applied sciences

MDPI

Review

Raman Spectroscopy, Machine Learning and Pancreatic Tumors

Riccardo Del Gratta [1,*], Gianmarco Lazzini [1], Raffaele Gaeta [2], Annalisa Comandatore [3], Giuseppe Succurro [4], Luca Morelli [3] and Mario D'Acunto [1,*]

[1] CNR-IBF, National Research Council, Institute of Biophysics, 56124 Pisa, Italy; gianmarco.lazzini@ibf.cnr.it (G.L.);
[2] Second Division of Surgical Pathology, University Hospital of Pisa, 56124 Pisa, Italy; raffaele.gaeta@med.unipi.it
[3] General Surgery Unit, Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, 56126 Pisa, Italy
[4] CNR-IIT, National Research Council, Institute of Informatics and Telematics, 56124 Pisa, Italy
* Correspondences: riccardo.delgratta@ibf.cnr.it (R.D.G.) and mario.dacunto@ibf.cnr.it (M.D.)

Abstract

Raman spectroscopy provides label-free, molecularly specific measurements of tissue, cells, and biofluids. Chemometric, machine-learning, and deep-learning methods are used to interpret these measurements for diagnostic purposes. This review critically examines an index sample of 15 studies applying such methods to pancreatic samples (human tissue and biofluids, animal models, cell lines), focusing on preprocessing, classification, and validation design. Few studies are restricted to the pancreas, so the sample includes other tumor types; "pancreatic cancer" denotes the whole group, and pancreatic ductal adenocarcinoma is named only where a study reports it. Across these studies, high internal performance is often accompanied by small numbers of independent biological units, spectral pseudoreplication, unclear patient- or specimen-level partitioning, data-dependent processing outside training folds, and limited external validation; instrument and pipeline differences further prevent direct comparison. We distinguish evidence of spectral separability within a dataset from evidence of diagnostic generalization to independent patients; our conclusions concern validation design, not diagnostic accuracy for any tumor type. Priorities include transparent reporting, patient-grouped evaluation, training-only preprocessing and model selection, harmonized acquisition and processing, and shared benchmarks with sequestered test sets. Raman spectroscopy remains a promising source of biochemical information, but clinical translation requires reproducible performance across independent patients, laboratories, and acquisition conditions.

Keywords: Raman spectroscopy; machine learning; deep learning; pancreatic cancer; pancreatic ductal adenocarcinoma; pancreatitis; data leakage; validation

Check for updates

Received: 1 October 2026
Revised: 29 October 2026
Accepted: 30 October 2026
Published: 6 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Pancreatic cancer is among the leading causes of cancer death, and global deaths have risen 2.3-fold since 1990 [1]. It also has one of the poorest prognoses: the 5-year survival rate is approximately 13% in the United States [2]. Survival improved only modestly over the past four decades, but has risen faster in the last ten years, a gain attributed mainly to higher resection rates and improved perioperative safety [3]. Despite this progress, pancreatic cancer has the lowest 5-year survival rate of all major cancers [2].

Clinicians classify pancreatic cancer into two main groups: exocrine tumors and neuroendocrine tumors. Each group contains several subtypes that differ in symptoms

and prognosis [4,5]. Exocrine tumors account for roughly 95% of all pancreatic cancers. They originate from the exocrine cells of the gland whose ducts carry digestive enzymes to the duodenum. pancreatic ductal adenocarcinoma (PDAC) is the most common exocrine subtype and accounts for over 90% of diagnoses. It originates from the ductal epithelium [6]. Several mutations drive PDAC; the most common affects Kirsten rat sarcoma viral oncogene homolog (KRAS), which encodes K-Ras, a key protein in the rat sarcoma/mitogen-activated protein kinase (RAS/MAPK) signaling pathway, and is altered in more than 80% of cases [7].

Less common exocrine subtypes include squamous cell carcinoma and adenosquamous carcinoma; the latter occurs in 1–4% of exocrine cancers and combines features of both histologies. Colloid carcinoma (1–3%) develops from an intraductal papillary mucinous neoplasm (IPMN) [8]. Tumor cells float in a mucin-rich matrix, so they spread more slowly and respond better to treatment. Neuroendocrine tumors originate in the pancreatic islets. They represent under 2% of pancreatic cancer cases and are typically less aggressive than exocrine forms [9].

Raman spectroscopy (RS) is used for the molecular characterization of biomedical samples. When incident photons interact with matter, a small fraction is scattered inelastically, with a frequency shift relative to the excitation radiation. This Raman shift reflects the vibrational properties, and hence the chemical composition, of the molecules probed [10]. The frequency distribution of the photons scattered by a complex biological sample therefore yields a characteristic spectral profile, commonly called the sample's "molecular fingerprint" [11]. RS is generally non-destructive at appropriate laser powers and label-free [11]. A single spectrum covers a range that includes bands from many potential biomarkers, such as lipids, nucleic acids, and proteins. Several variants of RS, described in the following sections, also map the signal across a surface and so show the spatial distribution of specific classes of molecules, with a resolution that can reach the subcellular level. Numerous studies have proposed RS as a diagnostic approach in cancer research, although this line of work is still at an early stage [12–15].

The Raman signal is intrinsically weak, particularly when compared with other potentially concomitant contributions, such as autofluorescence [16]. In this regard, stimulated Raman scattering (SRS), a member of the broader family of coherent Raman scattering (CRS) techniques, represents a potential solution to this limitation. The basic principle underlying SRS consists of illuminating the sample with two laser pulses, typically with durations of a few picoseconds or less [17], which are spatially and temporally overlapped [18]. This interaction generates a Raman signal characterized by spatial and temporal coherence, resulting in an improved signal-to-noise ratio. Moreover, the technique enables video-rate acquisition, at the expense of a narrow instantaneous bandwidth, typically on the order of $\sim$10 cm^{-1} in single-frequency implementations [17]. Nevertheless, recent studies have demonstrated the potential of SRS as a diagnostic approach in several tumor types [19–23]. In particular, in the case of brain tumors, SRS combined with deep learning has reached intraoperative molecular classification accuracy exceeding 90% [22], and, more recently, an accuracy above 96% has been reported for pancreatic biopsies obtained via endoscopic ultrasound-guided fine-needle aspiration [23]. Coherent nonlinear optical microscopy combining SRS with second-harmonic generation (SHG) has shown high diagnostic concordance (reported Cohen's kappa, $\kappa > 0.97$) with hematoxylin and eosin (H&E) staining on frozen sections from pancreatic resection margins, and produced comparable image quality on fresh, unprocessed tissue without a quantified concordance score [24].

Although CRS is particularly effective for the rapid imaging of complex biological samples, the high cost of the required instrumentation limits its potential for engineering and integration into compact systems [25]. Moreover, it does not fully overcome the limited

detection sensitivity often associated with biomarkers of interest. On the other hand, surface-enhanced Raman spectroscopy (SERS) appears to be a Raman-based approach combining high efficacy with the possibility of being engineered into compact, point-of-care devices [26]. SERS exploits the enhancement of the electromagnetic field in the vicinity of metallic nanostructures, such as nanostructured surfaces or nanoparticles. By placing the analyte within these regions of enhanced electromagnetic field, it is therefore possible to induce Raman scattering with a substantially enhanced intensity [27]. Two complementary mechanisms are generally invoked to explain this enhancement [28]. In the electromagnetic enhancement (EE) model, incident radiation resonant with the collective oscillations of the conduction electrons excites a localized surface plasmon resonance in the metallic nanostructure, strongly amplifying the local electric field. This amplification is maximal at the so-called "hot spots", typically located at junctions, gaps, or aggregates of metallic nanoparticles [29], where the enhancement can be sufficient to enable single-molecule detection [30]. The chemical enhancement (CE) mechanism, by contrast, arises from changes in the molecular polarizability caused by the interaction between the analyte and the metal surface [31]. EE is generally regarded as the dominant contribution, with enhancement factors of 10^4–10^8, whereas CE rarely exceeds 10^1–10^3 [27,32]; CE is, however, molecule- and surface-specific and thus contributes to spectral selectivity [31].

All the aforementioned advantages make SERS suitable for the detection of cancer biomarkers at very low concentrations [33]. However, their quantification remains challenging because experimental factors interfere with standard detection. Banaei et al. recently tested a SERS-based immunoassay on serum extracellular vesicles (EVs) from 15 patients (5 controls, 5 chronic pancreatitis, 5 PDAC) and reported sensitivity and specificity above 0.95 in discriminating chronic pancreatitis from pancreatic cancer [34]. Traditional methods for isolating tumor-derived EVs are laborious and often destructive. SERS immunoassays avoid this problem by detecting EVs directly in serum. Valera et al. used SERS to study immunosuppression in stratified pancreatic cancer–stroma spheroids (STAM) and in monocultures of cancer cells or fibroblasts. In the co-culture spheroids, indoleamine 2,3-dioxygenase 1 (IDO1) converted tryptophan into kynurenine at much higher rates than in either monoculture, and the SERS signal tracked this difference accurately when validated against the enzyme-linked immunosorbent assay (ELISA) [35].

Although the high information content of individual Raman spectra represents a major strength of the approach, translating that richness into diagnostically useful information is not straightforward. A Raman spectrum is, numerically, a table of intensities across hundreds or thousands of wavenumbers: neighboring channels are highly correlated because the finite linewidth of each vibrational mode is sampled at a much finer digital resolution, many channels are dominated by noise, and the differences between specific tumor histotypes, healthy tissue, and pancreatitis are often too subtle to distinguish by eye. Ordinary statistics struggle with this geometry, so advanced computational strategies are required to suppress noisy or redundant contributions and extract the diagnostically relevant signal [13,36]. In this context, chemometric methods distinguish patterns associated with the classes of interest from irrelevant variation or noise [37]. As in machine learning (ML) more broadly, chemometric methods applied to Raman data fall into two classes: unsupervised models, which explore the data without labels, and supervised models, which learn from samples whose class is already known [38].

In supervised learning (SL), the entire spectrum is associated with one label describing a property of the sample as a whole, such as the tissue histotype (tumor, pancreatitis, healthy) from which it was acquired. This label may take continuous or discrete (categorical) values; the model is then termed a regressor or a classifier, respectively. Labels are used during training, validation, and testing to build a model able to predict them on new

spectra. Among the most common regression models is multiple linear regression (MLR), while widely used classifiers include partial least squares discriminant analysis (PLS-DA), support vector machine discriminant analysis (SVM-DA), and artificial neural networks (ANNs), among others [15].

Unsupervised learning (UL) comprises a class of ML methods designed to autonomously uncover latent structures, patterns, and relationships within data in the absence of predefined class labels. Rather than learning an explicit mapping between input features and target outputs, these models characterize the intrinsic structure and distribution of the data, thereby revealing underlying patterns and organization. Representative UL techniques include dimensionality-reduction methods, such as principal component analysis (PCA) and *t*-distributed stochastic neighbor embedding (*t*-SNE), which project high-dimensional data into lower-dimensional spaces while preserving relevant features of their underlying structure, as well as clustering algorithms, such as *k*-means and hierarchical clustering, which partition observations according to predefined measures of similarity. In spectroscopic applications, these approaches are particularly valuable at the exploratory stage of data analysis, as they can reveal latent substructures, sample heterogeneity, or outliers without requiring prior assumptions regarding the classes of interest, such as tumor histotypes.

Raman-ML studies on pancreatic cancer are comparatively few. In a 2022 review of artificial intelligence (AI) applications to pancreatic cancer, Huang et al. mentioned only one Raman study among dozens of papers on computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), endoscopic ultrasound, and biomarkers [39]. More broadly, Raman-ML applications have been reported more extensively in other tumor types, including breast, lung, and cervical cancer [40].

To address this landscape, the present review examines the application of ML techniques to Raman data from pancreatic samples, mapping what has been achieved so far and assessing how far the field is from clinical translation. Because studies restricted to the pancreas are few, the review analyzes an index sample of studies that differ in sample type and tumor type, described in Section 3, and not an exhaustive survey of the literature. "Pancreatic cancer" denotes the whole group, and PDAC is used only where a study reports it. The review therefore does not estimate the diagnostic accuracy of Raman spectroscopy for any tumor type; it assesses whether the validation design of each study supports the performance it reports. This work builds upon a previous broad survey [15], which cataloged Raman-AI applications across seven cancer types and included seven of the pancreatic studies revisited here. While the two reviews overlap, they follow distinct methodological frameworks, which are disclosed here for transparency. The analysis conducted in the first review explicitly identifies leave-one-patient-out cross-validation, small-sample overfitting, and validation-strategy selection as central concerns for the field, but the authors do not apply these criteria to evaluate the individual studies. The present review takes up that gap directly, applying a fixed taxonomy of ML failure modes (structural leakage, preprocessing leakage, feature-selection bias, model-selection bias, confounded class labels, and spectral pseudoreplication) to each study of the sample.

2. Machine Learning Methods for Raman Investigations of Pancreatic Cancers

ML is increasingly applied to Raman spectroscopy to classify biological samples and identify spectral features associated with pathological conditions [15,41]. In studies of pancreatic cancer, researchers have investigated several supervised learning approaches. For instance, Mandrell et al. and Ziwei et al. applied support vector machines (SVMs) and *k*-nearest neighbors (*k*-NN) to Raman spectra for the classification of pancreatic cancer-

related samples [42,43], while Li et al. used convolutional neural networks (CNNs) to analyze Raman spectra and identify features relevant to pancreatic cancer classification [44]. Feature-selection procedures have also been employed to reduce the number of spectral variables and improve the interpretability of classification models [42,45]. Dimensionality reduction and feature extraction methods, including PCA and partial least squares (PLS), are generally used to transform the original spectra into lower-dimensional representations before classification.

SVMs are among the supervised learning methods commonly used for the classification of Raman spectra in cancer research, with applications to different tumor types and biological samples [46–48]. In the binary classification setting, an SVM determines a decision hyperplane that maximizes the margin between the two classes [49]. When the classes are not linearly separable in the original feature space, kernel functions allow SVMs to construct non-linear decision boundaries by implicitly mapping the data into a higher-dimensional feature space.

Random Forests (RFs) are ensemble learning methods based on collections of decision trees [50]. Each tree is trained on a bootstrap sample drawn from the original training set, and a random subset of features is considered at each node when determining the split [50]. Let N_p denote the total number of features; a common choice for classification is to consider $\sqrt{N_p}$ randomly selected features at each split [51]. The feature and threshold defining the split are chosen according to the reduction in node impurity, often measured using the Gini index [51,52]. The final class is assigned by majority voting across the individual trees.

A recent approach combines topological data analysis (TDA) with ML to obtain a different representation of Raman spectra for classification [53]. Instead of using the spectral variables directly, TDA is used to extract descriptors based on the topological structure of the data, which are eventually provided to a conventional machine-learning classifier. In a study of pancreatic tissue, Conti et al. applied this approach to Raman spectra from biopsies of normal tissue, chronic pancreatitis, and PDAC [53].

3. Review Scope and Methods

3.1. Review Type

This is a critical review of the validation design of studies that apply machine learning to Raman spectra of pancreatic samples. It has a systematic component: a fixed taxonomy of failure modes is applied to every study, the coverage search is documented, and the eligibility criteria are stated explicitly (Sections 3.3 and 3.4). It was not, however, conducted as a systematic review in the sense of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [54]: no protocol was registered, screening was not duplicated, and the aim is to examine how validation can fail, not to estimate how well the methods perform. We therefore do not claim adherence to PRISMA and use its 2020 flow diagram only as a template to report the selection of records.

3.2. Index Sample

The review analyzes an index sample of 15 studies. It starts from the pancreatic studies of the review by Manganelli Conforti et al. [15], seven in total, to which the group added its own 2025 study [41]; these eight studies formed an earlier, unpublished draft of this review. The authors then extended the sample to the 15 studies analyzed here, choosing those closest to the research interests of the group. The index sample therefore reflects what the authors found and chose, and not every study that meets the eligibility criteria below. Studies on pancreatic samples that apply no classifier to the spectra, such as those cited in the Introduction as clinical or technical context [24,35], are not part of the index sample.

The sample covers human tissue and biofluids, animal models, and cell lines. Because studies restricted to the pancreas are few, it also contains studies of other tumor types and of pancreatic lesions other than ductal adenocarcinoma, provided that pancreatic samples were part of the data analyzed with a machine-learning classifier. Pancreas-specific performance is not always reported: it is absent in two multi-cancer serum studies [55,56], and can be reconstructed from a confusion matrix in only one of them. "Pancreatic cancer" denotes the whole group, and PDAC is used only where a study reports it.

3.3. Coverage Search

To make the process verifiable and to estimate what the index sample leaves out, we searched PubMed and Scopus. For each database, we ran a broad query combining Raman, SERS, or stimulated Raman terms with a pancreas term, and a narrower query that also required ML and classification terms. The strings are given in the Supplementary Materials, with no restriction on language or publication dateat retrieval. The two sets were merged and deduplicated with a Python 3.12 script that matches the records first by DOI and then, among those not matched, by normalized title together with the year; the script is given in Section "Deduplication of PubMed and Scopus records" of Supplementary File S2. The procedure left 394 unique records, of which 164 were found in both databases, 3 in PubMed only and 227 in Scopus only. The 15 studies of the index sample had been identified before the search and independently of the queries. The broad queries retrieved all of them, and the narrower queries retrieved 14 of 15 (The exception is Carmicheal et al. [57]: the abstract names the classifier "principal component differential function analysis" (sic) and contains none of the machine-learning or classification terms of query A. The 19 October sets of query A could not be reproduced exactly (Supplementary File S2, Queries)). The check shows that the queries are not blind to the studies we had selected (The searches were run on 19 October 2026. The publication window (up to 30 June 2026) was applied at screening. Figure 1 describes the flow).

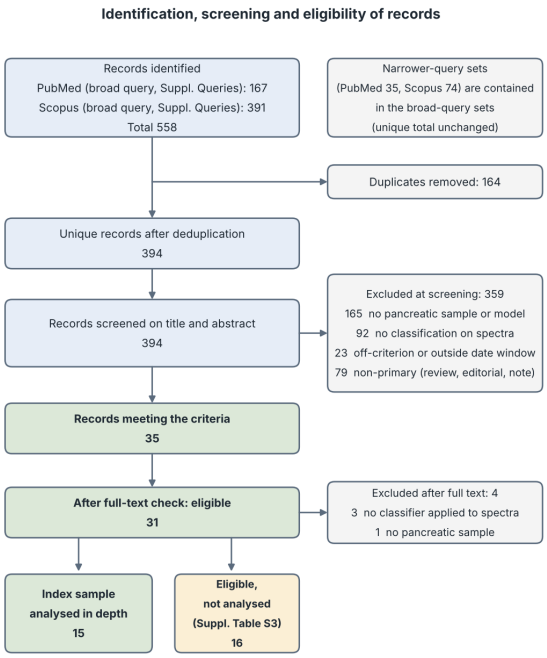


Figure 1. Identification, screening, and eligibility of records. The broad and narrower queries are reported in Supplementary File S2 (Queries). The authors selected the 15 studies in the index sample from the 31 eligible studies. The 16 eligible studies not analyzed were not excluded by any criterion and are listed in Table S3 of Supplementary File S1.

3.4. Eligibility and Screening

A record was eligible if it (i) reported primary research (journal article, conference paper, or book chapter), (ii) used Raman-based measurements, including SERS and stimulated Raman scattering, (iii) included pancreatic samples (human tissue or biofluid, animal model, or cell lines) among the data analyzed, and (iv) applied a statistical or machine-learning classifier to the spectra. Of the 394 unique records, 359 were excluded based on their titles and abstracts, and 35 met the criteria. Three were excluded after full-text review because no classifier was applied to the spectra; one lacked a pancreatic sample. Of 31 eligible studies, 15 form the index sample.

The index sample was read from the perspective of translational diagnostics, that is, of whether the reported validation would support a tool that a pathologist could use; this is the perspective of the review, not a criterion by which the studies were selected. Accordingly, the index sample contains more studies on human tissue directed to histopathology (four of 15 [23,41,58,59]) than the 16 studies not analyzed (one of 16 [60]).

The remaining 16 are described in Table S3 of Supplementary File S1; they were not excluded by any criterion, but were not selected for the analysis carried out in Section 4. We retrieved the full text for 14 of the 16 eligible records and used it to confirm eligibility. For the remaining two (Bunganič [61] and Chen [62]), the decision was based on the abstract alone.

Table S3 also allows a descriptive check, without appraising the studies, of whether the 16 would alter the weakness we found most often in the index sample, namely a partition of training and test data below the level of the patient or not stated (Supplementary File S1, Table S2). For 14 of the 16 the separation could be extracted. In six of them at least one analysis divided the data at the level of single spectra or readouts, or the unit of the split is not stated; in two, no held-out set was used and the model was evaluated on all samples, with leave-one-out cross-validation or permutation tests; in five, the data were divided by sample, by tissue block, or with an independent test set; in one, a model parameter was selected on the accuracy of the cross-validation also used for evaluation. The 16 studies therefore show the same weakness, but not universally. This description reports what the studies state; it is not an appraisal of their validity, and it does not cover weaknesses that Table S3 does not report, such as preprocessing or feature selection outside the training folds.

Titles and abstracts were screened by one author (R.D.G.), who read the full text where the abstract did not settle eligibility. Of the 35 records that met the criteria on title and abstract, 33 were read in full; the other two were assessed on the abstract (see above). The query strings combined the concepts of Raman spectroscopy (including SERS and stimulated Raman scattering) and pancreas, with the addition of machine learning in the narrower query; they were proposed by the AI assistant, and the author ran them in PubMed and Scopus and exported the records. To limit the risk of omitting relevant studies, the broad query combined only the concepts of Raman and pancreas, and two databases were searched with no restriction on language or publication date at retrieval. The AI assistant did not exclude any record on its own: it was used to organize the exports, to apply the deduplication rule (first by DOI, then by normalized title together with the year; Supplementary File S2, Section “Deduplication of PubMed and Scopus records”), and to propose an exclusion code for each record. The author took every inclusion and exclusion decision and checked the proposed codes on a sample. There was no independent duplicate screening. The screening log, with the exclusion codes X1 to X6, is provided as Supplementary File S2. To assess the risk of false exclusion, a random stratified sample of 30 of the 359 records excluded on title and abstract was re-assessed by R.D.G., blind to the assigned code. None of the 30 was found to be eligible, so the exclusion decision

was confirmed for all of them; at first reading, the assigned code matched in only 14 of the 30 records (stratified sample, not a population estimate). The disagreements arose because the exclusion criteria are not mutually exclusive (e.g., a review also lacks a classifier) and because no code was dedicated to the absence of Raman measurements. The codes in Supplementary File S2 should therefore be read as indicative of the main reason for exclusion, not as a unique classification. Because the re-assessor was also the screener, the check tests the consistency of the decisions, not their independence; with no false exclusion in 30 records, a rate of up to about 10% remains compatible with the data. The check does not cover studies that the queries did not retrieve or that the two databases do not index, and we did not search other databases or the reference lists of the retrieved studies.

3.5. Data Extraction and Appraisal

We read each of the 15 studies in full and extracted the sample type, the unit of analysis, the number of patients, animals, or preparations, as well as the number of spectra, the preprocessing and feature-selection steps, the model-selection and validation schemes, and the origin of the class labels. Each study was then assessed and described in Section 4. The extraction tables are provided in Supplementary File S1 (Tables S1 and S2), and a synopsis is presented in Table 1. The assessment describes potential failures in validation design; the review does not estimate how prevalent these failure modes are in the literature.

Table 1. Synopsis of the 15 Raman–ML studies reviewed in Section 4. "Units" are the independent biological or experimental units; the number of spectra follows the arrow. "Documented" leakage means that the paper itself describes a procedure that lets test information reach training; "unresolved" means that grouping or nesting is not reported, so leakage can be neither confirmed nor excluded. Headline figures are internal unless stated otherwise.

Study	Sample, Modality	Units → Spectra	Internal Validatio
Zhang et al., 2026 [55]	Human serum exosomes, SERS	655 samples (45 pancreatic) → ≥131,000	5-fold CV, groupin not stated
Lazzini et al., 2025 [41]	Human surgical tissue, confocal Raman	15 patients → 1140 per class after SMOTE	5-fold CV; SMOTI not stated
Tan et al., 2025 [23]	Human EUS-FNA biopsies, SRS imaging	76 patients → 2000 blocks per class	8:2 block-level spli grouping
Aslam et al., 2024 [45]	Mouse tissue, spontaneous Raman	20 mice → 2529	10% hold-out + 15 SVM-RFE nesting not shown
Kralova et al., 2024 [64]	Human plasma, Raman + ROA	191 samples (sample-level analysis)	LOOCV; band sel full cohort, nestin not stated
Szymoński et al., 2023 (EJNMMI) [59]	Human FFPE tissue, RHM + CNN	17 patients (24 slides) → 53,879	70:30 split at spect
Szymoński et al., 2023 (ABC) [58]	Human FFPE tissue, RHM + CNN	15 patients (16 slides) → 42,098 annotated, 126,294 after augmentation	70:30 split at spect augmentation bef
Sezer et al., 2022 [63]	Single cell line, confocal Raman	One cell line → 254	One 67:33 split; P on all spectra
Uthamacumaran et al., 2022 [56]	Human serum EVs, Raman + FTIR	9 subjects (1 pancreatic) → 18 Raman	50:50 split + 5-fol grouping unclear
Banaei et al., 2021 [34]	Human serum EVs, targeted SERS immunoassay	15 patients → 150 readouts	Random 5-fold CV on readouts

3.6. Limitations of the Approach

Several limitations apply. First, the index sample is not exhaustive and reflects the authors' choices, so conclusions about the field cannot be drawn from it. Second, the eligibility criteria were refined during the work rather than fixed beforehand. Third, one author screened titles and abstracts, and the coauthors checked the screening output; however, screening was not independently duplicated. The coauthors also validated each study appraisal, but did not conduct the appraisals in duplicate. Fourth, two authors coauthored one of the appraised studies [41]; an author who did not participate in that study conducted its appraisal.

4. Review of Raman Spectroscopy and Machine Learning Applications in Pancreatic Oncology

The 15 studies reviewed here are an index sample. How they were selected, and what the coverage search shows about the eligible studies that are not analyzed, is described in Section 3.

This section reviews Raman spectroscopy and machine-learning techniques for cancer diagnosis. Many studies combine pancreatic cancer with other malignancies, making it difficult to extract pancreas-specific accuracy figures.

Three measurement strategies appear across this literature. Spontaneous confocal Raman is used in roughly half of the studies, for measurements of serum [43], tissue [41,44], serum-derived extracellular vesicles [56], and cultured cells [63].

Other studies use SERS. Mandrell et al. [42] applied it to cultured cells, while Carmicheal et al. [57] applied it first to cell lines and later to serum-derived exosomes. Banaei et al. [34] used a targeted serum-EV immunoassay, and a later study applied SERS to serum-derived exosomes captured on a functionalized chip [55]. Szymoński and colleagues reported two companion studies combining Raman hyperspectral mapping [58,59] with CNN classification.

From a preprocessing perspective, the selected papers use distinct strategies. Aslam et al. [45] do not explicitly report preprocessing, suggesting that they either adopted the protocol from Li et al. [44] or used the preprocessed dataset directly. In contrast, Mandrell et al. [42] and Lazzini et al. [41] evaluate multiple dimensionality-reduction or feature-selection pathways before selecting a final classifier. The two companion papers by Szymoński combine polynomial baseline correction and Savitzky–Golay smoothing with spectral-range restriction; the Analytical and Bioanalytical Chemistry (ABC) study also uses noise-based data augmentation. The Sezer et al. [63] study processes the spectra shown in its main figures using background subtraction and software-level smoothing, but describes the spectra used for classification as unprocessed. Banaei et al. [34] reduce each targeted SERS measurement to the intensity of one reporter peak before decision-tree classification. Finally, the study [55] adopts a conventional four-step pipeline.

The selected papers are grouped by the sample type on which the main classifier was applied and, within each group, reviewed in reverse chronological order. The first group comprises biofluid and extracellular-vesicle studies [34,43,55–57,64] (Sections 4.1–4.6); the second comprises human and animal tissue studies [14,23,24,41,44,45,58,59] (Sections 4.7–4.12); and the third comprises cultured-cell studies [42,63] (Sections 4.13 and 4.14). A study that used more than one sample type was assigned to the type on which its reported performance was obtained. The SRS study by Zhang et al. [24] which has no classifier, is grouped by sample type. This grouping organizes the narrative only: Tables S1–S3 in Supplementary File S1 and the synopsis Table 1 retain chronological order.

4.1. Serum-Exosome SERS with a Hybrid CNN–MLP Classifier

Zhang et al. [55] developed an integrated SERS chip functionalized with a CP05 peptide for automatic capture of serum exosomes, coupled to a hybrid CNN–multilayer perceptron (MLP) classifier for the diagnosis of ten cancer types (breast, thyroid, esophageal, kidney, pancreatic, duodenal, lung, colorectal, ovarian, and gastric) versus non-cancer controls.

The cohort comprised 655 serum samples: 200 non-cancer controls and 455 patients with pathologically confirmed cancer. The authors collected at least 200 SERS spectra per sample and reported 97.4% accuracy for cancer vs. non-cancer discrimination. The paper also reports 97.08% accuracy and an area under the receiver operating characteristic curve (AUC) of 0.99 for early-stage cancer detection, and 93.89% accuracy, 93.1% sensitivity, and 99.03% specificity for the eleven-class task comprising ten cancer types and controls. None of these figures is broken down by cancer type in the text. The reported confusion matrix indicates that the pancreatic class included 45 patients, 41 of whom were correctly classified; the manuscript does not state pancreas-specific accuracy, sensitivity, or specificity.

4.1.1. Spectrum-Level Leakage Masquerading as Patient-Level Validation

The study contains at least 131,000 spectra but only 655 independent serum samples. The manuscript declares that five-fold cross-validation was implemented, but does not say how the spectra are split: the text states only that "the dataset was split into training and validation subsets" in each fold, with no mention of sample or patient identity. When spectra from the same serum sample are distributed across training and validation folds, the classifier can memorize sample-specific SERS fingerprints rather than generalizable cancer biology. Since no patient-holdout protocol is described or reported, the results must be considered potentially affected by spectral pseudoreplication.

4.1.2. Small Per-Class Cohort and Unstable Early-Stage Metrics

With 455 malignant samples across 10 cancer types, the average class size is about 45, and the pancreatic class is close to this average. With class sizes this small, accuracy is an unstable point estimate rather than a robust measure of screening performance. Moreover, the 93.89% pooled eleven-class accuracy cannot be assumed to hold uniformly across classes of such different sizes. The non-cancer class alone (200 samples) is more than four times the size of the average malignant class.

4.1.3. Overall Opinion

The integrated SERS–AI platform is an elegant engineering concept, but the reported metrics reflect spectrum-level separability within a limited cohort, not validated diagnostic accuracy on unseen patients. Thus, these very high accuracy estimates should not yet be interpreted as evidence of independent patient generalization or pancreas-specific clinical performance.

4.2. Raman Spectroscopy and Raman Optical Activity of Blood Plasma

Kralova et al. [64] applied Raman spectroscopy and its polarized variant, Raman optical activity (ROA), to blood plasma for the differential diagnosis of three gastrointestinal cancers—hepatocellular, colorectal, and pancreatic—against healthy controls. The study comprised 191 plasma samples; Raman and ROA spectra were acquired simultaneously and combined for classification. The paper addressed two diagnostic tasks: distinguishing cancer from controls, and differentiating among the three cancer types. Two classification strategies were compared: PCA–linear discriminant analysis (LDA) on the entire spectral range, and LDA on a manually curated selection of spectral bands, both evaluated by leave-one-out cross-validation (LOOCV). Supplementary File S1 reports these results: Band-based LDA achieved 88.5% sensitivity, 89.7% specificity, and 89.0% accuracy (170/191) for cancer-versus-control classification. For the three-cancer task, overall accuracy was 86.7% for band-based LDA versus 76.1% for PCA–LDA.

4.2.1. Unclear Nesting of Feature and Band Selection

The main methodological concern is the choice of classification variables. The authors used Welch's *t*-test to retain principal components whose scores differed significantly between sample groups. The initial spectral bands for the LDA model were chosen from PCA loading plots and then tuned using LDA-generated standardized canonical discriminant coefficients. These procedures involve supervised choices, but the authors do not state whether PCA, principal-component selection, and band optimization were repeated within each LOOCV training fold rather than performed once on the full dataset. The paper reports a separate 20-fold cross-validation for the PCA step itself, but does not establish whether the full selection procedure was nested within the final classifier evaluation. If selection used all samples before LOOCV, a given patient's label would have influenced the

representation later used to classify that same patient once held out. Therefore, the reported sensitivity, specificity, and accuracy figures carry a substantial risk of model-selection artifacts, cf. Kapoor and Narayanan [65].

4.2.2. Decoupled Evaluation of Cancer Detection and Cancer Typing

The cancer-versus-control and three-way cancer-type models are trained and reported separately; their combination into a single two-stage diagnostic procedure is proposed but never evaluated end-to-end, so errors propagating from the first stage into the second are not quantified.

4.2.3. Potential Demographic Confounding

The paper describes the three cancer groups as "age-balanced," but control subjects are 7–9 years younger on average than each cancer group (hepatocellular carcinoma (HCC) 68, colorectal carcinoma (CC) 66, pancreatic cancer (PaC) 67, controls 59 years). Sex distributions also differ across groups. The authors attribute the sex imbalance to the underlying clinical population at the recruiting network rather than to sampling choices, but do not report accuracy figures stratified by age or sex. The gaps in age and sex represent potential confounding and can partially inflate the 89% cancer-versus-control accuracy, which may reflect these gaps rather than a disease-specific signature.

4.2.4. Unreported Acquisition-Failure Rates and Potential Selection Bias

The authors acknowledge difficulties in collecting several blood plasma samples from cancer patients despite the fluorescence-suppression protocol, but do not report the number of failures, their distribution across diagnoses, or their relationship to the final 191-sample cohort. It cannot therefore be concluded that the unmeasurable samples were the hardest to classify, or that their omission affects the reported accuracy.

4.2.5. Internal Validation and Limited Clinical Scope

All reported classifier evaluations derive from LOOCV within one recruitment network across affiliated Prague hospitals and departments, which follow a common plasma-collection protocol; no independent test cohort, held-out-center evaluation, or second-instrument validation is reported. The comparison is also limited to cancer patients versus healthy controls: the study does not report specificity against pancreatitis or other clinically relevant, non-malignant conditions.

4.2.6. Overall Opinion

Kralova et al. worked on a 191-sample cohort and addressed both cancer-versus-control discrimination and differentiation among three malignancies. The analysis is reported at the sample level, without describing individual acquisition scans as independent observations, which bypasses the explicit scan-level pseudoreplication. The combined use of Raman and ROA, and the biologically motivated exclusion of non-specific carotenoid bands from the band-based classifiers, are genuine strengths. Nevertheless, the reported accuracies rest on a feature-selection procedure whose nesting within cross-validation is not established; the cancer-versus-control contrast is entangled with an unaddressed age and sex difference between groups. These issues suggest the 87% three-way and 89% binary accuracies are better read as upper bounds on achievable performance than as direct estimates of it.

4.3. Raman and FTIR Spectroscopy of Patient-Derived Extracellular Vesicles

Uthamacumaran et al. [56] proposed a method for differentiating cancer patients from healthy controls using Raman spectroscopy combined with Fourier transform infrared spec-

troscopy (FTIR) spectroscopy. The authors analyzed spectra from serum-derived EVs. The cohort comprised nine subjects: four cancer patients, each diagnosed with a different cancer type (colorectal, hepatocellular, breast, or pancreatic cancer), and five healthy controls. The authors selected 18 Raman spectra from the EVs, while FTIR measurements were obtained from a subset of 14 spectra. Three machine-learning classifiers—an AdaBoost-boosted Random Forest ensemble, Decision Trees, and Support Vector Machines—were evaluated using a single 50:50 training–test split, with an additional five-fold cross-validation analysis. In the limited 1800–1940 cm^{-1} interval, Decision Tree and SVM classifiers reported 100% hold-out accuracy on baseline-corrected Raman spectra, whereas the corresponding cross-validation scores drop to 30% and 90%, respectively.

4.3.1. Spectral Pseudoreplication and Class Label

The 18 Raman and 14 FTIR spectra do not correspond to 18 and 14 independent patients, respectively, since they originate from only nine biologically independent subjects. The paper describes spectrum-level splitting for both the 50:50 partition and the five-fold cross-validation, but does not indicate how spectra were grouped. If spectra from the same preparation occur in both training and test sets, the classifiers may learn patient-specific features rather than cancer-associated signatures. Moreover, the cancer class is biologically and histologically heterogeneous, with four different cancer types represented by only four patients. The classifier may therefore distinguish (from healthy controls) patient- or cancer-type-specific spectral signatures rather than a general cancer signature. In particular, only one patient had pancreatic cancer, making any conclusion about pancreatic cancer classification highly uncertain.

4.3.2. Validation-Protocol Ambiguity and Reporting Inconsistency

The reported standard deviations of approximately 20–40 percentage points indicate substantial variability in the performance estimates. The interpretation of the reported metrics is further complicated by the small number of subjects, the spectrum-level splitting, the preprocessing choices, and the restriction to a selected spectral interval. These factors make it difficult to assess the robustness and generalizability of the reported classification performance.

4.3.3. Overall Opinion

The study provides a proof of concept that standard classifiers can separate EV spectra under the experimental conditions considered. It also illustrates the effect of preprocessing, through the comparison of raw, baseline-corrected, and frequency-restricted spectra. However, the very small and heterogeneous cohort, together with the unclear grouping of spectra during validation, prevents the reported 100% hold-out accuracy from being interpreted as evidence of robust cancer, and in particular pancreatic cancer, classification.

4.4. Targeted SERS Immunoassay of Tumor-Derived Extracellular Vesicles

Banaei et al. [34] combined a targeted SERS immunoassay with shallow decision trees to distinguish pancreatic cancer from pancreatitis and healthy controls. The assay captured circulating EVs from serum using a cluster of differentiation 81 (CD81)-functionalized gold substrate and selectively labeled vesicles expressing ephrin type-A receptor 2 (EphA2) with antibody-functionalized gold nanoparticles. The nanoparticles carried 4-nitrobenzenethiol as an extrinsic Raman reporter, and the intensity of its 1336 cm^{-1} peak was used as a surrogate measure of tumor-derived EV abundance. Thus, the classifier operates on a targeted immunoassay readout rather than on an intrinsic serum or EV Raman fingerprint.

The clinical dataset comprised serum from five patients with pancreatic cancer, five patients assigned to the pancreatitis group, and five healthy controls. Four cancer patients

had metastatic pancreatic adenocarcinoma and one had locally advanced disease. Ten Raman readouts were obtained from each serum sample, producing 150 measurements from only 15 biologically independent subjects. The authors trained depth-two decision trees for the pancreatic cancer-versus-pancreatitis and pancreatic cancer-versus-control comparisons and reported five-fold cross-validated sensitivity of 0.95 and specificity of 0.96. The EphA2-positive EV signal was also significantly higher in the cancer group than in either comparison group.

4.4.1. Documented Readout-Level Pseudoreplication and Leakage

The five-fold cross-validation randomly partitions the 150 Raman measurements, not the 15 patients. Consequently, technical readouts from the same serum sample can occur in both training and test folds. The model can therefore exploit patient-, substrate-, incubation-, or batch-specific signal characteristics, and the 150 measurements do not constitute 150 independent diagnostic cases. Because there are only five patients per group, the reported sensitivity and specificity cannot be interpreted as estimates of performance on unseen patients. The same problem affects the inferential analysis: the reported $F(2, 72)$ statistic indicates that technical measurements, rather than the 15 patients, were treated as independent observations.

4.4.2. Ambiguous Performance Reporting

Figure 7 of the original article presents two binary trees trained on the full datasets, each containing 100 technical readouts, whereas the text reports only one sensitivity and one specificity value. It is not stated whether these metrics refer to one task, an average over the two tasks, or a pooled calculation. No fold-level results, confusion matrices, confidence intervals, or patient-level predictions are provided. The printed specificity equation is also internally inconsistent because its final numerator is labeled “true positive” rather than “true negative”. Restricting tree depth may reduce model complexity, but it does not correct leakage or the non-independence of technical replicates.

4.4.3. Cohort and Label Limitations

The pancreatitis group is clinically heterogeneous and includes one patient with acute gallstone pancreatitis despite the study’s stated focus on chronic pancreatitis. More importantly, the claim of potential early cancer detection is not supported by the cohort: every pancreatic cancer case was locally advanced or metastatic. Serial dilution of serum from one metastatic patient tests analytical signal response, not biological sensitivity in early-stage disease. No independent cohort, early-stage cases, benign pancreatic lesions, or additional clinical center was used for validation.

4.4.4. Experimental Burden and Protocol Inconsistency

Although the method is described as rapid and requiring minimal sample preparation, it depends on a multi-stage immunoassay with substrate functionalization, blocking, repeated incubations, washing, nanoparticle labeling, and drying. The methods also give inconsistent incubation times: one section reports 6 h for serum capture and 10 h for reporter labeling, whereas another reports 2 and 8 h, respectively. This ambiguity complicates reproduction and weakens the claim of a rapid clinical workflow.

4.4.5. Overall Opinion

This study is relevant to Raman–ML because it applies a decision-tree classifier to a SERS-derived clinical signal and addresses the important differential diagnosis between pancreatic cancer and pancreatitis. It provides useful proof-of-concept evidence that EphA2-positive EVs can be measured using a targeted SERS immunoassay. Nevertheless, the ML

task is effectively a threshold classification of one reporter-peak intensity, and its validation is dominated by technical-replicate leakage from a 15-patient cohort. The reported 0.95 sensitivity and 0.96 specificity should therefore be described as internal readout-level results, not as evidence of patient-level diagnostic performance or early cancer detection.

4.5. Serum Raman Spectroscopy for Distinguishing Benign and Malignant Pancreatic Tumors

Yan et al. [43] used confocal Raman micro-spectroscopy of blood serum to distinguish benign from malignant pancreatic tumors. The study included serum samples from 18 patients with pancreatic cancer and 10 patients with benign pancreatic tumors. Raman spectra were acquired in the 400–1800 cm^{-1} range; the authors attributed the observed spectral differences to changes in alcohols, lipids, amides, and nucleic acids. To reduce dimensionality, PLS was fitted to the original 815-wavenumber spectra using both the spectral matrix and the benign/malignant labels, yielding ten latent components (the first six accounted for 97.045% of cumulative variance, with third partial least-squares component (PLS3) alone explaining 48.9%). The resulting components were then classified using LDA, SVMs, and k-NN under five-fold cross-validation. The linear, quadratic, and cubic SVM kernels, as well as fine k-NN ($k = 1$), converged on the same peak accuracy of 96.4%; LDA performed slightly worse.

4.5.1. Supervised Preprocessing Leakage

PLS is a legitimate dimensionality-reduction tool for Raman data (cf. Sec. 1 for the specific nature of Raman data) before classification, but it is supervised. Indeed, the paper affirms that PLS was fitted using "the independent and dependent variables" — i.e., the full spectral matrix together with the benign/malignant labels — as a single step applied to the entire dataset, and that only the subsequent classification stage was cross-validated. Because the labels influenced the feature space before any train–test split, information from the nominal validation folds leaked into the training representation. This is not a potential risk; it is a confirmed case of supervised preprocessing leakage.

4.5.2. Statistical Inadequacy

With only 28 patients, each misclassification modifies accuracy by approximately 3.6 percentage points; the 96.4% figure (27/28) is thus inherently brittle. Given the 18:10 class ratio, a majority-class classifier would already achieve 64.3%. This factor limits the precision of the estimated classification performance and its generalizability to independent patient populations.

4.5.3. Ambiguity in the Experimental Unit

A source of potential data leakage is that with 28 patients, five-fold cross-validation creates test sets of five to six patients per fold. The authors collected multiple Raman measurements from different locations within each serum sample, yet they do not specify how these spectra were aggregated or assigned to folds. If spectra were split across the training and test sets, accuracy would increase unnecessarily and patient-specific information would be leaked. Moreover, the paper does not state whether the five-fold split was stratified at the patient level. Without this confirmation, the 96.4% figure cannot be assumed to reflect patient-level generalization.

4.5.4. Composition of the Benign Class

The benign group is clinically heterogeneous, and its exact composition is unspecified. It may encompass histologically distinct entities (e.g., IPMN, serous cystadenoma) with divergent molecular profiles. Grouping these into a single class means the classifier could

learn to separate one particular benign entity from cancer rather than benignity per se, further limiting generalizability.

4.5.5. Overall Opinion

The study [43] offers an interesting biological scenario for serum-based Raman diagnostics, although its ML implementation remains procedurally naive. The 96.4% accuracy is undermined by PLS leakage, unclear experimental-unit handling, and a minute cohort. The results should be read as a feasibility sketch rather than a validated diagnostic estimate.

4.6. Exosome SERS with Discriminant Function Analysis

Carmicheal et al. [57] studied the combination of SERS with principal component–discriminant function analysis (PC-DFA) to distinguish pancreatic cancer from normal tissue. SERS spectra were acquired from exosomes isolated from normal pancreatic epithelial and pancreatic cancer cell lines. The authors reduced the original spectral space (1004 wavenumbers) to 20 principal components. These components were then used for discriminant function analysis (DFA). The PC-DFA model achieved 90% accuracy on the cell-line spectra. To investigate the potential diagnostic efficacy, the model was applied to serum samples from 10 controls and 10 patients with early-stage cancer.

4.6.1. Validation Protocol and Potential Information Leakage

The manuscript does not provide enough detail to determine whether all data transformations, particularly PCA, were independently fitted within each training fold. The dimensionality reduction on the entire dataset before cross-validation is a possible source of information leakage. Moreover, the 121 specimens are not all biologically independent, since multiple SERS measurements originate from the same exosome preparation. There is no evidence that spectra from the same source were confined to a single fold.

4.6.2. Overall Opinion

Carmicheal et al. correctly recognize exosome SERS as a promising liquid-biopsy modality. However, given the issues described above, while the study supports the potential of Raman/SERS combined with multivariate analysis as a discovery and proof-of-concept tool, it does not yet provide sufficient evidence for robust clinical diagnostic performance.

4.7. Classical Classifiers on Raman Spectra of Surgical Pancreatic Specimens

Lazzini et al. [41] applied LDA, Gaussian naive Bayes (GNB), and a random forest (RF) classifier to Raman spectra of surgical pancreatic specimens, combined with either PCA or a custom spectral-band selection procedure (SPSEL) for dimensionality reduction, to distinguish normal pancreas from pancreatitis and ductal adenocarcinoma across a set of fifteen patients. The study claims a maximum accuracy of approximately 96% with RF. Another important result achieved by this study is the evidence that the Raman peak of methionine (at 695 cm^{-1}) is *less* intense in PDAC than in healthy tissues or in pancreatitis: the authors interpret this trend as consistent with deoxyribonucleic acid (DNA) hypomethylation, a process common in many tumors including PDAC [66] and mechanistically linked to methionine-cycle metabolism [67].

4.7.1. Spectral Pseudoreplication and Undisclosed Cross-Validation Grouping

Of the fifteen patients, four had normal tissue, three had pancreatitis, and ten had PDAC, with two patients providing both normal and pancreatitis regions from the same tissue sample. Therefore, patient identity is strongly confounded with class membership ab initio. In addition, the authors did not specify how the 5-fold cross-validation is applied to the pooled spectra. Without a grouping constraint, spectra from a single patient can be

distributed across training and test folds, allowing a classifier to exploit patient-specific features rather than tissue biology.

4.7.2. Label-Informed Oversampling Before Cross-Validation

The study leverages synthetic minority oversampling technique (SMOTE) to address class imbalance, yielding a balanced set of 1140 spectra per class. As described in the previous paragraph, the authors do not indicate whether the synthetic minority instances were generated separately within each training fold. Following the way SMOTE constructs synthetic spectra, such a missing protocol creates label-correlated information across the train–test boundary with reported inflations to metrics [68]. Moreover, the small error reported for the best configuration (e.g., $\pm$1.8% for RF) is consistent with cross-fold contamination from correlated or synthetic spectra.

4.7.3. Overall Opinion

This work addresses a genuinely three-class diagnostic problem, which is close to a realistic clinical decision. The spectral assignments proposed for the discriminating Raman bands are consistent with the literature but not independently assessed. The risk of data leakage and correlated spectra limits the validity of the reported accuracies. The 96% figure should be read as evidence of spectrum-level separability within this cohort, not as an estimate of performance on unseen patients.

4.8. Deep Learning-Assisted Stimulated Raman Histology of EUS-FNA Specimens

Tan et al. [23] developed a dual-stage EfficientNetV2-L pipeline for intraoperative histologic classification of pancreatic endoscopic ultrasound-guided fine-needle aspiration (EUS-FNA) specimens imaged by SRS microscopy. The SRS images—from 76 patients, 55 malignant, 6 benign, 5 non-diagnostic, and 10 resected-margin patients imaged by simulated puncture—were arranged into 512 $\times$ 512-pixel blocks (2000 patches per class). The authors used the blocks to train a network that first separates diagnostic from non-diagnostic blocks and then benign from malignant blocks. The model achieved a block-level AUC of 0.996 for both stages and a confusion-matrix diagonal above 0.98, and was further evaluated on an independent, four-hospital external set of 33 cases, reaching 96.97% case-level accuracy, comparable to four human pathologists (90.91–100%).

4.8.1. Block-Level Pseudoreplication and Non-Grouped Splitting

The central methodological weakness concerns the internal train/validation split. The authors state that blocks from all 76 patients were pooled and divided into train/test with a ratio of 8:2. However, the paper does not describe patient- or slide-level grouping. The absence of an explicit patient- or slide-level grouping allows a random block-level split from one patient's slide to appear on both sides of the partition.

This is structurally the same failure mode documented by Yagis et al. for slice-level (as opposed to subject-level) cross-validation in two-dimensional (2D) CNNs, especially for small cohorts [69]. The very high block-level AUC (0.996) and confusion-matrix diagonal ($>$0.98) should accordingly be read as evidence of within-slide separability rather than patient generalization.

4.8.2. Preparation-Route Confound in the Benign Training Class

Ten of the seventy-six training patients contribute benign material exclusively through ex-vivo "simulated EUS-FNA" puncture of resected tumor margins. This yields 1130 patches, which constitute the majority (57%) of the 2000 benign training patches. In contrast, malignant blocks come from genuine in-vivo fine-needle aspiration (FNA) aspirates. Since ex-vivo and in-vivo aspirates originate from different protocols, the network

learns the benign vs. malignant discrepancy that may somewhat track protocol instead of histopathological distinction, an example of a labeling confound.

4.8.3. Small, Imbalanced External Validation Set

The four-hospital, 33-case external cohort is a genuine strength relative to most of this corpus—it is the only prospectively independent, multi-site test set among the reviewed studies—but it contains only 5 benign and 2 non-diagnostic cases against 26 malignant.

4.8.4. Qualitative Rather than Co-Registered Histologic Comparison

The visual correspondence between SRS and H&E reported by the study is based on 17 sample pairs imaged at "nearby locations" with one member of each pair processed by SRS and the other by standard H&E. This supports a qualitative demonstration that SRS captures comparable histologic features; however, since the two images are not the same tissue section, the comparison does not ensure a pixel- or structure-level ground-truth validation of the SRS representation.

4.8.5. Overall Opinion

This study is externally validated. Its quantitative gradient-weighted class activation mapping (Grad-CAM)/pathologist concordance ($\kappa = 0.87$ on 31 cases, not explicitly identified as the same 33-case external cohort) is remarkable. However, the headline block-level performance is a typical case of non-grouped-split pseudoreplication, and the benign training class is confounded with tissue-acquisition protocol. The paper's claim of "comparable diagnostic accuracy to experienced pathologists" is best supported by the external case-level result alone.

4.9. Explainable Machine Learning on Mouse Tissue Spectra

Aslam et al. [45] proposed an explainable supervised machine-learning approach for detecting PDAC from Raman spectra. The study used a mouse model comprising 20 animals injected with human pancreatic cancer cells. The authors collected 2529 Raman spectra: 1305 from tumor tissue and 1224 from normal pancreas. Notably, Li et al. [44] used the same cohort. Rather than using the spectral intensities directly, the authors derived an interpretable feature set from three complementary approaches: statistical descriptors, Raman peak features, and extended empirical mode decomposition. A total of 162 features were extracted and subsequently reduced using support vector machine recursive feature elimination (SVM-RFE) combined with correlation bias reduction (CBR). The reduced feature sets were then used for classification with three baseline algorithms—*k*-NN, LDA, and linear support vector machine (LSVM)—alongside a nonlinear support vector machine (NLSVM). The latter achieved the highest reported accuracy, 98.5%. The authors also associated selected Raman peaks with mutations in KRAS and tumor protein p53 (TP53), presenting this as a biologically interpretable fingerprint.

4.9.1. Pseudoreplication and Spectral-Level Leakage

The dataset contains 2529 spectra but only 20 biological subjects. The authors partitioned the data into a 10% hold-out test set and performed 15-fold cross-validation on the remainder, but the manuscript does not clearly establish that spectra originating from the same animal were kept within the same partition. If spectra from the same animal were allowed to fall into both training and test folds, the classifier could learn animal-specific optical or anatomical signatures rather than tumor biology. With only 20 independent subjects, any claim of 98.5% mouse-level accuracy is unsustainable; the effective sample size is an order of magnitude smaller than the number of spectral trials suggests, on average

126 spectra per subject. Therefore, the reported results do not constitute a validation of the approach in human clinical samples.

4.9.2. Supervised Feature-Selection Leakage

SVM-RFE is a wrapper method that uses class labels to rank and eliminate features. The manuscript does not describe an inner cross-validation loop in which SVM-RFE is refitted exclusively on the training folds. If the 162 features were ranked once on the full dataset before any train–test split, the selected variables were shaped by information from the nominal validation sets. This introduces supervised preprocessing leakage.

4.9.3. Overall Opinion

Aslam et al. make an honest effort to move beyond black-box classification by linking selected features to spectral bands. Yet the pipeline is technically fragile: SVM-RFE leakage, pseudoreplication, and an optimistic benchmarking protocol mean that the reported 98.5% accuracy is best interpreted as an upper bound under idealized conditions; these limitations preclude direct inference to human PDAC detection.

4.10. Raman Hyperspectral Mapping and CNN Classification of Tumor Subtypes

Szymoński et al. published two related studies in 2023 applying Raman hyperspectral mapping (RHM) and CNN classification to overlapping pancreatic cancer (PC) subtypes using the same institutional tissue archive. In the paper published in ABC [58], the authors classified subcellular components (nuclei and cytoplasm) of conventional PDAC, intraductal papillary mucinous carcinoma (IPMC), and ampulla of Vater adenocarcinoma (AVAC) across sixteen slides from fifteen patients using a CNN combined with a secondary dense classifier fed by hand-crafted spectral ratios. In an earlier study published in *European Journal of Nuclear Medicine and Molecular Imaging (EJNMMI)* [59], twenty-four slides from seventeen patients (including two benign cases) were used to classify the three tumor types plus stroma and benign tissue by coupling FTIR imaging with RHM and training a CNN directly on raw spectra.

However, neither study specifies whether their respective patient cohorts overlap. Both projects used the same institutional archive over a near-identical timeframe and did not state any specific exclusion criteria, so patient reuse across the two datasets remains possible. Regarding performance, while the ABC workflow delivers over 98% accuracy thanks to random noise augmentation and the ratio-based classifier, the EJNMMI study reports 94% validation accuracy on a coarser five-class task without implementing either technique.

4.10.1. Spectrum-Level Splitting and Absence of Patient-Stratified Validation

Both studies rely on a 70:30 train-test split executed at the spectrum level. Because individual slides yield 10,000 to 18,000 spectra, this approach introduces significant spatial autocorrelation; thousands of highly correlated spectra from identical tissue structures are split between training and evaluation subsets, artificially inflating performance. While the ABC paper attempts a modest external validation by testing the CNN on an additional conventional pancreatic ductal adenocarcinoma (cPDAC) slide (13,924 spectra) from a new patient, the EJNMMI study does not evaluate its model on unseen patients at all. Therefore, the 94% accuracy reflects spectrum-level correlation within a fixed set of formalin-fixed, paraffin-embedded (FFPE) blocks rather than generalization to independent pathology. Without robust patient-level validation, the EJNMMI findings offer limited generalizability, denoting a weaker methodology compared to the companion study.

4.10.2. Classification Task and Cohort Scope Change Between Papers

A further inconsistency lies in the tasks assigned to the two architectures, which are presented as equivalent evidence despite solving non-comparable problems. The EJNMMI study focuses on five coarse classes, including benign conditions, whereas the ABC paper targets seven sub-cellular compartments without accounting for a benign category. Consequently, the two models target different biological scales.

4.10.3. Circular Interpretation of Biomarkers of Two Different Kinds

The validation approach in the ABC study uses a circular design. The workflow analyses DNA methylation and β-sheet ratios using spectra grouped by the same CNN under evaluation. This results in a near-automatic separation in the PCA space instead of providing independent biochemical confirmation. The EJNMMI study uses a similar form of circular reasoning, restricting the subsequent PCA to tissue regions already delineated by hierarchical cluster analysis (HCA) clustering. The distinct biochemical differences revealed by the PCA emerge as a mathematical consequence of the initial unsupervised partition, rather than providing an independent validation criterion.

4.10.4. Overall Opinion

The two papers present an ambitious technical pipeline (RHM + CNN + multivariate analysis) for pancreatic cancer subtyping, yet neither demonstrates patient-level generalization: both rest on spectrum-level pseudoreplication that inflates accuracy, and the single attempt at testing an unseen patient (ABC) is qualitative only. The drop from 98% to 94% suggests that the higher figure was partly purchased by the secondary classifier and by contaminated splitting, rather than a genuine methodological advance. The proposed spectroscopic "fingerprints" for DNA methylation and β-sheet content remain interesting hypotheses. However, their derivation from circular, inconsistent pipelines limits their clinical translatability.

4.11. CNN-Assisted Raman Spectroscopy of Mouse Pancreatic Tissue

Li et al. [44] presented a deep-learning approach, based on CNNs, for distinguishing cancerous from normal pancreatic tissue. The authors acquired 2529 spectra (1305 tumor, 1224 normal pancreas) from 20 mice, using the same cohort as Aslam et al. [45]. The classifier operates end-to-end on near-raw spectral representations: the original one-dimensional (1D) Raman spectrum, a rank-one 2D image obtained from the outer product of each spectrum with its transpose, and the first principal component (PC1) of that 2D image obtained via PCA. Each representation was passed to a custom four-layer CNN achieving test accuracy above 96% (up to 97.39% for 1D Raman), sensitivity above 97%, and AUC up to 0.993. The authors additionally associated cancer-relevant peaks with protein and collagen content and normal-tissue peaks with lipid and nucleic acid content.

4.11.1. Pseudoreplication and Specimen-Level Data Leakage

The most important methodological issue concerns the statistical independence of the spectra used for training and evaluation. The dataset contains 2529 spectra but only 20 biological animals (126 spectra per animal on average). The authors state that they split "the Raman data" without clarifying whether spectra from the same animal, tumor, or tissue specimen were kept within the same partition. A random distribution of spectra among the three partitions makes the measurements statistically dependent. This is particularly relevant for CNNs. Because they learn representations directly from the input rather than from a fixed set of engineered descriptors, and are therefore more capable than shallow classifiers of exploiting subject-specific artifacts—a mechanism known as shortcut

learning [70]. Yagis et al. report that slice-level rather than subject-level cross-validation in 2D CNNs for medical imaging inflates test accuracy by 29–55%, an effect most pronounced in small-sample regimes such as this twenty-animal cohort [69].

4.11.2. Test-Set-Driven Architecture Selection

The authors report comparing CNNs with three, four, and five convolutional layers and selecting the four-layer configuration because it obtained the higher testing accuracy, sensitivity, and specificity than the alternatives. Choosing an architecture by directly comparing test-set metrics across candidate models is a standard source of optimistic bias: once the test partition has informed a design decision, even informally, it is no longer a valid held-out estimate of generalization

4.11.3. Rank-One Redundancy of the 2D Representation

The 2D Raman image is constructed as the outer product of the normalized 1D spectrum with its own transpose, $M = \mathbf{w}\mathbf{w}^T$, where $\mathbf{w} \in \mathbb{R}^n$. By construction, M is a rank-one, symmetric $n \times n$ matrix and therefore contains the same information as $\mathbf{w}$. This has a direct consequence for the PCA step used to obtain 2D Raman PC1: because each 2D image is prefeatured by PCA individually rather than through a dataset-wide fit, 2D Raman PC1 is algebraically identical to the original 1D spectrum (up to sign and scale). The performance differences reported across the three CNN models (96.44–97.39% accuracy) therefore cannot reflect additional discriminative information carried by the 2D or PC1 representations; they are caused by the bias of feeding identical information through architecturally distinct convolutional filters.

4.11.4. Inconsistent Cross-Validation Reporting

The authors never reconcile the results of five- vs. ten-fold cross-validation (CV). The first was performed to test the CNN models, while the second is related to the extraction of CNN features from the max-pooling layers. Therefore, it is not possible to determine from the text whether this reflects a second, undocumented experiment or a reporting error. This ambiguity affects how the headline accuracy figures should be read.

4.11.5. Overall Opinion

The study is methodologically ambitious—to the authors' knowledge, the first attempt at CNN feature visualization in Raman-based cancer diagnosis—but the deep-learning pipeline is subject to the same categories of bias documented for classical ML pipelines in this review. The 2D representation adds architectural complexity without a demonstrable information gain over the 1D spectrum, making its construction unnecessary. The 97.39% accuracy is an overfitted estimate from a small-animal set, with the issues stated above, not a reproducible diagnostic metric.

4.12. Near-Infrared Raman Spectroscopy in a Mouse Xenograft Model

Pandya et al. [14] investigated near-infrared Raman spectroscopy for distinguishing normal pancreatic tissue from tumors in an orthotopic mouse xenograft model. The experiment involved 33 mice and intended normal/tumor sampling within animals, but resulted in 60 tissues: 31 normal and 29 malignant. From the tissues, the authors collected 920 spectra: 475 normal and 445 malignant. The final animal-to-specimen mapping was not provided. Classification used PCA and DFA, with separate analyses of individual spectra and one mean spectrum per tissue. For the 920 individual spectra, a random 50:50 training/test split repeated 20 times yielded reported average sensitivity and specificity of 96.4% and 96.2%, respectively. leave-one-out (LOO) evaluation yielded 91.0% sensitivity and 88.4% specificity. LOO on the small cohort of 60 tissue means reported sensitivity and specificity of 100%.

4.12.1. Incomplete Documentation of PCA Fitting and Component Use

For individual spectra, 13 principal components explained 96% of the variance, and analysis of variance (ANOVA) identified seven components with significant differences between diagnostic classes. For tissue means, eleven principal components explained approximately 99% of the variance. However, the study does not clarify whether PCA was refitted within each training partition. Its description suggests PCA was computed on the entire pooled cohort—all 920 individual spectra, or all 60 tissue means—before either the 50:50 split or the LOO partition was formed. In such a case, held-out observations would contribute to the representation used for training. Moreover, since the protocol uses the principal-component scores as DFA inputs directly, and ANOVA on the full cohort identified which components were diagnostically significant before that, label-informed selection leakage would also be possible.

4.12.2. Spectral Pseudoreplication in Individual-Spectrum Validation

The random split operates on spectra, and leave-one-out explicitly withholds one spectrum while retaining the others for training. Consequently, the latter evaluates an observation whose tissue is already represented in the training data by other measurements. The random split likewise permits spectra from the same tissue on both sides of the partition; no tissue- or animal-grouped allocation is described. The authors report low variability among measurements from a given tissue.

4.12.3. Tissue Averaging Does Not Establish Animal-Level Independence

The authors analyzed 60 mean spectra (from 60 tissues), leveraging the fact that averaging reduces the contribution of within-tissue spectral replicates. However, the experimental design permits normal and tumor tissues to originate from the same mouse, and so does the corresponding pair of mean spectra. Within-animal sampling is valuable for biological comparison, but it is not equivalent to testing on independent animals.

4.12.4. Author-Acknowledged Overestimation and Numerical Instability

The authors identify the flaws in their experiment, concluding that their results likely overestimate the true sensitivity and specificity. They link high prediction performance to low variability among spectra from the same tissue and expect greater heterogeneity in vivo. Moreover, they describe the individual-spectrum DFA with leave-one-out as unsatisfactory because of multicollinearity, reporting a condition number of $\kappa = 278{,}164.2$. Even if this diagnostic should not be treated as a measure of leakage or a numerical estimate of performance inflation, the authors' caution is nevertheless consistent with the limited independence of the individual spectral observations.

4.12.5. Restricted Biological Scope and Lack of External Validation

The experiment used two human carcinoma cell lines implanted into mice within one experimental setting. No independent animal cohort, human tissue test cohort, or second-instrument evaluation was reported in the paper. The paired normal/tumor design—with each animal contributing one tissue of each class—is a genuine methodological strength, but the model and sampling design offer limited evidence about performance across naturally occurring human disease.

4.12.6. Overall Opinion

This is a historically important paper—one of the earliest RS + ML investigations of pancreatic cancer. It provides exploratory evidence of spectral differences between normal pancreatic tissue and xenograft tumors. Its comparison of individual spectra with tissue means, and its explicit warning about overestimation, are remarkable features. However,

spectrum-level validation does not test independent tissues, tissue averaging does not ensure animal-level independence, and the description does not demonstrate training-only PCA fitting or clarify the role of ANOVA in determining classifier inputs. In conclusion, the reported figures, including 100% sensitivity and specificity for tissue means, should be viewed as internal exploratory results from a small animal study, not as validated clinical estimates or formal upper bounds on attainable accuracy.

4.13. Raman Classification of Pancreatic Cancer Cells Following EMT Modulation

In their study, Sezer et al. [63] examined the epithelial-to-mesenchymal transition (EMT), a key process in various physiological and pathological contexts. Using RS, the authors investigated how indomethacin, a nonsteroidal anti-inflammatory drug, affects transforming growth factor beta (TGF-β)-induced EMT in a pancreatic cancer cell line, tracking changes in peak intensity and position related to chemical composition. EMT-associated spectral changes were observed in treated cells compared to controls, with concentration-dependent effects at 100 and 300 μM. To reduce input dimensionality, the authors applied PCA and implemented four machine-learning classifiers. The PCA–SVM approach yielded a maximum accuracy of 97.6%; AdaBoost reached 96.4%, *k*-NN 95.2%, and Decision Tree 94%. These spectral trends were corroborated by molecular assays (immunostaining and quantitative reverse transcription polymerase chain reaction (qRT-PCR)).

4.13.1. Unprocessed Classification Branch, and PCA Leakage

The paper describes three distinct levels of spectral processing that are easy to confuse. The spectra shown in the main results figures underwent background subtraction and instrument-level smoothing, while the loading plots carry a Savitzky–Golay-smoothed overlay line, applied only for that visualization and not fed back into the analysis. The manuscript affirms that PCA was fitted on "unprocessed Raman signals" without preprocessing. Since no justification for using unprocessed spectra is given, the 97.6% figure is left exposed to instrumental artifacts. In addition, PCA itself was fitted on all 254 spectra as a single step before the 67/33 split described afterwards for classifier training. The test-set spectra therefore contributed to the covariance structure defining the principal components, so the projected scores are not independent between training and test partitions—a case of data leakage.

4.13.2. Single-Split Validation

The authors compared four classifiers on a single 67/33 train/test hold-out, reporting only the best result; this introduces an uncontrolled model-selection bias. Moreover, with 84 spectra in the test set, the performance is sensitive to the random partition and cannot be assumed stable.

4.13.3. Overall Opinion

Sezer et al. provide a sound pharmacological investigation in which Raman spectroscopy effectively tracks EMT-related biochemical changes. However, the machine-learning add-on is methodologically weak: an unjustified choice to classify on uncorrected spectra, PCA applied before splitting, and a single hold-out produce a 97.6% figure that should be treated as an informal separability index, not a validated classification metric.

4.14. SERS of Tumor-Repopulating Cells

Mandrell et al. [42] used Raman spectroscopy to distinguish tumor-repopulating cells (TRCs) from parental controls in the MIA PaCa-2 pancreatic cancer line, and sought the best combination of data type, dimensionality reduction, and classifier. The authors compared spontaneous Raman with gold- and silver-substrate SERS during instrument optimization and selected the silver SERS substrate, which gave the best signal-to-noise

ratio. All 51 subsequent spectra were acquired with this SERS configuration: 37 from 8 parental controls and 14 from 5 TRC preparations. Classification was performed using SVMs with several kernels and *k*-NN. The authors compared combinations of spectral representation, normalization, feature-selection method (*t*-statistic, signal-to-noise statistic (MIT correlation), Relief feature-selection algorithm (RELIEF)), dimensionality (including PCA), and classifier, and reported a best cross-validated accuracy of 0.982 using either *t*-score or MIT correlation-selected peaks without PCA.

The authors themselves acknowledge that the figure of 98.2% is fragile. Since the 14 TRC spectra set the granularity of the cross-validation scheme (one TRC spectrum per fold, giving 14 folds), the 0.982 accuracy corresponds to 13 folds classified perfectly and a single misclassified spectrum. The authors attribute this uniformity to the small sample size and the discrete cross-validation scheme, explicitly ruling out an analysis error; they do not describe it as data leakage. The same granularity, however, is consistent with structural leakage.

4.14.1. Non-Independent Spectra and Structural Leakage

The main methodological issue is that cross-validation was performed at the spectrum level rather than at the level of the 13 cell preparations. At least 3 spectra were collected from different locations in each preparation, and the averaged spectra were not used for machine learning. Each of the 14 folds was constructed to contain exactly one TRC spectrum together with two to three parental-control spectra—a design choice the authors adopted specifically to avoid leave-one-out validation. Since the 37 parental-control spectra come from only 8 preparations, no fold assignment of this kind can keep every preparation's spectra together: spectra from the same parental preparation necessarily recur across multiple folds. Unless all spectra from a preparation were assigned to the same fold, the classifier could exploit preparation- or substrate-specific signatures shared between training and test data. This alone is enough to inflate the reported accuracy. In addition, the headline 0.982 is the mean of the 14 per-fold accuracies (13 folds at 1.0, one fold at 0.75 for a subset of three parental-control and one TRC spectrum) rather than a pooled spectrum-level accuracy: a technical distinction that does not change the substance of the leakage argument.

4.14.2. Feature-Selection and Model-Selection Bias

Secondly, the article fails to mention whether the *t*-statistic, MIT correlation, or RELIEF scores were recalculated using only the training spectra within each cross-validation fold. If these label-dependent feature selection methods were applied to the full dataset once before cross-validation, the held-out labels would influence the selected Raman variables, introducing supervised preprocessing leakage.

4.14.3. Confounded Class Label

The TRC/control label is related to an unblinded experimental batch variable: TRCs underwent 3D fibrin-gel selection, whereas parental controls did not. There is no evidence that the classification can distinguish a signal driven by tumor-repopulating phenotype from one driven by culture condition alone.

4.14.4. Overall Opinion

In short, the study shows that SERS combined with ML techniques can separate these cell lines, but the 98.2% figure reflects an overfitted validation scheme rather than a reproducible diagnostic performance.

4.15. Methodological Summary and Limitations

Supplementary Table S2 reports the summary of Raman modalities, instrumentation, and preprocessing pipelines across the reviewed literature.

4.16. Recurring Methodological Weaknesses

The studies reviewed above share several recurring methodological limitations. The number of spectra is often treated as the sample size, even though the independent experimental unit is the patient, animal, tissue block, serum sample, or cell preparation. Data-dependent operations such as PLS, PCA, feature selection, data augmentation, and oversampling are not always fitted exclusively within the training partition. High internal accuracy is also rarely supported by an untouched external cohort. These limitations are not equivalent: some articles describe procedures that directly imply data leakage, whereas others provide insufficient information to determine whether patient- or specimen-grouped partitioning was used. Table 1 summarizes this classification, while Table S1 in Supplementary File S1 explains the reasoning for each study.

Table 1. *Cont.*

Study	Sample, Modality	Units → Spectra	Internal Validation	Headline Result	Leakage	External Cohort
Li et al., 2021 [44]	Mouse tissue, spontaneous Raman + CNN	20 mice → 2529 (cohort shared with Aslam)	Split level not stated; architecture chosen on test metrics	97.39% (1D input)	Unresolved	None
Mandrell et al., 2020 [42]	Cell-line preparations, SERS	13 preparations → 51	14-fold CV on single spectra	98.2% (mean of fold accuracies)	Documented	None
Yan et al., 2020 [43]	Human serum, confocal Raman	28 patients (18 malignant, 10 benign)	5-fold CV after PLS fitted on all data	96.4% (27/28 equivalent)	Documented	None
Carmicheal et al., 2019 [57]	Cell-line and serum exosomes, SERS	Cell-line preparations; serum 10 + 10	Fold structure and PCA fitting not documented	90% on cell lines	Unresolved	Serum application only (20 subjects)
Pandya et al., 2008 [14]	Mouse xenograft tissue, NIR Raman	33 mice, 60 tissues → 920	50:50 splits (20×), LOO; PCA on pooled cohort	96.4/96.2% (50:50); 91.0/88.4% (LOO), sens./spec.	Unresolved	None

CV, cross-validation; LOO(CV), leave-one-out (cross-validation); RF, random forest; NLSVM, nonlinear SVM; EV, extracellular vesicle; RHM, Raman hyperspectral mapping.

4.16.1. Cross-Study Interpretation

The most common weakness is not a small number of spectra but a small number of independent biological units relative to the number of correlated spectra. The correct denominator for uncertainty is therefore the number of patients, animals, or preparations, not the number of spectral acquisitions. It is also worth noting that two of the 15 reviewed studies (Aslam et al. [45], Li et al. [44]) share the same 20-mouse cohort (same instrument, same 2529 spectra). A defensible pipeline partitions these independent units first, then fits every downstream step—preprocessing, feature selection, balancing, model selection—using training data only. Nested validation becomes necessary once several processing or model alternatives are compared, and an external cohort remains the only way to test transportability across patients, instruments, and laboratories.

4.16.2. Consequences for Reported Accuracy

The weaknesses summarized in Table S1 of Supplementary File S1 do not imply that the observed Raman differences are spurious. Rather, they limit the strength and scope of the conclusions that can be drawn. Most studies provide evidence that their spectra contain class-associated structure under the acquisition conditions examined, but they do not yet provide unbiased estimates of prospective diagnostic performance. Therefore, we read the highest reported accuracies as internal spectrum-level or within-cohort results unless both patient-grouped validation and independent external validation are explicitly demonstrated.

5. Standardization Gaps, Benchmark Datasets and Future Directions

5.1. Reporting and Validation Standards

The weaknesses summarized in Section 4.16 recur across research groups, instruments, and publication years. This kind of pattern points to a structural gap in how the literature is reported and checked.

Reporting guidance for this type of study has existed since 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) [71] is a 22-item checklist for papers that develop or validate a multivariable diagnostic or prognostic model; it covers data source, predictors, sample size, validation strategy and performance measures. Its 2024 update, TRIPOD + AI [72], replaces it with 27 items and adds machine-learning requirements, including how validation folds were built and how class imbalance was handled. No reviewed study cites either document. The 2015 checklist predates all of them, whereas TRIPOD + AI appeared after most were written. Both target prediction models for individual patients, so they apply directly to the serum and tissue studies and only by analogy to the cell-line and animal work. Sezer et al. [63] illustrate what is at stake: their PCA was fitted on all 254 spectra, including the 84 later assigned to the test set, so the projection used for classification was not learned from training data alone. An item asking authors to state the order of fold construction and feature extraction would have made this visible in the paper itself, instead of leaving it to be inferred from the full text.

5.2. Instrumental and Computational Comparability

In a round-robin experiment within the European Cooperation in Science and Technology (COST) action Raman4clinics, aliquots of five reference substances were measured on 35 setups in 15 institutes [73]. The authors noticed changes in various metrics, including standard deviation and peak position. As a consequence, they expect models built on one setup to fail on another when the setups differ substantially. In addition, preprocessing the data adds a different kind of variation. Baseline correction, normalization and peak

extraction differ from study to study (Supplementary File S1 Table S2), so a band assigned to the same molecular group in two papers need not have been quantified in the same way. It is interesting to point out that quantitative imaging ran into this problem and addressed it with the Image Biomarker Standardization Initiative (IBSI) [74], which fixed feature definitions and processing workflows and released reference values for checking software. Raman spectroscopic imaging will need to address these gaps to become a tool for oncological diagnostics.

5.3. Benchmark Datasets: Lessons from Other Tumors

Shared data enable comparability testing and external validation. Unfortunately, pancreatic Raman research does not have such shared data. By contrast, glioma imaging represents a clear precedent. The Brain Tumor Segmentation (BraTS) challenge, active since 2012, combines multi-institutional MRI with expert annotation and withholds the test labels, so that participants are ranked on data they never saw [75,76]. More recent editions added survival prediction [77] and a genomic classification task [78]. Melanoma followed a similar route through the International Skin Imaging Collaboration (ISIC) [79], whose public archive allowed algorithms to be compared on images that none of the developing groups had collected.

The pancreas now has its own benchmark in radiology. Artificial Intelligence and Radiologists in Pancreatic Cancer Detection using standard-of-care CT scans (PANORAMA) released the training data and analysis code of a multicentre CT cohort of 3440 patients, together with a platform that scores algorithms on a sequestered, blinded test set, and used the same test data to compare an AI system with 68 radiologists [80]. The authors of this review recognize the PANORAMA initiative as the true lack in pancreatic Raman studies. Indeed, the closest Raman analogues are the stimulated Raman histology classifiers for glioma, developed on large intraoperative image collections and tested across institutions [22,81].

However, Raman benchmarks with different specimens are available. For example, the bacterial isolate dataset of Ho et al. [82] is employed as a common testbed for deep classifiers; RamanBench, a recent preprint, gathers 74 public datasets with 325,668 spectra and evaluates 28 models under one protocol [83], and a separate study compares Raman-specific deep networks on three open datasets under a unified training protocol [84]. Moreover, the small open library of reference spectra for individual biomolecules [85] serves a dissimilar purpose, since it catalogs pure compounds and holds no patient-derived spectra.

Metadata are also important: Every spectrum should carry patient, specimen, and acquisition identifiers, so that partitions can be grouped by patient and the pseudoreplication discussed in Section 4.16 becomes a queryable property of the dataset. Moreover, the functional metadata already compiled in Supplementary File Table S2 (instrument, excitation wavelength and power, SERS substrate, preprocessing pipeline) must accompany each record.

Standardized reporting makes internal validation verifiable, harmonized instruments and pipelines make results comparable across laboratories, and a shared benchmark with a sequestered test set makes external validation possible. All of them contribute to making accuracy figures reliable.

5.4. Future Directions

The immediate need is a multicentre patient cohort with predefined patient-level partitions, acquired on documented instruments and released with a held-out portion whose labels stay with the organizers, along the lines of BraTS and PANORAMA. Sensitivity studies that isolate the effect of each preprocessing choice on the extracted features would

show how much of the reported discrimination survives a change of pipeline. quantum machine learning (QML) has been proposed for cancer-related data [86,87]; none of the reviewed studies uses it, and any claimed benefit for Raman classification would have to be tested against tuned classical baselines on identical partitions, with computational cost and the distinction between simulation and quantum hardware reported.

6. Conclusions

Taken together, the studies reviewed here show that Raman measurements contain enough biochemical information to distinguish several pancreatic-related sample classes. This has been demonstrated in tissues, biofluids, cell preparations, and animal models, but the studies address different questions and cannot be combined into a single estimate of diagnostic performance. Most reported accuracies describe discrimination within a small cohort collected under one set of experimental conditions.

The main concern is the difference between the number of spectra and the number of independent biological samples. If spectra from the same patient, animal, or preparation are allowed to enter both training and test sets, the resulting accuracy may not reflect performance on a new case. The same problem arises when preprocessing, feature selection, or model tuning uses information from the test data. In several articles, the reported methods do not provide enough detail to determine whether these risks were avoided. This does not invalidate the observed spectral differences, but it limits their interpretation to separability within the studied cohort.

Future studies should therefore define the biological unit before partitioning the data, fit every data-dependent step using training data only, and evaluate the final model on independent patients and instruments. Patient-derived benchmark datasets, complete acquisition metadata, and clearer reporting would make results easier to compare across laboratories. The most urgent challenge is to develop a framework to examine how preprocessing choices affect extracted spectral features. Until such validation becomes routine, Raman spectroscopy should be regarded as a promising research tool for pancreatic cancer rather than as an established clinical diagnostic platform.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app162111019/s1, File S1 contains Table S1, Principal methodological weaknesses identified in the reviewed Raman and Raman–ML studies; Table S2, Summary of Raman modalities, instrumentation, and preprocessing pipelines across the reviewed literature; Table S3, Studies that met the eligibility criteria of the coverage search but are not part of the index sample; File S2 contains Figure S1, Publications per year; Table S1, Screening log of the 394 unique records retrieved by the coverage search. References [88–100] are cited in the supplementary materials.

Author Contributions: Conceptualization, R.D.G., G.L., R.G., A.C., G.S., L.M. and M.D.; methodology, R.D.G.; software, R.D.G.; validation, R.D.G.; formal analysis, R.D.G.; investigation, R.D.G.; data curation, R.D.G.; writing—original draft preparation, R.D.G., G.L. and M.D.; writing—review and editing, R.D.G., G.L., R.G., A.C., G.S., L.M. and M.D.; supervision, R.D.G. and M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this review.

Acknowledgments: During the preparation of this manuscript, the authors used Claude Pro (Anthropic; Sonnet 5.5) , Kimi K3 (Moonshot AI), and Prism (OpenAI; free version, accessed in October 2026) for preliminary article summarization, page-level source localization, methodological cross-

checking, and language editing. During the revision, Claude was also used to propose the queries of the coverage search, to organize the exported records, to apply the deduplication rule, and to propose an exclusion code for each record. The authors ran the searches and took every screening decision; the proposed codes were checked on a random stratified sample of 30 of the 359 exclusions, as described in the Review Scope and Methods section. The observations on the 15 analyzed studies were each verified by the authors against the cited pages of the original sources. All outputs were reviewed and edited by the authors, who take full responsibility for the final content. R.D.G, G.L, and M.D. wish to thank NanoBioTLab CNR-IBF for useful support.

Conflicts of Interest: Five of the authors (G.L., R.G., A.C., L.M, and M.D.) co-authored one of the studies included in the critical analysis (Lazzini et al., 2025 [41]). The authors declare no other conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

1D	one-dimensional
2D	two-dimensional
ABC	Analytical and Bioanalytical Chemistry
AI	artificial intelligence
ALS	asymmetric least squares
ANOVA	analysis of variance
ANN	artificial neural network
ANNs	artificial neural networks
AUC	area under the receiver operating characteristic curve
AVAC	ampulla of Vater adenocarcinoma
BraTS	Brain Tumor Segmentation
CBR	correlation bias reduction
CC	colorectal carcinoma
CD81	cluster of differentiation 81
CNN	convolutional neural network
CNNs	convolutional neural networks
COST	European Cooperation in Science and Technology
cPDAC	conventional pancreatic ductal adenocarcinoma
CRM	confocal Raman microscopy
CRS	coherent Raman scattering
CT	computed tomography
CV	cross-validation
DFA	discriminant function analysis
DL	deep learning
DNA	deoxyribonucleic acid
DTGS	deuterated triglycine sulfate
ECD	electronic circular dichroism
EJNMMI	European Journal of Nuclear Medicine and Molecular Imaging
ELISA	enzyme-linked immunosorbent assay
EM-CCD	electron-multiplying charge-coupled device
EMT	epithelial-to-mesenchymal transition
EphA2	ephrin type-A receptor 2
EUS-FNA	endoscopic ultrasound-guided fine-needle aspiration
EV	extracellular vesicle
EVs	extracellular vesicles
FFPE	formalin-fixed, paraffin-embedded
FFT	fast Fourier transform
FNA	fine-needle aspiration
FTIR	Fourier transform infrared spectroscopy
GNB	Gaussian naive Bayes
Grad-CAM	gradient-weighted class activation mapping
H&E	hematoxylin and eosin

HCA	hierarchical cluster analysis
HCC	hepatocellular carcinoma
IBSI	Image Biomarker Standardization Initiative
IDO1	indoleamine 2,3-dioxygenase 1
IPMC	intraductal papillary mucinous carcinoma
IPMN	intraductal papillary mucinous neoplasm
IR	infrared
ISIC	International Skin Imaging Collaboration
k-NN	*k*-nearest neighbors
KRAS	Kirsten rat sarcoma viral oncogene homolog
LDA	linear discriminant analysis
LOO	leave-one-out
LOOCV	leave-one-out cross-validation
LSVM	linear support vector machine
MAPK	mitogen-activated protein kinase
ML	machine learning
MLP	multilayer perceptron
MLR	multiple linear regression
MIT correlation	signal-to-noise statistic
MRI	magnetic resonance imaging
NLSVM	nonlinear support vector machine
NMF	non-negative matrix factorization
OPO	optical parametric oscillator
OPLS-DA	orthogonal partial least squares discriminant analysis
PaC	pancreatic cancer
PANORAMA	Artificial Intelligence and Radiologists in Pancreatic Cancer Detection using standard-of-care CT scans
PC	pancreatic cancer
PCA	principal component analysis
PC-DFA	principal component–discriminant function analysis
PC1	first principal component
PDAC	pancreatic ductal adenocarcinoma
PET	positron emission tomography
PLS	partial least squares
PLS-DA	partial least squares discriminant analysis
PLS3	third partial least-squares component
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
qRT-PCR	quantitative reverse transcription polymerase chain reaction
QML	quantum machine learning
RAS	rat sarcoma
RAS/MAPK	rat sarcoma/mitogen-activated protein kinase
RELIEF	Relief feature-selection algorithm
RF	random forest
RFs	random forests
RHM	Raman hyperspectral mapping
ROA	Raman optical activity
RS	Raman spectroscopy
SERS	surface-enhanced Raman spectroscopy
SHG	second-harmonic generation
SL	supervised learning
SMOTE	synthetic minority oversampling technique
SNV	standard normal variate
SPSEL	spectral-band selection procedure
SRS	stimulated Raman scattering

STAM	stratified pancreatic cancer–stroma spheroids
SVD	singular value decomposition
SVM	support vector machine
SVMs	support vector machines
SVM-DA	support vector machine discriminant analysis
SVM-RFE	support vector machine recursive feature elimination
TDA	topological data analysis
TGF-β	transforming growth factor beta
TML	topological machine learning
TP53	tumor protein p53
TRC	tumor-repopulating cell
TRCs	tumor-repopulating cells
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
t-SNE	*t*-distributed stochastic neighbor embedding
UL	unsupervised learning

References

1. GBD 2017 Pancreatic Cancer Collaborators. The global, regional, and national burden of pancreatic cancer and its attributable risk factors in 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol. Hepatol.* **2019**, *4*, 934–947. https://doi.org/10.1016/S2468-1253(19)30347-4.
2. Siegel, R.L.; Kratzer, T.B.; Wagle, N.S.; Sung, H.; Jemal, A. Cancer statistics, 2026. *CA A Cancer J. Clin.* **2026**, *76*, e70043. https://doi.org/10.3322/caac.70043.
3. Del Chiaro, M.; Ishida, H.; Schulick, R.D. Importance of increasing resection rates in pancreatic cancer treatment. *BJS Open* **2025**, *9*, zraf029. https://doi.org/10.1093/bjsopen/zraf029.
4. Park, W.; Chawla, A.; O'Reilly, E.M. Pancreatic Cancer: A Review. *JAMA* **2021**, *326*, 851–862. https://doi.org/10.1001/jama.2021.13027.
5. Stoop, T.F.; Javed, A.A.; Oba, A.; Koerkamp, B.G.; Seufferlein, T.; Wilmink, J.W.; Besselink, M.G. Pancreatic cancer. *Lancet* **2025**, *405*, 1182–1202. https://doi.org/10.1016/S0140-6736(25)00261-2.
6. Hu, Z.I.; O'Reilly, E.M. Therapeutic developments in pancreatic cancer. *Nat. Rev. Gastroenterol. Hepatol.* **2024**, *21*, 7–24. https://doi.org/10.1038/s41575-023-00840-w.
7. Luo, J. KRAS mutation in pancreatic cancer. *Semin. Oncol.* **2021**, *48*, 10–18. https://doi.org/10.1053/j.seminoncol.2021.02.003.
8. Ohtsuka, T.; Fernandez-del Castillo, C.; Furukawa, T.; Hijioka, S.; Jang, J.Y.; Lennon, A.M.; Miyasaka, Y.; Ohno, E.; Salvia, R.; Wolfgang, C.L.; Wood, L.D. International evidence-based Kyoto guidelines for the management of intraductal papillary mucinous neoplasm of the pancreas. *Pancreatology* **2024**, *24*, 255–270. https://doi.org/10.1016/j.pan.2023.12.009.
9. Pea, A.; Hruban, R.H.; Wood, L.D. Genetics of pancreatic neuroendocrine tumors: Implications for the clinic. *Expert Rev. Gastroenterol. Hepatol.* **2015**, *9*, 1407–1419. https://doi.org/10.1586/17474124.2015.1092383.
10. Orlando, A.; Franceschini, F.; Muscas, C.; Pidkova, S.; Bartoli, M.; Rovere, M.; Tagliaferro, A. A comprehensive review on Raman spectroscopy applications. *Chemosensors* **2021**, *9*, 262. https://doi.org/10.3390/chemosensors9090262.
11. Oshima, Y.; Haruki, T.; Koizumi, K.; Yonezawa, S.; Taketani, A.; Kadowaki, M.; Saito, S. Practices, Potential, and Perspectives for Detecting Predisease Using Raman Spectroscopy. *Int. J. Mol. Sci.* **2023**, *24*, 12170. https://doi.org/10.3390/ijms241512170.
12. Bocklitz, T.W.; Guo, S.; Ryabchykov, O.; Vogler, N.; Popp, J. Raman Based Molecular Imaging and Analytics: A Magic Bullet for Biomedical Applications!? *Anal. Chem.* **2016**, *88*, 133–151. https://doi.org/10.1021/acs.analchem.5b04665.
13. Ralbovsky, N.M.; Lednev, I.K. Raman spectroscopy and chemometrics: A potential universal method for diagnosing cancer. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *219*, 463–487. https://doi.org/10.1016/j.saa.2019.04.067.
14. Pandya, A.K.; Serhatkulu, G.K.; Cao, A.; Kast, R.E.; Dai, H.; Rabah, R.; Poulik, J.; Banerjee, S.; Naik, R.; Adsay, V.; et al. Evaluation of Pancreatic Cancer With Raman Spectroscopy in a Mouse Model. *Pancreas* **2008**, *36*, e1–e8. https://doi.org/10.1097/MPA.0b013e31815a3f1c.
15. Manganelli Conforti, P.; Lazzini, G.; Russo, P.; D'Acunto, M. Raman Spectroscopy and AI Applications in Cancer Grading: An Overview. *IEEE Access* **2024**, *12*, 54816–54852. https://doi.org/10.1109/ACCESS.2024.3388841.
16. Wei, D.; Chen, S.; Liu, Q. Review of Fluorescence Suppression Techniques in Raman Spectroscopy. *Appl. Spectrosc. Rev.* **2015**, *50*, 387–406. https://doi.org/10.1080/05704928.2014.999936.
17. Polli, D.; Kumar, V.; Valensise, C.M.; Marangoni, M.; Cerullo, G. Broadband Coherent Raman Scattering Microscopy. *Laser Photonics Rev.* **2018**, *12*, 1800020. https://doi.org/10.1002/lpor.201800020.

18. Alfonso-García, A.; Mittal, R.; Lee, E.S.; Potma, E.O. Biological Imaging with Coherent Raman Scattering Microscopy: A Tutorial. *J. Biomed. Opt.* **2014**, *19*, 071407. https://doi.org/10.1117/1.JBO.19.7.071407.
19. Sippl, C.; Stark, F.; Schneider, K.I.; Reyes Medina, B.; Schulz-Schaeffer, W.; Brinkmann, M.; Neumann, F.; Droop, R.; Ullmann, S.; Würthwein, T.; et al. Stimulated Raman Spectroscopy for Intraoperative Glioblastoma Diagnosis—A Complementary Tool to Frozen Section? *Cancers* **2026**, *18*, 1053. https://doi.org/10.3390/cancers18071053.
20. Steybe, D.; Poxleitner, P.; Metzger, M.C.; Rothweiler, R.; Beck, J.; Straehle, J.; Vach, K.; Weber, A.; Enderle-Ammour, K.; Werner, M.; et al. Stimulated Raman histology for histological evaluation of oral squamous cell carcinoma. *Clin. Oral Investig.* **2023**, *27*, 4705–4713. https://doi.org/10.1007/s00784-023-05098-9.
21. Zeng, J.; Wu, G.; Jia, H.; Liu, C.; Liu, C.; Zhang, Y.; Liu, S.; Wang, J.; Yue, S. Stimulated Raman Scattering Microscopy Reveals Aberrant Triglyceride Accumulation in Lymphatic Metastasis of Papillary Thyroid Carcinoma. *Anal. Chem.* **2024**, *96*, 5735–5740. https://doi.org/10.1021/acs.analchem.4c00306.
22. Hollon, T.C.; Jiang, C.; Chowdury, A.; Nasir-Moin, M.; Kondepudi, A.; Aabedi, A.; Adapa, A.; Al-Holou, W.N.; Heth, J.A.; Sagher, O.; et al. Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. *Nat. Med.* **2023**, *29*, 828–832. https://doi.org/10.1038/s41591-023-02252-4.
23. Tan, T.; Ma, L.; Guo, Y.; Chen, T.; Meng, L.; Luo, K.; Zhou, P.; Cai, M.; Ji, M.; Hu, H. Intelligent Diagnosis of Pancreatic Biopsy From Endoscopic Ultrasound-Guided Fine-Needle Aspiration Via Stimulated Raman Histopathology. *Lab. Investig.* **2025**, *105*, 104182. https://doi.org/10.1016/j.labinv.2025.104182.
24. Zhang, L.; Zou, X.; Huang, J.; Fan, J.; Sun, X.; Zhang, B.; Zheng, B.; Guo, C.; Fu, D.; Yao, L.; Ji, M. Label-Free Histology and Evaluation of Human Pancreatic Cancer with Coherent Nonlinear Optical Microscopy. *Anal. Chem.* **2021**, *93*, 15550–15558. https://doi.org/10.1021/acs.analchem.1c03861.
25. Abdullah, M.M.A.M.; Manisi, Y.S.; Elawami, F.M.E.Y.; Mohamad, A.F.; Zuraida, B.H.H.M.A. Critical assessment of Raman spectroscopy for clinical diagnostics: progress, limitations, and barriers to translation. *Clin. Chem. Lab. Med.* **2026**, *64*, 2141–2158. https://doi.org/10.1515/cclm-2026-0404.
26. Lee, S.; Dang, H.; Moon, J.I.; Kim, K.; Joung, Y.; Park, S.; Yu, Q.; Chen, J.; Lu, M.; Chen, L.; et al. SERS-based microdevices for use as in vitro diagnostic biosensors. *Chem. Soc. Rev.* **2024**, *53*, 5394–5427. https://doi.org/10.1039/d3cs01055d.
27. Lu, Z.; Wang, J.; Yan, S. Quantitative Surface-Enhanced Raman Spectroscopy: Challenges, Strategies, and Prospects. *Molecules* **2026**, *31*, 191. https://doi.org/10.3390/molecules31010191.
28. Pilot, R.; Signorini, R.; Durante, C.; Orian, L.; Bhamidipati, M.; Fabris, L. A Review on Surface-Enhanced Raman Scattering. *Biosensors* **2019**, *9*, 57. https://doi.org/10.3390/bios9020057.
29. Shiohara, A.; Wang, Y.; Liz-Marzán, L.M. Recent Approaches toward Creation of Hot Spots for SERS Detection. *J. Photochem. Photobiol. C Photochem. Rev.* **2014**, *21*, 2–25. https://doi.org/10.1016/j.jphotochemrev.2014.09.001.
30. Le Ru, E.C.; Meyer, M.; Etchegoin, P.G. Proof of single-molecule sensitivity in surface enhanced Raman scattering (SERS) by means of a two-analyte technique. *J. Phys. Chem. B* **2006**, *110*, 1944–1948. https://doi.org/10.1021/jp054732v.
31. Pérez-Jiménez, A.I.; Lyu, D.; Lu, Z.; Liu, G.; Ren, B. Surface-Enhanced Raman Spectroscopy: Benefits, Trade-Offs and Future Developments. *Chem. Sci.* **2020**, *11*, 4563–4577. https://doi.org/10.1039/d0sc00809e.
32. Stiles, P.L.; Dieringer, J.A.; Shah, N.C.; Van Duyne, R.P. Surface-enhanced Raman spectroscopy. *Annu. Rev. Anal. Chem.* **2008**, *1*, 601–626. https://doi.org/10.1146/annurev.anchem.1.031207.112814.
33. Prochazka, M. *Surface-Enhanced Raman Spectroscopy: Bioanalytical, Biomolecular and Medical Applications*; Biological and Medical Physics, Biomedical Engineering; Springer International Publishing: Cham, Switzerland, 2016. https://doi.org/10.1007/978-3-319-23992-7.
34. Banaei, N.; Moshfegh, J.; Kim, B. Surface-enhanced Raman spectroscopy-based immunoassay detection of tumor-derived extracellular vesicles to differentiate pancreatic cancers from chronic pancreatitis. *J. Raman Spectrosc.* **2021**, *52*, 1810–1819. https://doi.org/10.1002/jrs.6241.
35. Valera, P.S.; Henriques-Pereira, M.; Wagner, M.; Gaspar, V.M.; Mano, J.F.; Liz-Marzán, L.M. Surface-Enhanced Raman Scattering Monitoring of Tryptophan Dynamics in 3D Pancreatic Tumor Models. *ACS Sensors* **2024**, *9*, 4236–4247. https://doi.org/10.1021/acssensors.4c01210.
36. Wold, S. Chemometrics; what do we mean with it, and what do we want from it? *Chemom. Intell. Lab. Syst.* **1995**, *30*, 109–115. https://doi.org/10.1016/0169-7439(95)00042-9.
37. Geladi, P. Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochim. Acta Part B At. Spectrosc.* **2003**, *58*, 767–782. https://doi.org/10.1016/S0584-8547(03)00037-5.
38. Oancea, B.; Necula, M.; Milea, E.C.; Amărioarei, A.; Petre, I.; Păun, M.M. Machine Learning meets Raman spectroscopy: A systematic review of literature in cancer diagnostics. *Procedia Comput. Sci.* **2025**, *270*, 2666–2675. https://doi.org/10.1016/j.procs.2025.09.388.
39. Huang, B.; Huang, H.; Zhang, S.; Zhang, D.; Shi, Q.; Liu, J.; Guo, J. Artificial intelligence in pancreatic cancer. *Theranostics* **2022**, *12*, 6931–6954. https://doi.org/10.7150/thno.77949.

40. Blake, N.; Gaifulina, R.; Griffin, L.D.; Bell, I.M.; Thomas, G.M.H. Machine Learning of Raman Spectroscopy Data for Classifying Cancers: A Review of the Recent Literature. *Diagnostics* **2022**, *12*, 1491. https://doi.org/10.3390/diagnostics12061491.
41. Lazzini, G.; Gaeta, R.; Pollina, L.E.; Comandatore, A.; Furbetta, N.; Morelli, L.; D'Acunto, M. Raman Spectroscopy Based Diagnosis of Pancreatic Ductal Adenocarcinoma. *Sci. Rep.* **2025**, *15*, 13240. https://doi.org/10.1038/s41598-025-98122-9.
42. Mandrell, C.T.; Holland, T.E.; Wheeler, J.F.; Esmaeili, S.M.A.; Amar, K.; Chowdhury, F.; Sivakumar, P. Machine Learning Approach to Raman Spectrum Analysis of MIA PaCa-2 Pancreatic Cancer Tumor Repopulating Cells for Classification and Feature Analysis. *Life* **2020**, *10*, 181. https://doi.org/10.3390/life10090181.
43. Yan, Z.; Ma, C.; Mo, J.; Han, W.; Lv, X.; Chen, C.; Chen, C.; Nie, X. Rapid Identification of Benign and Malignant Pancreatic Tumors Using Serum Raman Spectroscopy Combined with Classification Algorithms. *Optik* **2020**, *208*, 164473. https://doi.org/10.1016/j.ijleo.2020.164473.
44. Li, Z.; Li, Z.; Chen, Q.; Ramos, A.; Zhang, J.; Boudreaux, J.P.; Thiagarajan, R.; Bren-Mattison, Y.; Dunham, M.E.; McWhorter, A.J.; et al. Detection of Pancreatic Cancer by Convolutional-Neural-Network-Assisted Spontaneous Raman Spectroscopy with Critical Feature Visualization. *Neural Netw.* **2021**, *144*, 455–464. https://doi.org/10.1016/j.neunet.2021.09.006.
45. Aslam, M.; Rajbdad, F.; Azmat, S.; Li, Z.; Boudreaux, J.P.; Thiagarajan, R.; Yao, S.; Xu, J. A Novel Method for Detection of Pancreatic Ductal Adenocarcinoma Using Explainable Machine Learning. *Comput. Methods Programs Biomed.* **2024**, *245*, 108019. https://doi.org/10.1016/j.cmpb.2024.108019.
46. Lazzini, G.; D'Acunto, M. Chondrogenic Cancer Grading by Combining Machine and Deep Learning with Raman Spectra of Histopathological Tissues. *Appl. Sci.* **2024**, *14*, 10555. https://doi.org/10.3390/app142210555.
47. Yan, L.; Su, H.; Liu, J.; Wen, X.; Luo, H.; Yin, Y.; Guo, X. Rapid Detection of Lung Cancer Based on Serum Raman Spectroscopy and a Support Vector Machine: A Case-Control Study. *BMC Cancer* **2024**, *24*, 791. https://doi.org/10.1186/s12885-024-12578-y.
48. Rao, S.; Sharma, N.; Bhat, V.G.; Kamath, V.; Thakur, M.; Melanthota, S.K.; Das, S.; Dehury, B.; Mazumder, N. Raman Spectroscopy and Machine Learning in the Diagnosis of Breast Cancer. *Lasers Med Sci.* **2025**, *40*, 348. https://doi.org/10.1007/s10103-025-04597-3.
49. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. https://doi.org/10.1007/BF00994018.
50. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
51. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New, York, , NY, USA, 2009. https://doi.org/10.1007/978-0-387-84858-7.
52. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: Oxfordshire, UK, 2017. https://doi.org/10.1201/9781315139470.
53. Conti, F.; Lazzini, G.; Gaeta, R.; Pollina, L.E.; Comandatore, A.; Furbetta, N.; Morelli, L.; D'Acunto, M.; Moroni, D.; Pascali, M.A. Topological Machine Learning for Raman Spectroscopy: Perspectives for Pancreatic Diseases. *Proceedings* **2025**, *129*, 61. https://doi.org/10.3390/proceedings2025129061.
54. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. https://doi.org/10.1136/bmj.n71.
55. Zhang, C.; Zhao, W.H.; Zuo, D.; Zhou, T.; Hou, W.; Wang, L.; Shi, S.; Yang, Y.; Liu, Y.; Sun, S.K.; et al. AI-Guided SERS Defines a Pan-Cancer Diagnostic Biomarker. *Adv. Sci.* **2026**, *13*, e16268. https://doi.org/10.1002/advs.202516268.
56. Uthamacumaran, A.; Elouatik, S.; Abdouh, M.; Berteau-Rainville, M.; Gao, Z.h.; Arena, G. Machine Learning Characterization of Cancer Patients-Derived Extracellular Vesicles Using Vibrational Spectroscopies: Results from a Pilot Study. *Appl. Intell.* **2022**, *52*, 12737–12753. https://doi.org/10.1007/s10489-022-03203-1.
57. Carmicheal, J.; Hayashi, C.; Huang, X.; Liu, L.; Lu, Y.; Krasnoslobodtsev, A.; Lushnikov, A.; Kshirsagar, P.G.; Patel, A.; Jain, M.; et al. Label-Free Characterization of Exosome via Surface Enhanced Raman Spectroscopy for the Early Detection of Pancreatic Cancer. *Nanomed. Nanotechnol. Biol. Med.* **2019**, *16*, 88–96. https://doi.org/10.1016/j.nano.2018.11.008.
58. Szymoński, K.; Skirlińska-Nosek, K.; Lipiec, E.; Sofińska, K.; Czaja, M.; Wilkosz, N.; Krupa, M.; Wanat, F.; Ulatowska-Białas, M.; Adamek, D. Combined Analytical Approach Empowers Precise Spectroscopic Interpretation of Subcellular Components of Pancreatic Cancer Cells. *Anal. Bioanal. Chem.* **2023**, *415*, 7281–7295. https://doi.org/10.1007/s00216-023-04997-w.
59. Szymoński, K.; Lipiec, E.; Sofińska, K.; Skirlińska-Nosek, K.; Czaja, M.; Seweryn, S.; Wilkosz, N.; Birarda, G.; Piccirilli, F.; Vaccari, L.; et al. Variabilities in global DNA methylation and β-sheet richness establish spectroscopic landscapes among subtypes of pancreatic cancer. *Eur. J. Nucl. Med. Mol. Imaging* **2023**, *50*, 1792–1810. https://doi.org/10.1007/s00259-023-06121-7.
60. Qi, Y.; Liu, B.; Shi, J.; Liu, Y. High-Precision Intelligent Diagnosis of Pancreatic Cancer: Flowing Diffuseness from Single to Whole. *Anal. Chem.* **2025**, *97*, 12548–12557. https://doi.org/10.1021/acs.analchem.5c00465.
61. Bunganic, B.; Sovickova, L.; Tatarkovic, M.; Kocourkova, L.; Suchanek, S.; Fric, P.; Setnicka, V.; Zavoral, M. Molecular spectroscopy of blood plasma - Towards the diagnostics of pancreatic cancer? *Gastroenterol. A Hepatol.* **2015**, *69*, 518–524. https://doi.org/10.14735/amgh2015518.

62. Chen, T.; Cheng, Q.; Lee, H.J. Machine-learning-mediated single-cell classification by hyperspectral stimulated Raman scattering imaging. In *Proceedings of the SPIE—The International Society for Optical Engineering*; SPIE: Bellingham, WA, USA, 2021; Volume 11900, p. 119000V. https://doi.org/10.1117/12.2602857.
63. Sezer, G.; Onses, M.S.; Sakir, M.; Sahin, F.; Çamdal, A.; Sezer, Z.; Inal, A.; Ciftci, Z. Indomethacin Prevents TGF-β-Induced Epithelial-to-Mesenchymal Transition in Pancreatic Cancer Cells: Evidence by Raman Spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *280*, 121493. https://doi.org/10.1016/j.saa.2022.121493.
64. Kralova, K.; Vrtelka, O.; Fouskova, M.; Hribek, P.; Bunganic, B.; Miskovicova, M.; Urbanek, P.; Zavoral, M.; Petruzelka, L.; Habartova, L.; et al. Raman Spectroscopy and Raman Optical Activity of Blood Plasma for Differential Diagnosis of Gastrointestinal Cancers. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2024**, *305*, 123430. https://doi.org/10.1016/j.saa.2023.123430.
65. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4*, 100804. https://doi.org/10.1016/j.patter.2023.100804.
66. Larsson, S.C.; Giovannucci, E.; Wolk, A. Methionine and Vitamin B_6 Intake and Risk of Pancreatic Cancer: A Prospective Study of Swedish Women and Men. *Gastroenterology* **2007**, *132*, 113–118. https://doi.org/10.1053/j.gastro.2006.10.017.
67. Ulrey, C.L.; Liu, L.; Andrews, L.G.; Tollefsbol, T.O. The impact of metabolism on DNA methylation. *Hum. Mol. Genet.* **2005**, *14*, R139–R147. https://doi.org/10.1093/hmg/ddi100.
68. Demircioğlu, A. Applying Oversampling Before Cross-Validation Will Lead to High Bias in Radiomics. *Sci. Rep.* **2024**, *14*, 11563. https://doi.org/10.1038/s41598-024-62585-z.
69. Yagis, E.; Atnafu, S.W.; García Seco de Herrera, A.; Marzi, C.; Scheda, R.; Giannelli, M.; Tessa, C.; Citi, L.; Diciotti, S. Effect of Data Leakage in Brain MRI Classification Using 2D Convolutional Neural Networks. *Sci. Rep.* **2021**, *11*, 22544. https://doi.org/10.1038/s41598-021-01681-w.
70. Geirhos, R.; Jacobsen, J.H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut Learning in Deep Neural Networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. https://doi.org/10.1038/s42256-020-00257-z.
71. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G.M. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD statement. *J. Clin. Epidemiol.* **2015**, *68*, 134–143. https://doi.org/10.1016/j.jclinepi.2014.11.010.
72. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378.
73. Guo, S.; Beleites, C.; Neugebauer, U.; Abalde-Cela, S.; Afseth, N.K.; Alsamad, F.; Anand, S.; Araujo-Andrade, C.; Aškrabić, S.; Avci, E.; et al. Comparability of Raman Spectroscopic Configurations: A Large Scale Cross-Laboratory Study. *Anal. Chem.* **2020**, *92*, 15745–15756. https://doi.org/10.1021/acs.analchem.0c02696.
74. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **2020**, *295*, 328–338. https://doi.org/10.1148/radiol.2020191145.
75. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J. et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med Imaging* **2015**, *34*, 1993–2024. https://doi.org/10.1109/TMI.2014.2377694.
76. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 170117. https://doi.org/10.1038/sdata.2017.117.
77. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M.; et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv* **2018**, arXiv:1811.02629. http://arxiv.org/abs/1811.02629.
78. Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F.C.; Pati, S.; et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv* **2021**, arXiv:2107.02314. http://arxiv.org/abs/2107.02314.
79. Gutman, D.; Codella, N.C.F.; Celebi, E.; Helba, B.; Marchetti, M.; Mishra, N.; Halpern, A. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2016**, arXiv:1605.01397. http://arxiv.org/abs/1605.01397.
80. Alves, N.; Schuurmans, M.; Rutkowski, D.; Saha, A.; Vendittelli, P.; Obuchowski, N.; Liedenbaum, M.H.; Haldorsen, I.S.; Molven, A.; Yakar, D.; et al. Artificial intelligence and radiologists in pancreatic cancer detection using standard of care CT scans (PANORAMA): An international, paired, non-inferiority, confirmatory, observational study. *Lancet Oncol.* **2026**, *27*, 116–124. https://doi.org/10.1016/S1470-2045(25)00567-4.

81. Kondepudi, A.; Pekmezci, M.; Hou, X.; Scotford, K.; Jiang, C.; Rao, A.; Harake, E.S.; Chowdury, A.; Al-Holou, W.; Wang, L.; et al. Foundation models for fast, label-free detection of glioma infiltration. *Nature* **2025**, *637*, 439–445. https://doi.org/10.1038/s41586-024-08169-3.
82. Ho, C.S.; Jean, N.; Hogan, C.A.; Blackmon, L.; Jeffrey, S.S.; Holodniy, M.; Banaei, N.; Saleh, A.A.E.; Ermon, S.; Dionne, J. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat. Commun.* **2019**, *10*, 4927. https://doi.org/10.1038/s41467-019-12898-9.
83. Koddenbrock, M.; Lange, C.; Legner, R.; Jäger, M.; Kögler, M.; Cruz Bournazou, M.N.; Neubauer, P.; Biessmann, F.; Rodner, E. RamanBench: A Large-Scale Benchmark for Machine Learning on Raman Spectroscopy. *arXiv* **2026**, arXiv:2605.02003. http://arxiv.org/abs/2605.02003.
84. Sineesh, A.; Kamsali, A. Benchmarking deep learning models for Raman spectroscopy across open-source datasets. *Digit. Discov.* **2026**, *5*, 3314.
85. Terán, M.; Ruiz, J.J.; Loza-Alvarez, P.; Masip, D.; Merino, D. Open Raman spectral library for biomolecule identification. *Chemom. Intell. Lab. Syst.* **2025**, *264*, 105476. https://doi.org/10.1016/j.chemolab.2025.105476.
86. Ghobadi, M.Z.; Afsaneh, E. Potential of Quantum Machine Learning for Solving the Real-World Problem of Cancer Classification. *Discov. Appl. Sci.* **2024**, *6*, 513. https://doi.org/10.1007/s42452-024-06220-6.
87. Rahimi, M.; Asadi, F. Oncological Applications of Quantum Machine Learning. *Technol. Cancer Res. Treat.* **2023**, *22*, 15330338231215214. https://doi.org/10.1177/15330338231215214.
88. Habartova, L.; Bunganic, B.; Tatarkovic, M.; Zavoral, M.; Vondrousova, J.; Syslova, K.; Setnicka, V. Chiroptical spectroscopy and metabolomics for blood-based sensing of pancreatic cancer. *Chirality* **2018**, *30*, 581–591. https://doi.org/10.1002/chir.22834.
89. Banaei, N.; Moshfegh, J.; Mohseni-Kabir, A.; Houghton, J.M.; Sun, Y.; Kim, B. Machine learning algorithms enhance the specificity of cancer biomarker detection using SERS-based immunoassays in microfluidic chips. *RSC Adv.* **2019**, *9*, 1859–1868. https://doi.org/10.1039/c8ra08930b.
90. Schie, I.W.; Kiselev, R.; Krafft, C.; Popp, J. Rapid acquisition of mean Raman spectra of eukaryotic cells for a robust single cell classification. *Analyst* **2016**, *141*, 6387–6395. https://doi.org/10.1039/c6an01018k.
91. Phyo, J.B.; Woo, A.; Yu, H.J.; Lim, K.; Cho, B.H.; Jung, H.S.; Lee, M.Y. Label-free SERS analysis of urine using a 3D-stacked AgNW-glass fiber filter sensor for the diagnosis of pancreatic cancer and prostate cancer. *Anal. Chem.* **2021**, *93*, 3778–3785. https://doi.org/10.1021/acs.analchem.0c04200.
92. Shin, H.; Choi, B.H.; Shim, O.; Kim, J.; Park, Y.; Cho, S.K.; Kim, H.K.; Choi, Y. Single test-based diagnosis of multiple cancer types using Exosome-SERS-AI for early stage cancers. *Nat. Commun.* **2023**, *14*, 1644. https://doi.org/10.1038/s41467-023-37403-1.
93. Linh, V.T.N.; Lee, M.Y.; Mun, J.; Kim, Y.; Kim, H.; Han, I.W.; Park, S.G.; Choi, S.; Kim, D.H.; Rho, J.; et al. 3D plasmonic coral nanoarchitecture paper for label-free human urine sensing and deep learning-assisted cancer screening. *Biosens. Bioelectron.* **2023**, *224*, 115076. https://doi.org/10.1016/j.bios.2023.115076.
94. Al Ja'farawy, M.S.; Linh, V.T.N.; Yang, J.Y.; Mun, C.; Lee, S.; Park, S.G.; Han, I.W.; Choi, S.; Lee, M.Y.; Kim, D.H.; et al. Whole urine-based multiple cancer diagnosis and metabolite profiling using 3D evolutionary gold nanoarchitecture combined with machine learning-assisted SERS. *Sens. Actuators B Chem.* **2024**, *412*, 135828. https://doi.org/10.1016/j.snb.2024.135828.
95. Zhang, Y.; Zhang, Y.; Gong, R.; Liu, X.; Zhang, Y.; Sun, L.; Ma, Q.; Wang, J.; Lei, K.; Ren, L.; et al. Label-Free Prediction of Tumor Metastatic Potential via Ramanome. *Small Methods* **2025**, *9*, 2400861. https://doi.org/10.1002/smtd.202400861.
96. Cho, M.; Koh, E.Y.; Kim, Y.; Kim, S.J.; Pack, C.G.; Jun, E.; Kim, J.K. Molecular Profiling of Inflammatory and Myofibroblast Cancer-Associated Fibroblast Subtypes Derived from Human Pancreatic Stellate Cells Using Machine Learning-Based Label-Free Raman Spectroscopy. *Biomater. Res.* **2025**, *29*, 0292. https://doi.org/10.34133/bmr.0292.
97. Cutshaw, G.; Demidova, E.V.; Czyzewicz, P., III; Quam, E.; Lorang, N.; Al Siyabi, A.W.; Batra, S.; Arora, S.; Bardhan, R. Metabolic and proteomic signatures differentiate inflammatory phenotypes from cancer and predict treatment response in patient sera. *Bioeng. Transl. Med.* **2025**, *10*, e70029. https://doi.org/10.1002/btm2.70029.
98. Wang, H.; Li, Y.; Chen, S.; Yan, J.; Zhang, J.; Li, P.; Schioth, H.B.; Zhang, C.; Yang, Y.; Li, J.; et al. New label-free serum exosomes detection method based on hierarchical SERS substrate for diagnosis of pancreatic cancer using AI. *Sens. Actuators B Chem.* **2025**, *433*, 137588. https://doi.org/10.1016/j.snb.2025.137588.
99. Khalid, E.; Ali, A.; Majeed, M.I.; Nawaz, H.; Alshammari, A.; Albekairi, N.A.; Zahid, L.; Ditta, A.; Rabbani, R.; Khan, R.; et al. Gold nanoparticles based surface-enhanced Raman spectroscopy with multivariate analysis for characterization of 100 kDa blood serum fractions of pancreatic cancer. *Photonics Nanostructures Fundam. Appl.* **2026**, *72*, 101621. https://doi.org/10.1016/j.photonics.2026.101621.
100. Sohng, W.; Noh, H.S.; Park, C.; Jang, E.B.; Han, I.W.; Chang, J.; Lee, M.Y.; Chung, H. Electrochemical separation-enhanced SERS signals of urinary metabolites for improved pancreatic cancer detection. *Chem. Eng. J.* **2026**, *529*, 173117. https://doi.org/10.1016/j.cej.2026.173117.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/app162111019