ability to discriminate various malignancies in a single serum test is crucial; otherwise, one cannot be sure if the detected malignancy is truly the PC or maybe a gastric or colon cancer, making the test not useful in early screening for PC. A comparative study of various malignancies is required. Recently, in [24], the authors showed significant spectral variations among differentiated main groups of PC regarding global DNA methylation and the secondary structure of proteins. The results of their analysis, as well as ours, focusing on further interpretation of subcellular components of PC, can be directly translated into the development of PC diagnostic technology, which will be specific to PC (not other malignancies). However, this could not be achieved without our initial step taken in this study of characterizing the PC tissues first. It is due to the cognitive visualization by RHM and CNN prediction plotting, followed by PCA of each class that the true (not assumed) spectral interpretation was possible. A natural next step in establishing PC early screening technology is implementing our results in liquid biopsy testing. Early screening of PC patients is crucial for lowering the percentage of PC treatment failure, as it leads to early-stage tumors, making them suitable for radical resection and longer survival of patients [2].

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s00216-023-04997-w.

Author contribution Conceptualization: KSz, KSN, and EL; CNN design and programming: KSz; pathologic evaluation and POI selection: KSz, MK, FW, and MUB; Raman imaging: EL, KSo, NW, and MC; multivariate data analysis: KSz, EL, KSN, KSo, and MC; writing, original draft preparation: KSz, EL, and KSo; writing, review and editing: KSz, EL, KSo, KSN, MUB, and DA; funding acquisition: EL; supervision: DA and MUB. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by the National Science Centre, Poland, under the "OPUS 19" project (Reg. No. UMO-2020/37/B/ST4/02990).

Data availability All data obtained during the studies are available from the corresponding authors upon reasonable request. The source codes for the CNN training, CNN-based prediction with map plotting, as well as PCA and other plotting, are available in a public git repository under: https://bitbucket.org/howkuen/avac_cpdac_ipmc_cnn/src/master/

Declarations

Ethics approval The study was conducted following recognized ethical guidelines (Declaration of Helsinki). The project protocol was approved by the Jagiellonian University Bioethics Committee (opinion no. 1072.6120.77.2021).

Source of biological material Tissue samples involved in the study were selected from the Cracow University Hospital's Pathomorphology Department's archive, normally stored as conventional formalin-fixed paraffin-embedded (FFPE) blocks.

Conflict of interest The authors declare no competing interests.

References

1. Szymoński K, Milian-Ciesielska K, Lipiec E, Adamek D. Current pathology model of pancreatic cancer. Cancers (Basel). 2022;14:2321. https://doi.org/10.3390/cancers14092321.
2. Rawla P, Sunkara T, Gaduputi V. Epidemiology of pancreatic cancer: global trends, etiology and risk factors. World J Oncol. 2019;10:10–27. https://doi.org/10.14740/wjon1166.
3. Szymoński K, Chmura Ł, Lipiec E, Adamek D. Vibrational spectroscopy – are we close to finding a solution for early pancreatic cancer diagnosis? World J Gastroenterol. 2023;29:96–109. https://doi.org/10.3748/wjg.v29.i1.96.
4. Takaori K, Bassi C, Biankin A, Brunner TB, Cataldo I, Campbell F, Cunningham D, Falconi M, Frampton AE, Furuse J, Giovannini M, Jackson R, Nakamura A, Nealon W, Neoptolemos JP, Real FX, Scarpa A, Sclafani F, Windsor JA, Yamaguchi K, Wolfgang C, Johnson CD. International Association of Pancreatology (IAP)/ European Pancreatic Club (EPC) consensus review of guidelines for the treatment of pancreatic cancer. Pancreatology. 2016;16:14–27. https://doi.org/10.1016/j.pan.2015.10.013.
5. Gui J-C, Yan W-L, Liu X-D. CA19-9 and CA242 as tumor markers for the diagnosis of pancreatic cancer: a meta-analysis. Clin Exp Med. 2014;14:225–33. https://doi.org/10.1007/s10238-013-0234-9.
6. Wu J, Hu S, Zhang L, Xin J, Sun C, Wang L, Ding K, Wang B. Tumor circulome in the liquid biopsies for cancer diagnosis and prognosis. Theranostics. 2020;10:4544–56. https://doi.org/10.7150/thno.40532.
7. Błogowski W, Deskur A, Budkowska M, Sałata D, Madej-Michniewicz A, Dąbkowski K, Dołęgowska B, Starzyńska T. Selected cytokines in patients with pancreatic cancer: a preliminary report. PLoS ONE. 2014;9: e97613. https://doi.org/10.1371/journal.pone.0097613.
8. Talar-Wojnarowska R, Gasiorowska A, Smolarz B, Romanowicz-Makowska H, Kulig A, Malecka-Panas E. Clinical significance of interleukin-6 (Il-6) gene polymorphism and Il-6 serum level in pancreatic adenocarcinoma and chronic pancreatitis. Dig Dis Sci. 2009;54:683–9. https://doi.org/10.1007/s10620-008-0390-z.
9. Mroczko B, Groblewska M, Gryko M, Kędra B, Szmitkowski M. Diagnostic usefulness of serum interleukin 6 (IL-6) and C-reactive protein (CRP) in the differentiation between pancreatic cancer and chronic pancreatitis. J Clin Lab Anal. 2010;24:256–61. https://doi.org/10.1002/jcla.20395.
10. Bressy C, Lac S, Nigri J, Leca J, Roques J, Lavaut M-N, Secq V, Guillaumond F, Bui T-T, Pietrasz D, Granjeaud S, Bachet J-B, Ouaissi M, Iovanna J, Vasseur S, Tomasini R. LIF drives neural remodeling in pancreatic cancer and offers a new candidate biomarker. Cancer Res. 2018;78:909–21. https://doi.org/10.1158/0008-5472.CAN-15-2790.
11. De Rubis G, Rajeev Krishnan S, Bebawy M. Liquid biopsies in cancer diagnosis, monitoring, and prognosis. Trends Pharmacol Sci. 2019;40:172–86. https://doi.org/10.1016/j.tips.2019.01.006.
12. Zhang Y, Mi X, Tan X, Xiang R. Recent progress on liquid biopsy analysis using surface-enhanced Raman spectroscopy. Theranostics. 2019;9:491–525. https://doi.org/10.7150/thno.29875.
13. Krafft C, Popp J. Micro-Raman spectroscopy in medicine. Physical Sciences Reviews. 2019;4(10):20170047. https://doi.org/10.1515/psr-2017-0047.
14. Krafft C, Sergo V. Biomedical applications of Raman and infrared spectroscopy to diagnose tissues. Spectroscopy. 2006;20:195–218. https://doi.org/10.1155/2006/738186.
15. Diem M, Miljković M, Bird B, Chernenko T, Schubert J, Marcsisin E, Mazur A, Kingston E, Zuser E, Papamarkakis K, Laver N. Applications of infrared and Raman microspectroscopy of cells and tissue in medical diagnostics: present status and future promises. Spectroscopy (New York). 2012;27:463–96. https://doi.org/10.1155/2012/848360.
16. Grzelak MM, Wróbel PM, Lankosz M, Stęgowski Z, Chmura Ł, Adamek D, Hesse B, Castillo-Michel H. Diagnosis of ovarian tumour tissues by SR-FTIR spectroscopy: a pilot study. Spectrochim Acta A Mol Biomol Spectrosc. 2018;203:48–55. https://doi.org/10.1016/j.saa.2018.05.070.
17. Koster HJ, Guillen-Perez A, Gomez-Diaz JS, Navas-Moreno M, Birkeland AC, Carney RP. Fused Raman spectroscopic analysis of blood and saliva delivers high accuracy for head and neck cancer diagnostics. Sci Rep. 2022;12:18464. https://doi.org/10.1038/s41598-022-22197-x.
18. Cameron JM, Sala A, Antoniou G, Brennan PM, Butler HJ, Conn JJA, Connal S, Curran T, Hegarty MG, McHardy RG, Orringer D, Palmer DS, Smith BR, Baker MJ. A spectroscopic liquid biopsy for the earlier detection of multiple cancer types. Br J Cancer. 2023. https://doi.org/10.1038/s41416-023-02423-7.
19. Sala A, Cameron JM, Jenkins CA, Barr H, Christie L, Conn JJA, Evans TRJ, Harris DA, Palmer DS, Rinaldi C, Theakstone AG, Baker MJ. Liquid biopsy for pancreatic cancer detection using infrared spectroscopy. Cancers (Basel). 2022;14:3048. https://doi.org/10.3390/cancers14133048.
20. Fischer H-P, Zhou H. Pathogenesis of carcinoma of the papilla of Vater. J Hepatobiliary Pancreat Surg. 2004;11:301–9. https://doi.org/10.1007/s00534-004-0898-3.
21. Chandrasegaram MD, Gill AJ, Samra J, Price T, Chen J, Fawcett J, Merrett ND. Ampullary cancer of intestinal origin and duodenal cancer - a logical clinical and therapeutic subgroup in periampullary cancer. World J Gastrointest Oncol. 2017;9:407–15. https://doi.org/10.4251/wjgo.v9.i10.407.
22. Ferchichi M, Jouini R, Koubaa W, Khanchel F, Helal I, Hadad D, Bibani N, Chadli-Debbiche A, BenBrahim E. Ampullary and pancreatic adenocarcinoma—a comparative study. J Gastrointest Oncol. 2019;10:270–5. https://doi.org/10.21037/jgo.2018.09.09.
23. Reid MD, Balci S, Ohike N, Xue Y, Kim GE, Tajiri T, Memis B, Coban I, Dolgun A, Krasinskas AM, Basturk O, Kooby DA, Sarmiento JM, Maithel SK, El-Rayes BF, Adsay V. Ampullary carcinoma is often of mixed or hybrid histologic type: an analysis of reproducibility and clinical relevance of classification as pancreatobiliary versus intestinal in 232 cases. Mod Pathol. 2016;29:1575–85. https://doi.org/10.1038/modpathol.2016.124.
24. Szymoński K, Lipiec E, Sofińska K, Skirlińska-Nosek K, Czaja M, Seweryn S, Wilkosz N, Birarda G, Piccirilli F, Vaccari L, Chmura Ł, Szpor J, Adamek D, Szymoński M. Variabilities in global DNA methylation and β-sheet richness establish spectroscopic landscapes among subtypes of pancreatic cancer. Eur J Nucl Med Mol Imaging. 2023. https://doi.org/10.1007/s00259-023-06121-7.
25. Szymoński K, Lipiec E, Sofińska K, Skirlińska-Nosek K, Milian-Ciesielska K, Szpor J, Czaja M, Seweryn S, Wilkosz N, Birarda G, Piccirilli F, Vaccari L, Szymoński M. Spectroscopic screening of pancreatic cancer. Clinical Spectroscopy. 2021;3: 100016. https://doi.org/10.1016/j.clispe.2021.100016.
26. Luo X, Xing Y, Galvan DD, Zheng E, Wu P, Cai C, Yu Q. Plasmonic gold nanohole array for surface-enhanced Raman scattering detection of DNA methylation. ACS Sens. 2019;4:1534–42. https://doi.org/10.1021/acssensors.9b00008.
27. Li L, Lim SF, Puretzky A, Riehn R, Hallen HD. DNA methylation detection using resonance and nanobowtie-antenna-enhanced Raman spectroscopy. Biophys J. 2018;114:2498–506. https://doi.org/10.1016/j.bpj.2018.04.021.
28. Morla-Folch J, Xie H, Gisbert-Quilis P, Pedro SG, Pazos-Perez N, Alvarez-Puebla RA, Guerrini L. Ultrasensitive direct quantification of nucleobase modifications in DNA by surface-enhanced Raman scattering: the case of cytosine. Angew Chem Int Ed. 2015;54:13650–4. https://doi.org/10.1002/anie.201507682.

29. Guerrini L, Krpetić Ž, van Lierop D, Alvarez-Puebla RA, Graham D. Direct surface-enhanced Raman scattering analysis of DNA duplexes. Angew Chem Int Ed. 2015;54:1144–8. https://doi.org/10.1002/anie.201408558.
30. Barhoumi A, Halas NJ. Detecting chemically modified DNA bases using surface-enhanced Raman spectroscopy. J Phys Chem Lett. 2011;2:3118–23. https://doi.org/10.1021/jz201423b.
31. Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, Fielden PR, Fogarty SW, Fullwood NJ, Heys KA, Hughes C, Lasch P, Martin-Hirsch PL, Obinaju B, Sockalingum GD, Sulé-Suso J, Strong RJ, Walsh MJ, Wood BR, Gardner P, Martin FL. Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc. 2014;9:1771–91. https://doi.org/10.1038/nprot.2014.110.
32. Oleszko A, Hartwich J, Wójtowicz A, Gąsior-Głogowska M, Huras H, Komorowska M. Comparison of FTIR-ATR and Raman spectroscopy in determination of VLDL triglycerides in blood serum with PLS regression. Spectrochim Acta A Mol Biomol Spectrosc. 2017;183:239–46. https://doi.org/10.1016/j.saa.2017.04.020.
33. Lima C, Ahmed S, Xu Y, Muhamadali H, Parry C, McGalliard RJ, Carrol ED, Goodacre R. Simultaneous Raman and infrared spectroscopy: a novel combination for studying bacterial infections at the single cell level. Chem Sci. 2022;13:8171–9. https://doi.org/10.1039/D2SC02493D.
34. Owens GL, Gajjar K, Trevisan J, Fogarty SW, Taylor SE, Da Gama-Rose B, Martin-Hirsch PL, Martin FL. Vibrational biospectroscopy coupled with multivariate analysis extracts potentially diagnostic features in blood plasma/serum of ovarian cancer patients. J Biophotonics. 2014;7:200–9. https://doi.org/10.1002/jbio.201300157.
35. Engel J, Gerretzen J, Szymańska E, Jansen JJ, Downey G, Blanchet L, Buydens LMC. Breaking with trends in pre-processing? TrAC, Trends Anal Chem. 2013;50:96–106. https://doi.org/10.1016/j.trac.2013.04.015.
36. Liu J, Osadchy M, Ashton L, Foster M, Solomon CJ, Gibson SJ. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. Analyst. 2017;142:4067–74. https://doi.org/10.1039/C7AN01371J.
37. Kazemzadeh M, Hisey CL, Zargar-Shoshtari K, Xu W, Broderick NGR. Deep convolutional neural networks as a unified solution for Raman spectroscopy-based classification in biomedical applications. Opt Commun. 2022;510: 127977. https://doi.org/10.1016/j.optcom.2022.127977.
38. Chen X, Li J, Zhang Y, Lu Y, Liu S. Automatic feature extraction in X-ray image based on deep learning approach for determination of bone age. Futur Gener Comput Syst. 2020;110:795–801. https://doi.org/10.1016/j.future.2019.10.032.
39. Shaheen F, Verma B, Asafuddoula M. Impact of automatic feature extraction in deep learning architecture. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). Gold Coast, QLD, Australia. 2016, pp. 1–8. https://doi.org/10.1109/DICTA.2016.7797053.
40. Kong K, Kendall C, Stone N, Notingher I. Raman spectroscopy for medical diagnostics — from in-vitro biofluid assays to in-vivo cancer detection. Adv Drug Deliv Rev. 2015;89:121–34. https://doi.org/10.1016/j.addr.2015.03.009.
41. Li Z, Li Z, Chen Q, Ramos A, Zhang J, Boudreaux JP, Thiagarajan R, Bren-Mattison Y, Dunham ME, McWhorter AJ, Li X, Feng J-M, Li Y, Yao S, Xu J. Detection of pancreatic cancer by convolutional-neural-network-assisted spontaneous Raman spectroscopy with critical feature visualization. Neural Netw. 2021;144:455–64. https://doi.org/10.1016/j.neunet.2021.09.006.
42. Barroso EM, ten Hove I, Bakker Schut TC, Mast H, van Lanschot CGF, Smits RWH, Caspers PJ, Verdijk R, Noordhoek Hegt V, Baatenburg de Jong RJ, Wolvius EB, Puppels GJ, Koljenović S. Raman spectroscopy for assessment of bone resection margins in mandibulectomy for oral cavity squamous cell carcinoma. Eur J Cancer. 2018;92:77–87. https://doi.org/10.1016/j.ejca.2018.01.068.
43. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, Koelsche C, Sahm F, Chavez L, Reuss DE, Kratz A, Wefers AK, Huang K, Pajtler KW, Schweizer L, Stichel D, Olar A, Engel NW, Lindenberg K, Harter PN, Braczynski AK, Plate KH, Dohmen H, Garvalov BK, Coras R, Hölsken A, Hewer E, Bewerunge-Hudler M, Schick M, Fischer R, Beschorner R, Schittenhelm J, Staszewski O, Wani K, Varlet P, Pages M, Temming P, Lohmann D, Selt F, Witt H, Milde T, Witt O, Aronica E, Giangaspero F, Rushing E, Scheurlen W, Geisenberger C, Rodriguez FJ, Becker A, Preusser M, Haberler C, Bjerkvig R, Cryan J, Farrell M, Deckert M, Hench J, Frank S, Serrano J, Kannan K, Tsirigos A, Brück W, Hofer S, Brehmer S, Seiz-Rosenhagen M, Hänggi D, Hans V, Rozsnoki S, Hansford JR, Kohlhof P, Kristensen BW, Lechner M, Lopes B, Mawrin C, Ketter R, Kulozik A, Khatib Z, Heppner F, Koch A, Jouvet A, Keohane C, Mühleisen H, Mueller W, Pohl U, Prinz M, Benner A, Zapatka M, Gottardo NG, Driever PH, Kramm CM, Müller HL, Rutkowski S, von Hoff K, Frühwald MC, Gnekow A, Fleischhack G, Tippelt S, Calaminus G, Monoranu C-M, Perry A, Jones C, Jacques TS, Radlwimmer B, Gessi M, Pietsch T, Schramm J, Schackert G, Westphal M, Reifenberger G, Wesseling P, Weller M, Collins VP, Blümcke I, Bendszus M, Debus J, Huang A, Jabado N, Northcott PA, Paulus W, Gajjar A, Robinson GW, Taylor MD, Jaunmuktane Z, Ryzhova M, Platten M, Unterberg A, Wick W, Karajannis MA, Mittelbronn M, Acker T, Hartmann C, Aldape K, Schüller U, Buslei R, Lichter P, Kool M, Herold-Mende C, Ellison DW, Hasselblatt M, Snuderl M, Brandner S, Korshunov A, von Deimling A, Pfister SM. DNA methylation-based classification of central nervous system tumours. Nature. 2018;555:469–74. https://doi.org/10.1038/nature26000.
44. Czaja M, Skirlińska-Nosek K, Adamczyk O, Sofińska K, Wilkosz N, Rajfur Z, Szymoński M, Lipiec E. Raman research on bleomycin-induced DNA strand breaks and repair processes in living cells. Int J Mol Sci. 2022;23:3524. https://doi.org/10.3390/ijms23073524.
45. Huang C-Y, Balakrishnan G, Spiro TG. Early events in apomyoglobin unfolding probed by laser T-jump/UV resonance Raman spectroscopy. Biochemistry. 2005;44:15734–42. https://doi.org/10.1021/bi051578u.
46. Notingher I, Bisson I, Polak JM, Hench LL. In situ spectroscopic study of nucleic acids in differentiating embryonic stem cells. Vib Spectrosc. 2004;35:199–203. https://doi.org/10.1016/j.vibspec.2004.01.014.
47. Takai Y, Masuko T, Takeuchi H. Lipid structure of cytotoxic granules in living human killer T lymphocytes studied by Raman microspectroscopy. Biochimica et Biophysica Acta (BBA) General Subjects. 1997;1335:199–208. https://doi.org/10.1016/S0304-4165(96)00138-9.
48. Schneider FW. Frank S. Parker: Applications of infrared, Raman, and resonance Raman spectroscopy in biochemistry. Ber Bunsenges Phys Chem. 1984;88:1167B – 1168. https://doi.org/10.1002/bbpc.198400034.
49. Ishigaki M, Kashiwagi S, Wakabayashi S, Hoshino Y. In situ assessment of mitochondrial respiratory activity and lipid metabolism of mouse oocytes using resonance Raman spectroscopy. Analyst. 2021;146:7265–73. https://doi.org/10.1039/D1AN01106E.
50. Wang S, Chen A, Yang L, Cai L, Xie Y, Fujimoto J, Gazdar A, Xiao G. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. Sci Rep. 2018;8:10393. https://doi.org/10.1038/s41598-018-27707-4.

51. Fu H, Mi W, Pan B, Guo Y, Li J, Xu R, Zheng J, Zou C, Zhang T, Liang Z, Zou J, Zou H Automatic pancreatic ductal adenocarcinoma detection in whole slide images using deep convolutional neural networks. Front Oncol. 2021;11:665929. https://doi.org/10.3389/fonc.2021.665929.
52. Mahmoudi T, Kouzahkanan ZM, Radmard AR, Kafieh R, Salehnia A, Davarpanah AH, Arabalibeik H, Ahmadian A. Segmentation of pancreatic ductal adenocarcinoma (PDAC) and surrounding vessels in CT images using deep convolutional neural networks and texture descriptors. Sci Rep. 2022;12:3092. https://doi.org/10.1038/s41598-022-07111-9.
53. Ma H, Liu Z-X, Zhang J-J, Wu F-T, Xu C-F, Shen Z, Yu C-H, Li Y-M. Construction of a convolutional neural network classifier developed by computed tomography images for pancreatic cancer diagnosis. World J Gastroenterol. 2020;26:5156–68. https://doi.org/10.3748/wjg.v26.i34.5156.
54. Laxminarayanamma K, Krishnaiah RV, Sammulal P. Enhanced CNN model for pancreatic ductal adenocarcinoma classification based on proteomic data. Ingénierie des systèmes d information. 2022;27:127–33. https://doi.org/10.18280/isi.270115.
55. Yang X, Ou Q, Qian K, Yang J, Bai Z, Yang W, Shi Y, Liu G Diagnosis of lung cancer by ATR-FTIR spectroscopy and chemometrics. Front Oncol. 2021;11:753791. https://doi.org/10.3389/fonc.2021.753791.
56. Chatchawal P, Wongwattanakul M, Tippayawat P, Kochan K, Jearanaikoon N, Wood BR, Jearanaikoon P. Detection of human cholangiocarcinoma markers in serum using infrared spectroscopy. Cancers (Basel). 2021;13:5109. https://doi.org/10.3390/cancers13205109.
57. Gajjar K, Trevisan J, Owens G, Keating PJ, Wood NJ, Stringfellow HF, Martin-Hirsch PL, Martin FL. Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. Analyst. 2013;138:3917. https://doi.org/10.1039/c3an36654e.
58. Sitnikova VE, Kotkova MA, Nosenko TN, Kotkova TN, Martynova DM, Uspenskaya M, v. Breast cancer detection by ATR-FTIR spectroscopy of blood serum and multivariate data-analysis. Talanta. 2020;214: 120857. https://doi.org/10.1016/j.talanta.2020.120857.
59. Butler HJ, Brennan PM, Cameron JM, Finlayson D, Hegarty MG, Jenkinson MD, Palmer DS, Smith BR, Baker MJ. Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. Nat Commun. 2019;10:4501. https://doi.org/10.1038/s41467-019-12527-5.
60. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18:500–10. https://doi.org/10.1038/s41568-018-0016-5.
61. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest. 2021;101:412–22. https://doi.org/10.1038/s41374-020-00514-0.
62. Strahl BD, Allis CD. The language of covalent histone modifications. Nature. 2000;403:41–5. https://doi.org/10.1038/47412.
63. Ferrolino MC, Zhuravleva A, Budyak IL, Krishnan B, Gierasch LM. Delicate balance between functionally required flexibility and aggregation risk in a β-rich protein. Biochemistry. 2013;52:8843–54. https://doi.org/10.1021/bi4013462.
64. Costa DCF, de Oliveira GAP, Cino EA, Soares IN, Rangel LP, Silva JL. Aggregation and prion-like properties of misfolded tumor suppressors: is cancer a prion disease? Cold Spring Harb Perspect Biol. 2016;8: a023614. https://doi.org/10.1101/cshperspect.a023614.
65. Baslan T, Morris JP, Zhao Z, Reyes J, Ho Y-J, Tsanov KM, Bermeo J, Tian S, Zhang S, Askan G, Yavas A, Lecomte N, Erakky A, Varghese AM, Zhang A, Kendall J, Ghiban E, Chorbadjiev L, Wu J, Dimitrova N, Chadalavada K, Nanjangud GJ, Bandlamudi C, Gong Y, Donoghue MTA, Socci ND, Krasnitz A, Notta F, Leach SD, Iacobuzio-Donahue CA, Lowe SW. Ordered and deterministic cancer genome evolution after p53 loss. Nature. 2022;608:795–802. https://doi.org/10.1038/s41586-022-05082-5.
66. Gao S, Sun Y, Zhang X, Hu L, Liu Y, Chua CY, Phillips LM, Ren H, Fleming JB, Wang H, Chiao PJ, Hao J, Zhang W. IGFBP2 activates the NF-κB pathway to drive epithelial–mesenchymal transition and invasive character in pancreatic ductal adenocarcinoma. Cancer Res. 2016;76:6543–54. https://doi.org/10.1158/0008-5472.CAN-16-0438.
67. Schmitz ML, dos Santos Silva MA, Altmann H, Czisch M, Holak TA, Baeuerle PA. Structural and functional analysis of the NF-kappa B p65 C terminus. An acidic and modular transactivation domain with the potential to adopt an alpha-helical conformation. J Biol Chem. 1994;269:25613–20.
68. Chen C-C, Feng W, Lim PX, Kass EM, Jasin M. Homology-directed repair and the role of BRCA1, BRCA2, and related proteins in genome integrity and cancer. Annu Rev Cancer Biol. 2018;2:313–36. https://doi.org/10.1146/annurev-cancerbio-030617-050502.
69. Lai E, Ziranu P, Spanu D, Dubois M, Pretta A, Tolu S, Camera S, Liscia N, Mariani S, Persano M, Migliari M, Donisi C, Demurtas L, Pusceddu V, Puzzoni M, Scartozzi M. BRCA-mutant pancreatic ductal adenocarcinoma. Br J Cancer. 2021;125:1321–32. https://doi.org/10.1038/s41416-021-01469-9.
70. Liu X-Y, Guo C-H, Xi Z-Y, Xu X-Q, Zhao Q-Y, Li L-S, Wang Y. Histone methylation in pancreatic cancer and its clinical implications. World J Gastroenterol. 2021;27:6004–24. https://doi.org/10.3748/wjg.v27.i36.6004.
71. Hessmann E, Johnsen SA, Siveke JT, Ellenrieder V. Epigenetic treatment of pancreatic cancer: is there a therapeutic perspective on the horizon? Gut. 2017;66:168–79. https://doi.org/10.1136/gutjnl-2016-312539.
72. Parkin A, Man J, Timpson P, Pajic M. Targeting the complexity of Src signalling in the tumour microenvironment of pancreatic cancer: from mechanism to therapy. FEBS J. 2019;286:3510–39. https://doi.org/10.1111/febs.15011.
73. Corcoran RB, Contino G, Deshpande V, Tzatsos A, Conrad C, Benes CH, Levy DE, Settleman J, Engelman JA, Bardeesy N. STAT3 plays a critical role in KRAS-induced pancreatic tumorigenesis. Cancer Res. 2011;71:5020–9. https://doi.org/10.1158/0008-5472.CAN-11-0908.
74. Payne SN, Maher ME, Tran NH, Van De Hey DR, Foley TM, Yueh AE, Leystra AA, Pasch CA, Jeffrey JJ, Clipson L, Matkowskyj KA, Deming DA. PIK3CA mutations can initiate pancreatic tumorigenesis and are targetable with PI3K inhibitors. Oncogenesis. 2015;4:e169–e169. https://doi.org/10.1038/oncsis.2015.28.
75. Dammann R, Schagdarsurengin U, Liu L, Otto N, Gimm O, Dralle H, Boehm BO, Pfeifer GP, Hoang-Vu C. Frequent RASSF1A promoter hypermethylation and K-ras mutations in pancreatic carcinoma. Oncogene. 2003;22:3806–12. https://doi.org/10.1038/sj.onc.1206582.
76. García-Heredia JM, Díaz-Quintana A, Salzano M, Orzáez M, Pérez-Payá E, Teixeira M, De la Rosa MA, Díaz-Moreno I. Tyrosine phosphorylation turns alkaline transition into a biologically relevant process and makes human cytochrome c behave as an anti-apoptotic switch. J Biol Inorg Chem. 2011;16:1155–68. https://doi.org/10.1007/s00775-011-0804-9.
77. Abramczyk H, Brozek-Pluska B, Kopeć M. Double face of cytochrome c in cancers by Raman imaging. Sci Rep. 2022;12:2120. https://doi.org/10.1038/s41598-022-04803-0.