Analyst

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2025, **150**, 3349

Raman spectroscopy in tandem with machine learning – based decision logic methods for characterization and detection of primary precancerous and cancerous cells†

Uraib Sharaha,[a,b] Daniel Hania,[c] Dima Bykhovsky,[d] Itshak Lapidot,‡[e,f] Mahmoud Huleihel‡[a] and Ahmad Salman *‡[g]

Early cancer detection improves patient outcomes, but most Raman spectroscopy research has focused on discriminating between normal and malignant cells, ignoring the essential precancerous stage. This study fills that gap by combining Raman spectroscopy with machine learning methods to characterize and categorize normal (primary fibroblast cells from mouse embryos), precancerous (murine fibroblast cell lines (NIH/3T3)), and malignant mouse fibroblast cells transformed by a murine sarcoma virus (MBM-T) as cancerous cells. Key spectral bands associated with malignancy progression were identified using ANOVA-based feature selection, while Log-likelihood estimation decision logic enhanced classification robustness across multiple measurements per cell. The method was 95.8% accurate in classifying normal from cancerous cells, 91% for normal *vs.* precancerous cells, and 86% for precancerous vs cancerous cells. These results show that Raman spectroscopy has the potential to be a valuable diagnostic tool for early cancer detection, offering insight into carcinogenesis spectrum indications. This study advances Raman-based diagnostics in oncology by strengthening spectrum analysis and classification algorithms.

Received 30th March 2025,
Accepted 24th June 2025

DOI: 10.1039/d5an00360a

rsc.li/analyst

Introduction

Cancer remains one of the most significant worldwide health challenges, accounting for an anticipated 10 million deaths by 2020.[1] According to the United States Cancer Statistics, in 2019, 1 752 735 new invasive cancer cases were reported in the United States. For all cancers combined, the incidence rate was 439 per 100 000 standard population overall.[2] Furthermore, the GLOBOCAN statistics show that an anticipated 19.3 million new cancer cases and about 10.0 million cancer deaths occurred in 2020.[3] Early detection is crucial since the cancer stage at diagnosis has a significant impact on patient survival and quality of life.[4,5]

The prognosis of cancerous patients is strongly dependent on early detection, yet many malignancies lack reliable screening methods for precancerous stages.[6,7] Conventional imaging techniques such as X-ray, MRI, and PET scans often fail to detect subtle biochemical changes preceding malignancy, limiting their effectiveness in early diagnosis.[8–10] This highlights the need for sensitive, label-free techniques to identify early molecular transformations associated with cancer development. Thus, developing novel cancer detection, diagnosis, and treatment methods is urgently required.[1]

Raman spectroscopy has emerged as an effective tool for biochemical characterization of cells and tissues, providing a noninvasive, water-insensitive method for detecting molecular changes.[9,11–18]

Raman spectroscopy enables label-free, hypothesis-free molecular profiling, circumventing antibody optimization and spectral overlap limitations inherent to fluorescence-based techniques.[13,19] Preserving samples in their native state facilitates re-analysis and is particularly suited for dried or archival specimens where labeling is impractical.[20] These attributes make Raman a complementary tool to targeted flow cytometry for broad biomolecular studies.[21]

[a]*Department of Microbiology, Immunology and Genetics, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel*
[b]*Department of Biology, Science and Technology College, Hebron University, Hebron P760, Palestine*
[c]*Department of Green Engineering, SCE - Shamoon College of Engineering, Beer-Sheva 84100, Israel*
[d]*Electrical and Electronics Engineering Department, SCE-Sami Shamoon College of Engineering, Beer-Sheva 84100, Israel*
[e]*Department of Electrical and Electronics Engineering, Afeka Tel-Aviv Academic College of Engineering, Tel-Aviv 69107, Israel*
[f]*LIA Avignon Université, 339 Chemin des Meinajaries, Avignon 84000, France*
[g]*Department of Physics, SCE-Sami Shamoon College of Engineering, Beer-Sheva 84100, Israel. E-mail: ahmad@sce.ac.il; Tel: +972-8-6475794*

†Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5an00360a

‡Contributed equally.

Several studies have effectively combined Raman spectroscopy with machine learning methods to classify normal and cancerous cells.[22–31] However, most of these studies focus on well-established malignancies, overlooking the transitional precancerous stage—a critical window for early intervention.[9,11–18] Precancerous cells exhibit progressive molecular changes, including metabolic shifts, nucleic acid modifications, and altered protein structures, yet the spectral biomarkers associated with this transformation remain poorly characterized.

To address this gap, we present a Raman spectroscopy-based approach that explicitly incorporates the precancerous stage, enabling the detection of molecular changes preceding malignancy. Our study examines three cell types: murine primary fibroblasts (normal), NIH/3T3 fibroblasts (precancerous), and MBM-T sarcoma-transformed cells (cancerous). We identify key Raman biomarkers associated with cancer progression by analyzing spectral differences across these states. For academic integrity, it is important to clarify that these cell lines are not meant to represent a direct oncogenic trajectory but instead serve as distinct phenotypic states used to evaluate our Raman-based approach's sensitivity and discriminatory power.

In addition, to improve classification robustness and account for multiple spectral measurements per cell, we combine powerful machine learning-based feature selection with decision logic techniques. This study advances Raman spectroscopy as a cutting-edge diagnostic tool for early cancer diagnosis. It reveals important spectral wavenumbers that disclose the transition from healthy to precancerous and, eventually, malignant states. Our findings demonstrate the enormous potential of Raman spectroscopy in tandem with machine learning as a sensitive, label-free diagnostic approach, paving the path for more effective screening methods and preventive healthcare interventions.

Materials and methods

Biological system preparation

Primary fibroblast cells from separate mice embryos, murine fibroblast cell lines (NIH/3T3), and malignant mouse fibroblast cells transformed by a murine sarcoma virus (MBM-T) were obtained from different cultures and biological replicates. Sample preparation and Raman measurements were conducted over one year, with each measurement session performed on the same day as sample preparation to maintain consistent physiological conditions.

While measurements for all three classes (normal, precancerous, and malignant) were not always performed on the same day, we minimized potential batch effects by:

- Adhering to strict, consistent preparation protocols,
- Using identical instrument settings for all sessions,
- All cells were used at early passages—passage 2 for primary cells and passages 3–4 for cell lines and transformed cells—with viability and normal morphology confirmed under a light microscope before drying.

The three cell type were maintained in T25 cell culture flasks using an RPMI medium supplemented with 10% FBS, two mM L-glutamine and 2% penicillin–streptomycin solution (50–100 μg ml^{-1}), in a 5% CO_2 environment at 37 °C s We harvested the cells from the culture flask by trypsin treatment, then centrifuged to form a pellet, washed three times with 500 μl of 0.9% NaCl, and re-suspended in 50 μL of 0.9% NaCl. Cell concentration was determined using a hemocytometer. The cells were then pelleted and resuspended to achieve a final 30–50 cells per μL concentration. A 2 μL drop of each sample was applied onto an aluminum-coated slide and left to dry for 10 minutes.

Dried cells were used to ensure sample stability and reproducibility during Raman spectral acquisition. While drying can alter cell morphology and may damage membranes, our analysis targets biochemical composition rather than structural features.

Raman measurements

The dried cell samples were measured using the single-spectrum mode of a Horiba LabRAM HR Evolution Raman microscope equipped with a sincerity CCD detector (deep-cooled to −60 °C, 1024 × 256 pixels). A 532 nm Nd: YAG green laser (10 mW, two μm spot size) was used for illumination. A 10% transmittance filter was placed at the laser probe station to prevent laser-induced sample heating.

Raman measurements were conducted after carefully adjusting the focal plane to ensure accurate focus. Spectra were recorded with a 60 s integration time for all measurements. The laser was focused onto the sample using a 50× NA objective lens (Olympus MPLAN), generating a diffraction-limited spot size of 1.54 μm. A 600 lines per mm grating optimized signal strength while minimizing background autofluorescence.

The wavenumber calibration was done using a silicon reference sample every two hours. For each cell, three measurements were taken from the center, cytoplasm, and membrane regions, as illustrated in Fig. 1. The Randomizing measurement regions were randomized within sessions to reduce systematic bias.

Although 785 nm excitation is standard for live-cell Raman spectroscopy (minimizing fluorescence and photodamage),[32–34] we selected 532 nm for its higher scattering efficiency ($\sim 1/\lambda^4$) and improved spatial resolution, critical for resolving fine biochemical details in fixed/dried cells. This wavelength has been successfully applied to study lipid–protein dynamics and subcellular structures in fixed systems,[35,36] with minimal fluorescence interference in processed samples. A low-power illumination further reduced potential artifacts, aligning with established *ex vivo* cellular analysis protocols.[37–39]

Over one year, 222 individual and different cells were analyzed using the Raman facility: 92 normal, 76 precancerous, and 54 cancerous cells. These different 222 cells were conducted across multiple biological replicates from different mice, cultures, and batches, ensuring experimental variability and enhancing the generalizability of the findings.

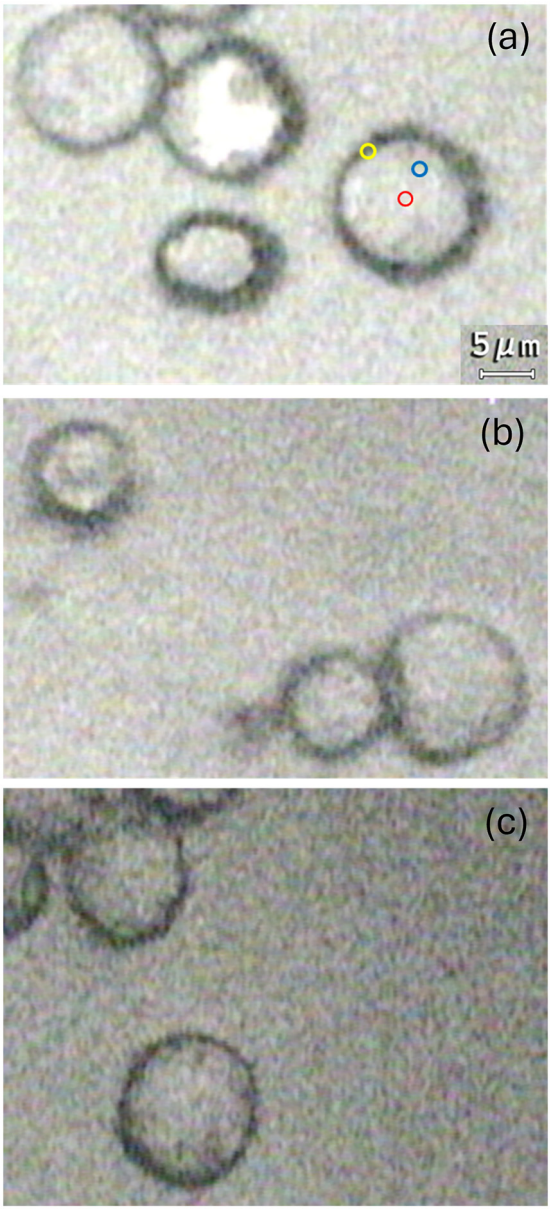

Fig. 1 Representative images of (a) NIH/3T3, (b) Primary fibroblast, and (c) MBM-T cells observed under the Raman microscope. The three measured areas are marked with circles: center (red), cytoplasm (blue), and edge (yellow).

Data preprocessing

Fig. S1a† presents typical raw Raman spectra of primary fibroblasts obtained from the Raman spectrometer. All acquired spectra were pre-processed before classification to enhance spectral quality, refine Raman shift bands, and facilitate comparison across different spectra. The pre-processing steps were implemented using our in-house Python code.

First, the spectra were cut to the 1800–600 cm^{-1} range and smoothed using the Savitzky–Golay algorithm (5-point window) to reduce instrumental noise and enhance spectral clarity. Next, baseline correction was applied to eliminate fluorescence-induced variations and spectral baseline shifts.

For baseline correction, each spectrum was divided into 64 equal-sized segments. The minimum y-value within each segment was identified, and these minima were used to fit a polynomial function representing the baseline. This polynomial was then subtracted from the original spectrum to obtain the baseline-corrected spectrum. The entire procedure was repeated five times to ensure optimal correction.

The final pre-processing step involved vector normalization. Each spectrum was treated as a vector, with the average intensity across all wavenumbers calculated and subtracted from the spectrum. The resulting spectrum was then normalized to a unit vector by computing the sum of the squared intensity values (*Y*-axis) and dividing by the square root of this sum. Since vector normalization can yield negative intensity values, all spectra were adjusted by shifting the minimum intensity to zero.

Machine learning analysis

This study examines classification among normal, precancerous, and cancerous cells, focusing on feature selection methods to identify the most relevant wavenumbers linked to malignancy progression.

The classification system

The analysis utilized the Raman spectra to enhance classification performance, as previous studies have demonstrated that feature selection improves data informativeness and, consequently, classifier accuracy.[40] The Raman spectra, consisting of 970 data points representing the wavenumbers in the 1800–600 cm^{-1} region, served as the initial feature vectors.

However, a substantial portion of these wavenumbers contributed little to no valuable information for classification. To refine the dataset and optimize classification performance, the ANOVA F-score was applied to these Raman spectra for feature selection.[41–43] This step is crucial for reducing data dimensionality while simultaneously enhancing classification accuracy.

Fig. 2 presents a comprehensive workflow of the machine learning pipeline, illustrating each stage—from feature extraction to model construction.

Validation

The classification system evaluates three binary tasks: normal *vs.* cancerous, normal *vs.* precancerous, and precancerous *vs.* cancerous, distinguishing cell types with high precision. As described in Fig. 1, three spectra were reordered from each cell; thus, the leave-one-group-out (LOGO) approach was adopted for validation. In this manner, each cell was treated as a set whose elements are the three spectra recorded by the Raman spectrometer. All spectra from a held-out cell were classified during validation, while the remaining cells trained the model, iterating until every cell was tested. A Log-Likelihood Ratio (LLR) decision framework aggregated predictions across a cell's three spectra, assigning a final label according to the used classifier: normal or precancerous; normal or cancerous; precancerous or cancerous, based on collective evidence.

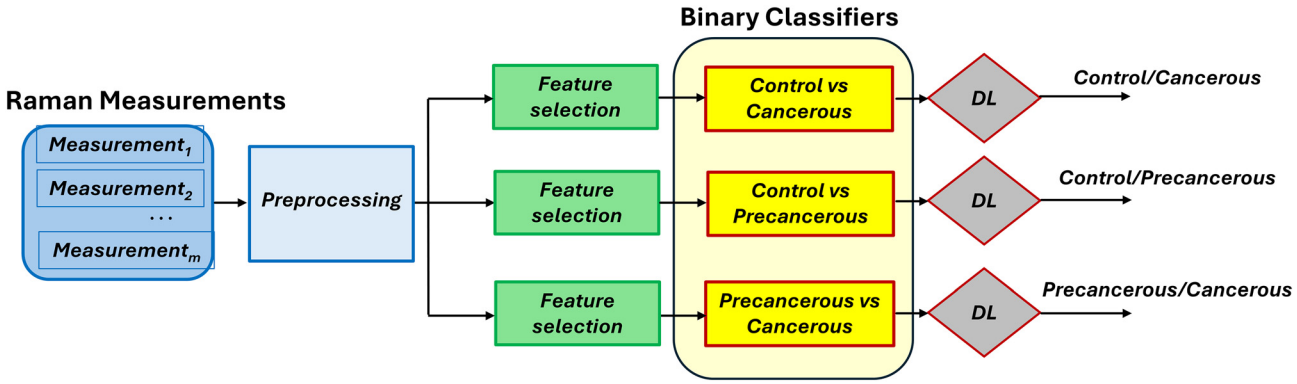

Fig. 2 Illustration of the system's training and testing process.

To further ensure robustness, 5-fold cross-validation was implemented in some of the classification experiments, enabling the estimation of the error as standard deviation. Here, cells—not individual spectra—were partitioned into five folds, ensuring no spectra from the same cell appear in both sets. We kept the same ratio of cells in each fold. In each iteration, four folds trained the model, while the fifth fold's cells were classified *via* the LLR decision system, which combines predictions from all three spectra of each validation cell to classify it.

By rotating the test fold across all partitions, the method generated an averaged performance metric with standard deviation, quantifying model stability.

This dual-validation strategy—LOGO for exhaustive per-cell assessment and 5-fold CV for error estimation—strengthened generalization while rigorously respecting the data's hierarchical structure.

Features selection

Using the ANOVA F-score approach, we identified the most diagnostically important spectral characteristics embedded in Raman spectra, revealing critical biomarkers for illness characterization.[44,45]

This approach assumes statistical independence between the features. It evaluates each feature's significance in differentiating between the different pairs,[46,47] ranking them from the most significant to lowest based on their F-scores. Higher-ranked features exhibit greater discriminatory power. The most significant 60 spectral features distinguishing between Controls and Precancerous, Controls and Cancerous, as well as Precancerous and Cancerous, are detailed in Table S1a.†

Since feature selection is critical for interpretability (explainable AI),[48] we prioritized features that reflect malignancy-driven biological alterations rather than unrelated variability. We also applied relative entropy as an alternative feature selection method to further validate our findings and compared the results with the ANOVA F-score approach (Table S1b†).

We repeat each classification procedure 20 times using a 5-fold cross-validation framework to evaluate classification performance, thus guaranteeing strong model generalization methodically and improving classification accuracy. The database was randomly divided into five folds each time, ensuring diverse training and testing sets. Different feature vector subsets were used each time, starting with the five most statistically significant wavenumbers and incrementally expanding the selection in five increments, up to 150 features.

The optimal feature subset for each classification task (Control *vs.* Cancerous, Control *vs.* Precancerous, and Precancerous *vs.* Cancerous) was selected through manual evaluation of feature importance rankings (Fig. 3a).

We plotted the accuracy *versus* the number of features in steps of 5 features. Fig. 3a shows the curve plateaus at 10 features for Control *vs.* Cancerous, 70 for Control *vs.* Precancerous, and 115 for Cancerous *vs.* Precancerous classification. Thus, we chose these values as the optimal feature numbers for their respective classifications. The standard deviation of accuracy was calculated and plotted in Fig. 3a. Moreover, in Fig. 3b, we compare the accuracy of the Logistic Regression (LR) classifier for the classification between Precancerous *vs.* Cancerous obtained using two approaches: (1) Log-Likelihood Ration (LLR)-based decision logic applied across three measurement sites (Fig. 1) and (2) classification performed separately for each site without decision logic.

For comparison, extra analysis was conducted using average spectra, where each cell was represented by the average Raman spectrum from the three sites shown in Fig. 1. The LR model was used to classify the two categories in each couple, with five-fold cross-validation.

As shown in Fig. 3b, the LR classifier achieved better performance using the first approach when the number of features was relatively small (10 features were chosen).

Fig. S2a and S2b† are similar to Fig. 3b but correspond to the classification of Normal *vs.* Cancerous and Normal *vs.* Precancerous, respectively.

Logistic regression (LR) classifier

The LR classifier was implemented as a linear classifier to classify samples into Normal, Precancerous, and Cancerous categories.[43,49–51] The loss function minimized the cross-entropy between the actual labels and the LR predictions to enhance classification accuracy.

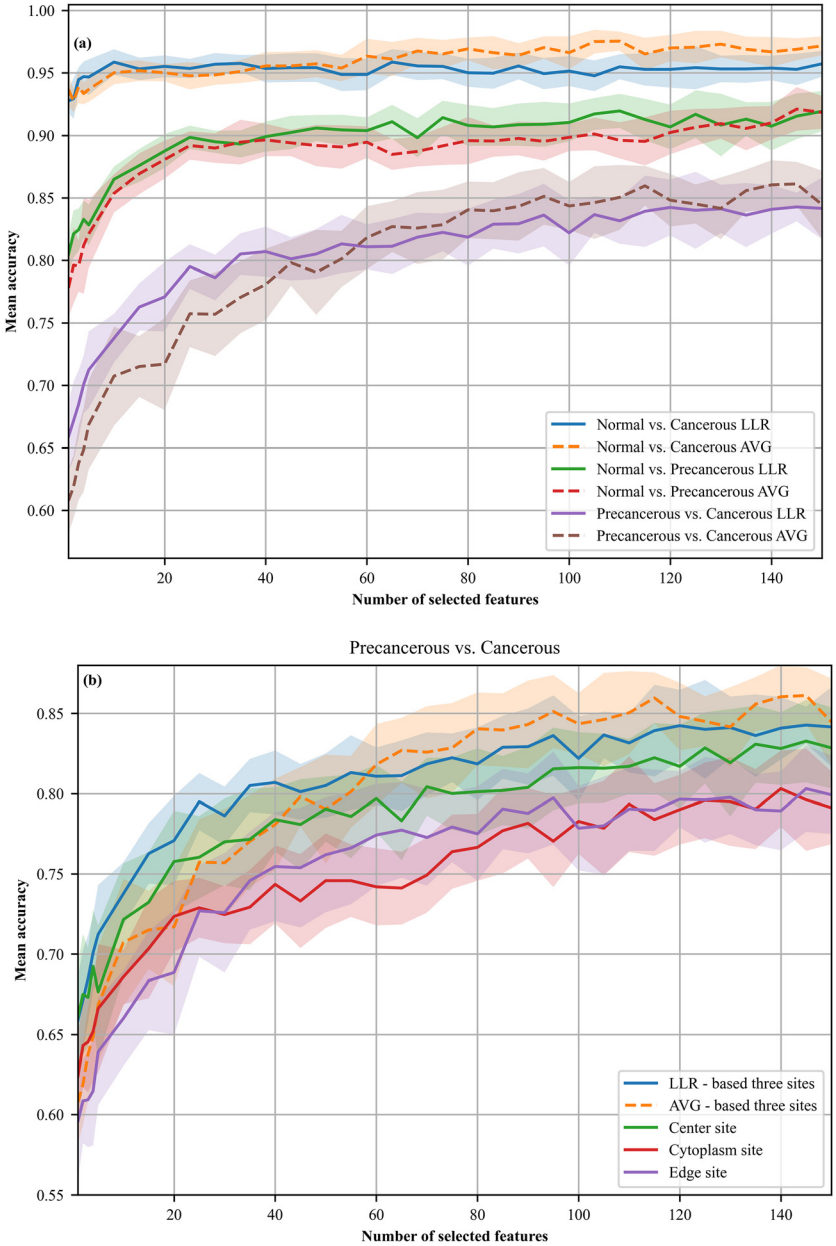

Fig. 3 (a) LR binary classification results for: Control *vs.* Cancerous, Control *vs.* Precancerous, and Precancerous *vs.* Cancerous. The analyses were conducted using individual spectra from the three measurement sites, incorporating the LLR DL method to classify samples as Precancerous *vs.* Cancerous. (b) LR binary classification for the Precancerous *vs.* Cancerous, comparing two approaches: (i) classification using LLR-based decision logic across three measurement sites and (ii) classification performed separately for each site without decision logic.

Decision logic system

Cell-level classification in this system is performed by a decision logic framework that interprets the classifier's output scores. For single-measurement samples, classification is straightforward, with labels assigned based on a predefined threshold.[52] However, in this study, each sample (cell) is represented by multiple measurements (three spectra), necessitating a decision considering all measurements.

To consolidate multiple measurements into a final classification, we explored the LLR,[53] which takes into account all the scores from all the measurements. LLR offers a refined approach by weighting predictions based on certainty, making it particularly valuable for high-precision classification tasks.

Evaluation

To assess the performance of the LR classifier, we applied the 5-fold cross-validation as described in the "Feature Selection" section, repeating the process twenty times. The feature vector sizes used were 10 features for Control *vs.* Cancerous, 70 for

Control *vs.* Precancerous, and 115 for Cancerous *vs.* Precancerous classification.

At the cell level, classification is determined using the LLR as a decision-logic approach, which relies on the classifier's scores for each spectrum within a given cell. This process is repeated for all cells in the dataset, with the final classification of each cell being based on the collective output of the classifier from its spectra. The performances of the binary classifier are summarized in a confusion matrix in Table 1.

When the classification was performed between couples of the three categories, the cancerous (MBM-T) category was determined to be the positive state in Normal–Cancerous and Precancerous–Cancerous. At the same time, when the classification was between Precancerous–Normal, the precancerous (NIH/3T3) category was determined as the positive state.

Statistical analysis

Statistical analysis in this study was conducted to evaluate the significance of spectral features and the performance of classification models applied to Raman spectral data from Normal, Precancerous, and Cancerous cells.

Evaluating the performance of the classifier

The performance of the classifiers was evaluated using several metrics: Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), Negative Predictive Value (NPV), and Positive Predictive Value (PPV).

Accuracy is the percentage of truly predicted both positive and negative states; Sensitivity is the percentage of actual positive states; Specificity is the percentage of truly predicted negative state samples out of actual negative state samples; PPV is the percentage of truly predicted positive state out all the samples predicted as positive by the classifier; and NPV is the percentage of truly predicted negative state out all the samples predicted as negative by the classifier.

t-Test and *P*-value

A *t*-test is a common hypothesis test used to compare the means of two groups and determine if they are significantly different from each other. It determines whether the compared categories are substantially different, whether observed variances are due to chance, or represent a meaningful difference.[54] The *p*-value represents the likelihood of obtaining outcomes as extreme as those observed under the null hypothesis.[55] A *p*-value of less than 0.05 denotes a statistically significant difference.

Table 1 Typical confusion matrix of a binary classifier obtained after validation

		Predicted	
		Positive	Negative
True	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Results and discussion

The Raman average spectra acquired from the cytoplasm, edge, and center sites of mouse embryo fibroblast cells as primary (normal) cells, NIH/3T3 (precancerous), and MBMT (cancerous) are presented in Fig. 4a, covering the 1800–600 cm^{-1} fingerprint region. This spectral range provides key biochemical information related to DNA, lipids, proteins, and nucleic acids, as detailed in Table S2,† which lists all prominent Raman bands based on published literature.[56–59] The major peaks are labeled in Fig. 4. Each peak in the Raman spectrum corresponds to specific vibrational modes of functional groups in key biomolecules, offering a detailed molecular fingerprint.[60] As shown in Fig. 4a, the Raman spectra of the three biological systems exhibit high similarity, with some differences in Raman intensities and subtle variations in spectral shape across certain regions. To emphasize these spectral changes, difference spectra (Δ) were calculated and plotted in Fig. 4b, representing Normal–Cancerous (blue line), Normal–Precancerous (orange line), and Precancerous–Cancerous (green line) comparisons.

However, not all spectral differences between the tested cell types directly reflect compositional and biochemical changes associated with cancer progression. Spectral variations can arise from two main sources: inter-variance, which represents true biological differences between Normal, Precancerous, and Cancerous states due to pathological abnormalities, and intra-variance, which includes both biological variability and technical factors within each group (not relevant to malignant transformation), such as batch-to-batch variations.[61] Additionally, due to the spatial resolution of Raman, different organelles and components could be measured depending on the measured site; these variations in the same cell types are also considered as intra-variance.

Therefore, careful feature selection is essential to isolate biomarkers linked to malignant transformation.

By correlating these wavenumbers with their corresponding biomolecules, we aim to identify the underlying biological changes associated with malignancy, relating spectral features to molecular alterations that may drive cancer development and offering insights into key biochemical processes underlying tumorigenesis.[48] A key advantage of feature selection over dimensionality reduction methods such as PCA, UMAP, and Diffusion Map is its interpretability, allowing for a more direct biological understanding of the spectral variations.[62,63]

Our approach enhances diagnostic accuracy and interpretability by selecting the most informative spectral features, ensuring they reflect malignancy-driven alterations rather than unrelated variability. This framework strengthens the reliability of spectral classification, providing deeper insights into key biochemical processes involved in cancer progression.[49,64–66] In contrast, other methods do not give specific information regarding the contribution of specific wavenumbers to the classification. The information derived using these methods is spread across the entire spectroscopic range.

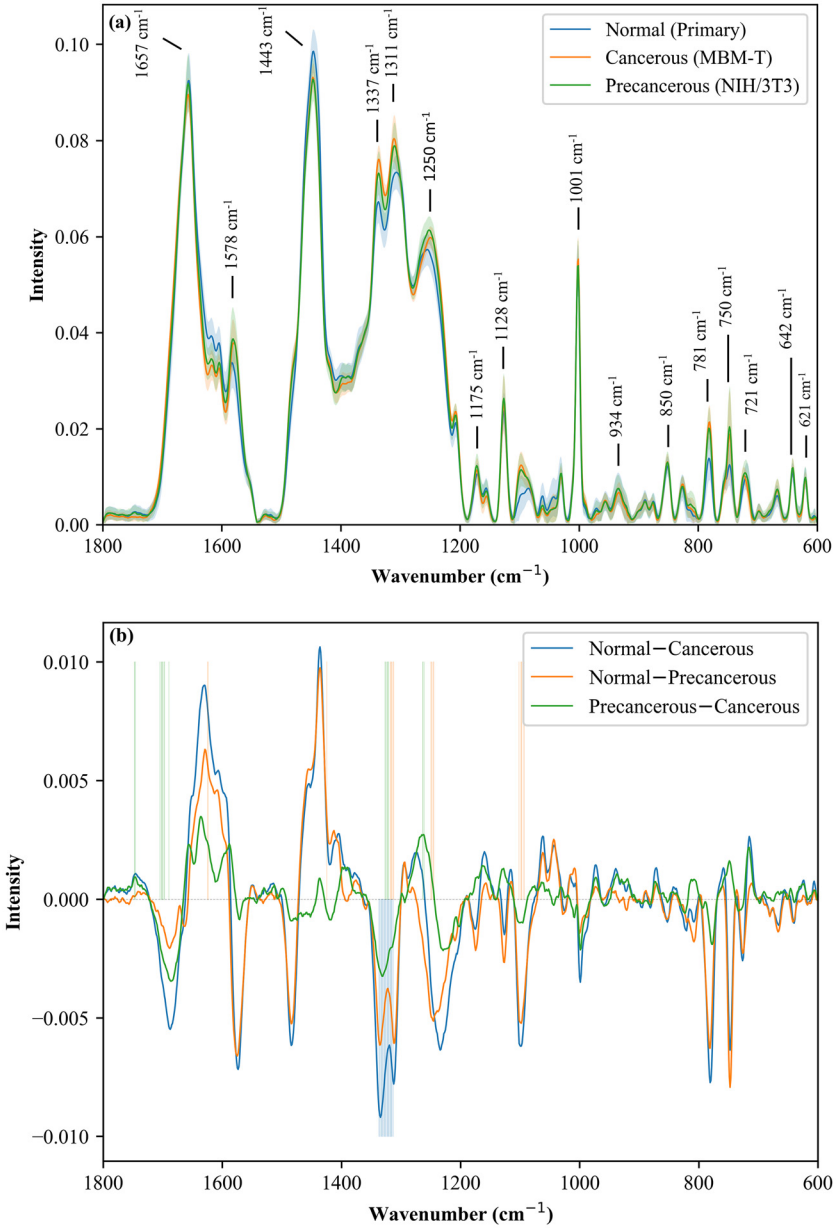

Fig. 4 (a) Average Raman spectra in the 1800–600 cm^{-1} region from Normal (primary), Precancerous (NIH/3T3), and Cancerous (MBM-T) cells measured from the three sites: cytoplasm, edge, and center. (b) Difference spectra (Δ) for Normal–Cancerous, Normal–Precancerous, and Precancerous–Cancerous comparisons. The top twenty discriminative features (Table S1†) are marked in blue, orange, and green shading.

Table 2 Metrics derived from the top 20 selected wavenumbers, along with major contributors to the absorption at these wavenumbers. Each metric was calculated as the sum of the absorption intensities at the corresponding wavenumbers

Pair	Metric	Wavenumbers (cm^{-1})	Major contributed molecules	*p*-Value
Normal–Cancer	I	1337, 1336, 1334, 1333, 1332, 1331, 1329, 1328, 1327, 1326, 1324, 1323, 1322, 1321, 1320, 1318, 1317, 1316, 1315, 1313	Lipid/protein/nucleic acids	1.2×10^{-36}
Normal–Precancer	II	1625, 1624	Amide I	1.6×10^{-22}
	III	1425	Deoxyribose	2.9×10^{-22}
	IV	1318, 1317, 1316, 1315, 1313,1312	Lipid/protein/nucleic acids	2.5×10^{-26}
	V	1250, 1249, 1248, 1246, 1245	Amide III, guanine, cytosine	2.0×10^{-23}
	VI	1102, 1099, 1098, 1097, 1094, 1093	PO_2^-/DNA/lipids	4.4×10^{-22}
Precancer–Cancer	VII	1748, 1747, 1746	Lipids/phospholipids	1.6×10^{-8}
	VIII	1705, 1703, 1702, 1701, 1700, 1698, 1697, 1690	amino acids aspartic & glutamic acid/Amide I	1.2×10^{-8}
	IX	1327, 1326, 1324, 1323, 1322, 1321	Proteins/nucleic acids	4.1×10^{-9}
	X	1264, 1263, 1261, 1260	Amide III	2.2×10^{-8}

To demonstrate the efficiency of the selected feature techniques, we employed the *t*-test to compare the averages of predefined metrics between two groups of the compared categories and evaluate whether they differed substantially.

A *t*-test determines whether the compared categories are substantially different, whether observed variances are due to chance, or represent a meaningful difference.[54] The *p*-value represents the likelihood of obtaining outcomes as extreme as those observed under the null hypothesis. A *p*-value of less than 0.05 denotes a statistically significant difference.

Before performing *t*-test analysis, Raman spectra acquired from different subcellular regions (center, cytoplasm, membrane) were rigorously averaged within each individual cell to avoid pseudo-replication. This process generated one representative spectrum per cell, resulting in 222 independent cell-level data points and ensuring that each biological replicate contributed a single, independent measurement to the analysis.

We applied a *t*-test to compare the two category pairs based on different metrics, as presented in Table 2: Normal–Cancerous, Normal–Precancerous, and Precancerous–Cancerous.

The number of metrics defined for each pair depends on the dispersion of these top 20 wavenumbers across the 1800–600 cm^{-1} region (Fig. 4b). These predefined metrics are calculated as the sum of absorption intensities at the corresponding wavenumbers within each metric and are associated with their specific vibrational modes and contributing molecules (Table 2).

A significant difference between the two averages suggests a substantial alteration in the contents of the corresponding molecules. For example, for the "Normal–Cancer" pair, we summed the absorption intensities at the wavenumbers corresponding to metric I (Table 2). The resulting *p*-value indicated a statistically significant differentiation between the normal and cancer groups. Furthermore, the observed alteration in lipid, protein, and nucleic acid contents are significant in the transformation from normal to cancerous state.

We presented the *t*-test results as a violin plot based on each metric in Fig. 5. As seen in Table 2, significant differences were observed between the averages of the compared groups in all the metrics, in the specific wavenumber and corresponding biomolecules. The significant differences observed in all the evaluated metrics confirm the effectiveness of the selected feature selection techniques in distinguishing between the compared categories. The identified spectral features and their corresponding biomolecules provide biologically relevant insights into the underlying differences between the groups. These findings reinforce the importance of feature selection in enhancing classification accuracy and improving the interpretability of spectral-based diagnostics.

The ANOVA F-score analysis identified the most informative spectra features, effectively distinguishing among the compared categories in three pairs of the groups: Cancerous *vs.* Normal, Precancerous *vs.* Normal, and Cancerous *vs.* Precancerous (Table S1†). These key features, visually marked in Fig. 4b with blue, orange, and green shading, enhance early cancer detection and understanding of tumorigenesis.

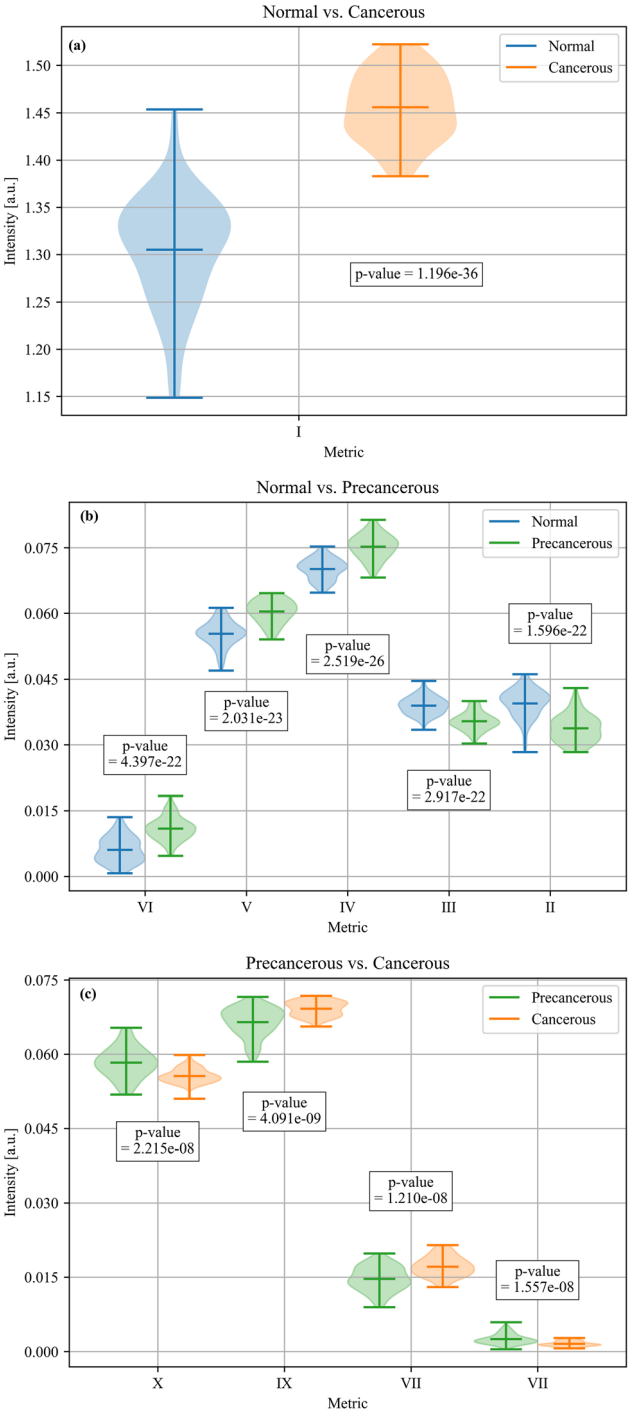

Fig. 5 *t*-Test statistical calculations presented as Violin plots for: Normal–Cancer (a), Normal–Precancer (b), and Precancer–cancer (c) based on each metric. The calculation was performed for the I–X metrics defined in Table 2. For each spot appears, three horizontal lines, the middle line is the average, and the upper and lower lines represent the minimum and maximum of the calculated metrics. The shadowed plot is the kernel density distribution estimate of the metrics values.

The transition from normal to cancerous states involves progressive molecular changes that can be detected using Raman spectroscopy.[67] We systematically analyze molecular

changes based on the best 20 selected features associated with precancerous transformation and the transition between these states (Table 3).

We correlated the top 20 selected spectral features (wavenumbers) with the functional groups of the biomolecules that compose the cells, identifying specific biochemical alterations associated with malignant transformation based on literature (Table 3).

These spectral markers reflect key molecular changes, including lipid membrane remodeling, shifts in protein secondary structures, and alterations in nucleic acid composition. The observed spectral variations suggest disruptions in lipid saturation, indicative of altered membrane fluidity and cellular signaling in tumorigenesis. Additionally, protein-related features point to changes in β-sheet and α-helix structures linked to cytoskeletal remodeling and protein folding dynamics. Moreover, the neoplastic cells generate more lactate than healthy cells,[80,81] while spectral bands corresponding to nucleic acids highlight transcriptional and epigenetic alterations characteristic of precancerous and cancerous states (reference). The upregulation of specific Raman bands associated with oxidative stress suggests an imbalance in cellular redox homeostasis, a hallmark of cancer progression. Systematically mapping these spectral features to biochemical processes provides deeper insight into the molecular events driving tumorigenesis. This approach enhances our ability to differentiate between normal, precancerous, and cancerous states, reinforcing the potential of Raman spectroscopy for early cancer detection and classification.

Cancerous and precancerous cells express oncoproteins that resemble normal cytoplasmic proteins, disrupting DNA–protein interactions and modifying nuclear proteins involved

Table 3 Top 20 Raman spectral features *via* ANOVA-based selection: correlating functional group vibrations with biochemical alterations in Normal, Precancerous, and Cancerous cells

Wavenumber (cm^{-1})	Vibrational mode	Associated biomolecules	Cell type comparison	Biochemical significance
1690–1698	C═C stretch	Lipid	***	● **Indicates** altered lipid metabolism, with increased unsaturated fatty acids supporting cell division, membrane fluidity, and metastasis.[68]
	Amide I C═O stretch	Protein		● **Protein** structure changes (misfolding, aggregation) are linked to disrupted proteostasis, oxidative stress, and oncogenic activation.[69] (see 1687–1680 for more on protein aggregation and structural shifts.)
1705–1700	C═C OH	Amino acids aspartic & glutamic acid	***	May reflect altered amino acid metabolism, protein modifications, and microenvironmental shifts, supporting tumor growth and progression[70,71]
1748–1746	C═C stretch	Lipid	***	lipid changes in cancer progression, including increased unsaturation, membrane remodeling, enhanced lipid synthesis, oxidative stress, and altered signaling, all supporting tumor growth and metastasis.[72]
1624, 1625	Amide I C═O stretch	Protein	**	Protein misfolding and post-translational modifications (glycosylation, oxidation) are associated with oxidative stress and the transition from normal to precancerous stages.[73]
1337–1313, 1327–1321	CH_2/CH_3 deformation and torsion	Lipids	*, ***	Increased intensity in cancer cells suggests altered lipid metabolism and membrane composition, supporting rapid cell division and higher membrane fluidity, critical for tumor growth.[68]
1321	CH_2 bending	Lipid–protein interactions	*	May reflects altered lipid–protein interactions in cancer cell membranes, contributing to membrane dynamics essential for tumor growth and metastasis.[74]
1318–1312	CH_2/CH_3 bending vibrations	Lipids and proteins	**	Increased intensity in precancerous cells suggests lipid raft formation, which is involved in signaling pathways that promote proliferation, survival, and membrane remodeling.[75]
1264–1261	PO_2^- stretch	Nucleic acids	***	● **Nucleic acids**: Extensive modifications in DNA, including mutations, chromosomal instability, and increased replication, driving tumor growth.[76]
	CH_2/CH_3 Deformation	Lipids/proteins		● **Lipids**: Metabolism alterations and membrane remodeling enhance oncogenic signaling and metastasis.[77]
1250–1245	C–C stretching	Proteins/lipids	**	Alterations in lipid metabolism and raft formation are critical in early cancer progression. In precancerous cells, membrane changes facilitate oncogenic signaling.[77]
1100	C–N stretch	Proteins/nucleic acids	**	Reflects protein backbone modifications and nucleic acid integrity changes (*e.g.*, DNA methylation, oxidative damage), common in early-stage mutations.[73,78]
1102–1093	PO_2^- stretch	Nucleic acids	**	Shifts and intensity changes correspond to genomic instability (epigenetic changes, mutations), marking early DNA disruptions linked to cancer initiation.[79]

Normal–Cancer (*); Normal–Precancerous (**); Precancerous–Cancerous (***).

in cell division and DNA replication, resulting in uncontrolled proliferation.[81]

Genes associated with cell cycle regulation (*e.g.*, *CCNB1*, *MCM3*, *MCM4*, *MCM7*) and oxidative phosphorylation (*e.g.*, *ATP5B*, *ATP5G3*) are upregulated in precancerous and cancerous cells, leading to increased protein levels.[82] These molecular changes correspond to specific Raman spectral shifts, particularly in the Amide II and Amide III regions at 1250 and 1267 cm^{-1}.[83]

The spectral differences between the abnormal and normal categories are larger than the spectral differences between the cancerous and precancerous categories, as seen in Table S1,† Fig. 2 and 3. This trend is further supported by the *p*-values in Table 2, where the Normal–Cancer comparison exhibits significantly lower *p*-values than the Normal–Precancerous and Precancerous–Cancerous comparisons. This is not surprising and makes sense from a biological point of view. As mentioned above, the precancerous and cancerous biological systems have many similar properties and characteristics. As is known, the precancerous (NIH/3T3) cells have undergone many changes due to various mutations throughout the multiple transfers. However, they are still not considered cancerous cells because they do not have all the properties of a cancerous cell.[84]

The classification analyses in all the classification experiments were based on selected features derived from Raman spectra by ANOVA F-score in the 1800–900 cm^{-1}. The classification was performed separately as a binary classification between all the different pairs: Normal–Cancerous, Normal–Precancerous, and Precancerous–Cancerous.

In this analysis, leave-one-group-out cross-validation (LOGOCV) was used to generate the receiver operating characteristic (ROC) curve to optimize the classification operating point (threshold) and to estimate the classifier's performance as the area under the curve (AUC) of the ROC curve. Each group consisted of three measurements taken from the center, cytoplasm, and edge of the same cell (Fig. 1).

For example, when the classification is between the normal and cancerous categories, the ROC curve evaluates the tests' accuracy quantitatively in terms of correct determination for a certain sample as normal or cancerous by calculating the AUC of the ROC curve.

Fig. 6 presents the LR model's ROC curves and the operating points for the binary classification: Normal *vs.* Cancerous, Normal *vs.* Precancerous, and Precancerous *vs.* Cancerous. The LOGOCV was used at the spectrum level, while the sample category was determined using LLR.

The LR classification's performance for the discrimination between the different categories of each pair: Normal–Cancerous, Normal–Precancerous, and Precancerous–Cancerous is summarized in Table 4.

As shown in Fig. 3, S2a, S2b,† and Table 4, the performance of the LR logic across three measurement sites, incorporating the LLR DL method for classifying samples as Precancerous *vs.* Cancerous, surpasses the performance obtained when each site is analyzed separately without decision logic.

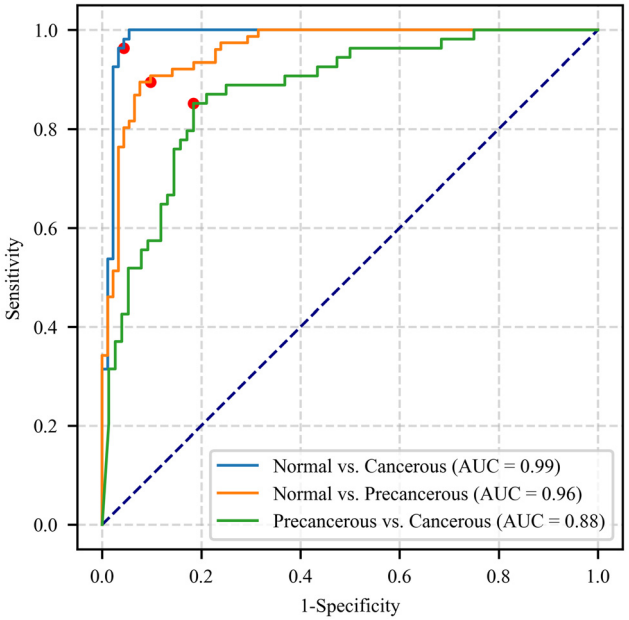

Fig. 6 ROC curves for classifying the binary classification: Normal *vs.* Cancerous, Normal *vs.* Precancerous, and Precancerous *vs.* Cancerous. The LOGOCV was used at the spectrum level, while the category of the sample was determined using LLR. The operating points are labeled as red points.

Therefore, for future studies, acquiring spectra from additional sites and analyzing cells as a group using spectra from all sites is recommended. This approach involves evaluating data at the spectrum level and applying a decision logic method to determine the classification at the cell level.

The results presented in Table 4 demonstrate the powerful capability of Raman spectroscopy machine learning for excellent differentiation between the normal and precancerous or cancerous cells, based on the changes in the cells' biomolecules, with 91% and 95.1% accuracy, respectively. In addition, this study shows the high potential of this method to differentiate between precancerous and cancerous cells with 90% accuracy, where the spectral differences between precancerous and cancerous categories are minute. This finding is significant since detecting precancerous cells before developing cancerous cells is critical for the effective prevention/treatment of cancer development.

The data is unbalanced and was threatened by training the LR with a weighted loss to emphasize the smaller class. This is a widely used strategy to overcome the unbalanced data problem.[85]

When comparing classifier performance using ANOVA F-score selected features with and without class-weighted (Columns I and II), we observe no significant advantage from imbalance adjustment (Column II). The performance metrics (AUC, Accuracy, Sensitivity, Specificity, PPV, and NPV) remain statistically equivalent within their respective error margins.

This finding aligns with expectations, as our dataset's minority class representation (24%) substantially exceeds the

Table 4 LR classifier performance in discriminating: (a) Normal–Cancerous, (b) Normal–Precancerous, and (c) Precancerous–Cancerous. Analysis used single-site measurements and averaged spectra (trivial threshold) or multi-site LLR fusion. Comparison between LR with and without class-weighted (to address imbalance). Features were selected *via* ANOVA-F and relative entropy

	AUC			Accuracy (%)			Sensitivity (%)			Specificity (%)			PPV (%) Positive Predictive Value			NPV (%) Negative Predictive Value		
(a) Normal–Cancerous classes (10 selected features)																		
Site	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
Three sites	0.99 ± 0.00	0.99 ± 0.01	0.99 ± 0.00	95.9 ± 1.0	96.3 ± 0.8	96.9 ± 0.9	93.4 ± 2.5	95.8 ± 1.6	96.9 ± 1.6	97.3 ± 0.5	96.6 ± 0.8	96.8 ± 0.8	95.3 ± 0.9	94.4 ± 1.3	94.8 ± 1.2	96.2 ± 1.4	97.5 ± 1.0	98.1 ± 0.9
Center	0.93 ± 0.01	0.94 ± 0.01	0.96 ± 0.01	93.6 ± 1.1	94.2 ± 1.2	95.8 ± 1.2	91.2 ± 2.3	94.5 ± 1.9	94.8 ± 2.7	94.9 ± 1.3	94.0 ± 1.6	96.3 ± 1.2	91.4 ± 1.9	90.3 ± 2.4	93.8 ± 1.9	94.9 ± 1.3	96.7 ± 1.1	97.0 ± 1.5
Cytoplasm	0.93 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	93.8 ± 1.2	94.3 ± 1.1	93.8 ± 1.1	91.9 ± 2.2	94.6 ± 2.0	94.4 ± 1.9	94.7 ± 1.0	94.2 ± 1.2	93.5 ± 1.2	91.4 ± 1.7	90.6 ± 1.7	89.6 ± 1.7	95.3 ± 1.3	96.8 ± 1.2	96.6 ± 1.1
Membrane	0.92 ± 0.02	0.92 ± 0.02	0.92 ± 0.01	92.3 ± 1.5	92.4 ± 1.4	91.6 ± 1.1	88.8 ± 2.4	91.5 ± 2.8	91.9 ± 2.0	94.4 ± 2.0	92.9 ± 1.5	91.5 ± 1.7	90.4 ± 3.1	88.4 ± 2.2	86.4 ± 2.3	93.5 ± 1.3	94.9 ± 1.6	95.1 ± 1.1
Av. Sp.	0.98 ± 0.00	0.98 ± 0.01	0.98 ± 0.00	95.0 ± 0.9	94.9 ± 1.1	96.5 ± 0.7	93.1 ± 2.1	94.0 ± 1.8	96.2 ± 1.8	96.1 ± 1.0	95.4 ± 1.0	96.7 ± 0.2	93.4 ± 1.6	92.3 ± 1.6	94.5 ± 0.4	95.9 ± 1.2	96.4 ± 1.1	97.8 ± 1.0
(b) Normal–Precancerous (70 selected features)																		
Site	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
Three sites	0.95 ± 0.02	0.96 ± 0.01	0.97 ± 0.01	89.8 ± 1.6	91.1 ± 1.9	91.4 ± 1.3	87.6 ± 2.4	89.7 ± 2.3	89.6 ± 2.1	91.6 ± 1.6	92.2 ± 2.1	92.8 ± 1.5	89.7 ± 1.9	90.5 ± 2.4	91.2 ± 1.7	91.0 ± 1.8	91.5 ± 1.8	91.6 ± 1.6
Center	0.87 ± 0.02	0.87 ± 0.01	0.88 ± 0.02	86.9 ± 1.5	87.3 ± 1.1	88.7 ± 1.5	83.1 ± 2.9	84.9 ± 1.8	86.6 ± 2.2	90.1 ± 1.6	89.2 ± 1.7	90.4 ± 2.1	87.4 ± 1.8	86.7 ± 1.8	88.2 ± 2.3	87.2 ± 2.0	87.7 ± 1.3	89.1 ± 1.6
Cytoplasm	0.87 ± 0.02	0.87 ± 0.01	0.91 ± 0.02	87.6 ± 1.7	87.5 ± 1.3	90.6 ± 1.7	85.1 ± 2.9	86.1 ± 2.2	90.0 ± 2.5	89.7 ± 1.9	88.6 ± 1.6	91.1 ± 1.8	87.2 ± 2.1	86.3 ± 1.7	89.3 ± 2.0	89.0 ± 1.8	88.6 ± 1.6	91.7 ± 1.9
Membrane	0.84 ± 0.01	0.85 ± 0.02	0.89 ± 0.02	84.5 ± 1.4	85.0 ± 2.0	89.0 ± 1.8	79.7.6 ± 2.3	81.9 ± 2.9	87.6 ± 3.6	88.5 ± 1.8	87.6 ± 1.9	90.1 ± 1.6	85.1 ± 2.0	84.5 ± 2.2	88.0 ± 1.7	84.1 ± 1.5	85.4 ± 2.1	89.8 ± 2.6
Av. Sp.	0.96 ± 0.01	0.96 ± 0.01	0.97 ± 0.01	88.7 ± 1.2	88.9 ± 1.5	92.7 ± 1.0	85.7 ± 2.5	87.3 ± 2.1	91.3 ± 1.5	91.2 ± 1.8	90.2 ± 1.7	94.0 ± 1.2	89.0 ± 1.9	88.1 ± 1.9	92.6 ± 1.4	88.6 ± 1.7	89.6 ± 1.6	92.9 ± 1.1
(c) Precancerous–Cancerous (115 selected features)																		
Site	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
Three sites	0.89 ± 0.03	0.89 ± 0.01	0.87 ± 0.02	83.9 ± 2.9	84.4 ± 2.0	82.0 ± 2.3	82.2 ± 4.2	83.1 ± 3.0	79.9 ± 3.8	85.1 ± 3.3	85.3 ± 2.9	83.6 ± 3.3	79.8 ± 3.9	80.2 ± 3.2	77.7 ± 3.5	87.1 ± 2.8	87.7 ± 1.9	85.5 ± 2.3
Center	0.82 ± 0.02	0.83 ± 0.02	0.80 ± 0.03	82.2 ± 2.2	82.7 ± 2.0	80.0 ± 2.9	79.5 ± 3.1	81.5 ± 3.8	78.1 ± 4.7	84.1 ± 2.9	83.6 ± 2.1	81.3 ± 3.7	78.2 ± 3.4	78.0 ± 2.4	74.9 ± 3.8	85.3 ± 1.9	86.5 ± 2.4	84.0 ± 2.9
Cytoplasm	0.78 ± 0.02	0.78 ± 0.02	0.75 ± 0.03	78.4 ± 2.3	78.2 ± 2.0	75.1 ± 2.6	76.2 ± 3.6	78.1 ± 2.6	73.1 ± 4.6	79.9 ± 2.8	78.3 ± 3.2	76.4 ± 3.2	73.0 ± 3.1	72.0 ± 3.0	68.9 ± 3.0	82.6 ± 2.3	83.4 ± 1.6	80.1 ± 2.8
Membrane	0.79 ± 0.03	0.79 ± 0.01	0.75 ± 0.02	79.0 ± 2.5	78.6 ± 1.5	75.6 ± 2.3	77.6 ± 4.0	79.1 ± 2.3	73.7 ± 3.6	79.9 ± 3.2	78.2 ± 3.2	77.0 ± 3.2	73.4 ± 3.3	72.2 ± 2.5	69.6 ± 3.1	83.4 ± 2.5	84.1 ± 1.2	80.5 ± 2.2
Av. Sp.	0.92 ± 0.02	0.92 ± 0.02	0.89 ± 0.02	86.0 ± 1.6	84.8 ± 2.5	82.3 ± 2.3	83.7 ± 2.6	84.0 ± 2.6	82.6 ± 3.9	87.6 ± 2.1	85.5 ± 3.2	82.2 ± 3.4	82.8 ± 2.5	80.5 ± 3.7	76.8 ± 3.4	88.3 ± 1.6	88.2 ± 1.9	87.0 ± 2.5

Cells: Normal (n = 92), Precancerous (NIH/3T3, n = 76), Cancerous (MBM-T, n = 54). **Methods**: (I) ANOVA F-score (no reg.), (II) ANOVA F-score + class-weighted (to address imbalance), (III) Relative entropy + class-weighted (to address imbalance). The averaged spectrum across the three measurement sites (Av Sp).

10–15% threshold where class imbalance typically impairs performance.[86,87] Existing literature demonstrates that class-weighted LR with feature selection maintains robustness for minority classes ≥20%,[85] which supports our observed metric stability (Table 4).

The consistent outperformance of average spectra analysis and multi-site LLR fusion over single-site methods across all classification pairs (Normal–Cancerous, Normal–Precancerous, Precancerous–Cancerous) and performance metrics stems from noise reduction, enhanced statistical power, and compensation for site-specific variability. Integrating data across sites improves signal-to-noise ratio, increases generalizability, and mitigates biases, leading to more robust and biologically meaningful classification. This aligns with established data fusion principles and confirms their superiority for diagnostic applications requiring high reliability.

Our comparative analysis reveals that multi-site LLR fusion demonstrates superior performance for Normal–Cancerous and Normal–Precancerous classification by leveraging discriminative weighting of pronounced spectral differences. In contrast, average spectra analysis shows marginally better, though error-bound, results for Precancerous-Cancerous discrimination.

Comparing the performance of the logistic regression (LR) classifier using features selected by ANOVA F-score and relative entropy (columns II and III), we find that the key metrics—AUC, accuracy, sensitivity, specificity, PPV, and NPV—are statistically equivalent within their respective error margins. This alignment is expected, as Table S2b† shows that both methods identify nearly the same 60 features, differing primarily in their ranking order.

Table 4 and Fig. 6 demonstrate that the selected features enable clear discrimination between the two compared categories in each pair.

For decades, scientists have been looking for distinct features that can help them to discriminate between a cancerous cell and a normal cell. Our findings in this study support the possible use of this spectroscopic method to detect precancerous and cancerous cells early.

Conclusion

This study emphasizes the critical necessity to characterize the precancerous stage, which has been largely overlooked in Raman spectroscopy studies. This study uses Raman spectroscopy, ANOVA-based feature selection, and log-likelihood ratio decision logic to distinguish between normal, precancerous, and cancerous cells systematically. The findings indicate different spectrum indicators associated with essential biological changes in cancer progression, such as changes in nucleic acid contents and protein structures. Combining various analytical methods improves classification accuracy and provides a deeper understanding of the chemical alterations that drive cancer development. Combining Raman spectroscopy with feature selection methods and machine learning creates a powerful diagnostic tool capable of characterizing and accurately diagnosing the precancerous stage. Early cancer diagnosis has the potential to revolutionize public health, profoundly transform public health, elevate the quality of life, and drive significant economic benefits.

Conflicts of interest

There is no conflict of interest.

Data availability

The data and code supporting this study are publicly available at: **https://github.com/Biomedical-Diagnosis-SCE/RamanAI-Cancer-Precancer-Detect**.

References

1 M. D. Keller, E. M. Kanter and A. Mahadevan-Jansen, Raman spectroscopy for cancer diagnosis, *Spectroscopy*, 2006, **21**(11), 33.
2 K. D. Miller, L. Nogueira, A. B. Mariotto, J. H. Rowland, K. R. Yabroff, C. M. Alfano, A. Jemal, J. L. Kramer and R. L. Siegel, Cancer treatment and survivorship statistics, 2019, *CA-Cancer J. Clin.*, 2019, **69**(5), 363–385.
3 H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA-Cancer J. Clin.*, 2021, **71**(3), 209–249.
4 K.-l. Edward, M. Chipman, J.-A. Giandinoto and K. Robinson, Quality of life and personal resilience in the first two years after breast cancer diagnosis: systematic integrative review, *Br. J. Nurs.*, 2019, **28**(10), S4–S14.
5 R. M. Kaplan, The significance of quality of life in health care, *Qual. Life Res.*, 2003, **12**(1), 3–16.
6 P. M. Das and R. C. Bast Jr, *Early detection of ovarian cancer*, 2008.
7 C. Hamashima, Current issues and future perspectives of gastric cancer screening, *World J. Gastroenterol.*, 2014, **20**(38), 13767.
8 C. A. Lieber, S. K. Majumder, D. D. Billheimer, D. L. Ellis and A. Mahadevan-Jansen, Raman microspectroscopy for skin cancer detection in vitro, *J. Biomed. Opt.*, 2009, **13**(2), 024013.
9 D. Chaturvedi, S. A. Balaji, V. K. Bn, F. Ariese, S. Umapathy and A. Rangarajan, Different phases of breast cancer cells: Raman study of immortalized, transformed, and invasive cells, *Biosensors*, 2016, **6**(4), 57.
10 C. Krafft, S. B. Sobottka, G. Schackert and R. Salzer, Near infrared Raman spectroscopic mapping of native brain tissue and intracranial tumors, *Analyst*, 2005, **130**(7), 1070–1077.

11 I. P. Santos, E. M. Barroso, T. C. B. Schut, P. J. Caspers, C. G. van Lanschot, D.-H. Choi, M. F. Van Der Kamp, R. W. Smits, R. Van Doorn and R. M. Verdijk, Raman spectroscopy for cancer detection and cancer surgery guidance: translation to the clinics, *Analyst*, 2017, **142**(17), 3025–3047.
12 A. Downes, Raman microscopy and associated techniques for label-free imaging of cancer tissue, *Appl. Spectrosc. Rev.*, 2015, **50**(8), 641–653.
13 I. Pence and A. Mahadevan-Jansen, Clinical instrumentation and applications of Raman spectroscopy, *Chem. Soc. Rev.*, 2016, **45**(7), 1958–1979.
14 M. Jermyn, J. Desroches, K. Aubertin, K. St-Arnaud, W.-J. Madore, E. De Montigny, M.-C. Guiot, D. Trudel, B. C. Wilson and K. Petrecca, A review of Raman spectroscopy advances with an emphasis on clinical translation challenges in oncology, *Phys. Med. Biol.*, 2016, **61**(23), R370–R400.
15 L. A. Austin, S. Osseiran and C. L. Evans, Raman technologies in cancer diagnostics, *Analyst*, 2016, **141**(2), 476–503.
16 S. Cui, S. Zhang and S. Yue, Raman spectroscopy and imaging for cancer diagnosis, *J. Healthcare Eng.*, 2018, **2018**, 8619342.
17 S. Kumar, N. Matange, S. Umapathy and S. S. Visweswariah, Linking carbon metabolism to carotenoid production in mycobacteria using Raman spectroscopy, *FEMS Microbiol. Lett.*, 2015, **362**(3), 1–6.
18 P. Crow, B. Barrass, C. Kendall, M. Hart-Prieto, M. Wright, R. Persad and N. Stone, The use of Raman spectroscopy to differentiate between different prostatic adenocarcinoma cell lines, *Br. J. Cancer*, 2005, **92**(12), 2166–2170.
19 H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, M. R. McAinsh, N. Stone and F. L. Martin, Using Raman spectroscopy to characterize biological materials, *Nat. Protoc.*, 2016, **11**(4), 664–687.
20 M. J. Baker, H. J. Byrne, J. Chalmers, P. Gardner, R. Goodacre, A. Henderson, S. G. Kazarian, F. L. Martin, J. Moger, N. Stone and J. Sulé-Suso, Clinical applications of infrared and Raman spectroscopy: state of play and future challenges, *Analyst*, 2018, **143**(8), 1735–1757.
21 C. García–Timermans, P. Rubbens, J. Heyse, F. M. Kerckhof, R. Props, A. G. Skirtach, W. Waegeman and N. Boon, Discriminating bacterial phenotypes at the population and single–cell level: a comparison of flow cytometry and Raman spectroscopy fingerprinting, *Cytometry, Part A*, 2020, **97**(7), 713–726.
22 M.-J. Jeng, M. Sharma, L. Sharma, T.-Y. Chao, S.-F. Huang, L.-B. Chang, S.-L. Wu and L. Chow, Raman spectroscopy analysis for optical diagnosis of oral cancer detection, *J. Clin. Med.*, 2019, **8**(9), 1313.
23 N. A. Correia, L. T. Batista, R. J. Nascimento, M. C. Cangussú, P. J. Crugeira, L. G. Soares, L. Silveira Jr and A. L. Pinheiro, Detection of prostate cancer by Raman spectroscopy: A multivariate study on patients with normal and altered PSA values, *J. Photochem. Photobiol., B*, 2020, **204**, 111801.
24 H. Nargis, H. Nawaz, H. Bhatti, K. Jilani and M. Saleem, Comparison of surface enhanced Raman spectroscopy and Raman spectroscopy for the detection of breast cancer based on serum samples, *Spectrochim. Acta, Part A*, 2021, **246**, 119034.
25 J. Zhao, H. Zeng, S. Kalia and H. Lui, Using Raman spectroscopy to detect and diagnose skin cancer in vivo, *Dermatol. Clin.*, 2017, **35**(4), 495–504.
26 L. E. Kamemoto, A. K. Misra, S. K. Sharma, M. T. Goodman, H. Luk, A. C. Dykes and T. Acosta, Near-infrared micro-Raman spectroscopy for in vitro detection of cervical cancer, *Appl. Spectrosc.*, 2010, **64**(3), 255–261.
27 F. M. Lyng, D. Traynor, I. R. Ramos, F. Bonnier and H. J. Byrne, Raman spectroscopy for screening and diagnosis of cervical cancer, *Anal. Bioanal. Chem.*, 2015, **407**, 8279–8289.
28 J. Desroches, M. Jermyn, M. Pinto, F. Picot, M.-A. Tremblay, S. Obaid, E. Marple, K. Urmey, D. Trudel, G. Soulez, M.-C. Guiot, B. C. Wilson, K. Petrecca and F. Leblond, A new method using Raman spectroscopy for in vivo targeted brain cancer tissue biopsy, *Sci. Rep.*, 2018, **8**(1), 1792.
29 W. Lee, A. Nanou, L. Rikkert, F. A. Coumans, C. Otto, L. W. Terstappen and H. L. Offerhaus, Label-free prostate cancer detection by characterization of extracellular vesicles using Raman spectroscopy, *Anal. Chem.*, 2018, **90**(19), 11290–11296.
30 L. Zhang, C. Li, D. Peng, X. Yi, S. He, F. Liu, X. Zheng, W. E. Huang, L. Zhao and X. Huang, Raman spectroscopy and machine learning for the classification of breast cancers, *Spectrochim. Acta, Part A*, 2022, **264**, 120300.
31 D. Traynor, I. Behl, D. O'Dea, F. Bonnier, S. Nicholson, F. O'Connell, A. Maguire, S. Flint, S. Galvin, C. M. Healy, C. M. Martin, J. J. O'Leary, A. Malkin, H. J. Byrne and F. M. Lyng, Raman spectral cytopathology for cancer diagnostic applications, *Nat. Protoc.*, 2021, **16**(7), 3716–3735.
32 R. Smith, K. L. Wright and L. Ashton, Raman spectroscopy: an evolving technique for live cell studies, *Analyst*, 2016, **141**(12), 3590–3600.
33 M. Hardy and H. O. M. Chu, Laser wavelength selection in Raman spectroscopy, *Analyst*, 2025, **150**(10), 1986–2008.
34 C. Chen, Z. Zhao, N. Qian, S. Wei, F. Hu and W. Min, Multiplexed live-cell profiling with Raman probes, *Nat. Commun.*, 2021, **12**(1), 3405.
35 A. A. Fung, K. Hoang, H. Zha, D. Chen, W. Zhang and L. Shi, Imaging sub-cellular methionine and insulin interplay in triple negative breast cancer lipid droplet metabolism, *Front. Oncol.*, 2022, **12**, 858017.
36 C. Ma, L. Zhang, T. He, H. Cao, X. Ren, C. Ma, J. Yang, R. Huang and G. Pan, Single cell Raman spectroscopy to identify different stages of proliferating human hepatocytes for cell therapy, *Stem Cell Res. Ther.*, 2021, **12**(1), 555.
37 X. Yuan, Y. Song, Y. Song, J. Xu, Y. Wu, A. Glidle, M. Cusack, U. Z. Ijaz, J. M. Cooper, W. E. Huang and H. Yin, Effect of Laser Irradiation on Cell Function and Its Implications in Raman Spectroscopy, *Appl. Environ. Microbiol.*, 2018, **84**(8), e02508–e02517.

38 S. Wang, H. Wang and Y. Han, Study on effects of neoadjuvant chemotherapy on ovarian cancer tissues using Raman spectroscopy, *Laser Phys.*, 2025, **35**(3), 035601.
39 B. Brozek-Pluska, K. Miazek, J. Musiał and R. Kordek, Label-free diagnostics and cancer surgery Raman spectra guidance for the human colon at different excitation wavelengths, *RSC Adv.*, 2019, **9**(69), 40445–40454.
40 A. H. Agbaria, G. Beck Rosen, I. Lapidot, D. H. Rich, M. Huleihel, S. Mordechai, A. Salman and J. Kapelushnik, Differential Diagnosis of the Etiologies of Bacterial and Viral Infections Using Infrared Microscopy of Peripheral Human Blood Samples and Multivariate Analysis, *Anal. Chem.*, 2018, **90**(13), 7888–7895.
41 C. Kerepesi, B. Daroczy, A. Sturm, T. Vellai and A. Benczur, Prediction and characterization of human ageing-related proteins by using machine learning, *Sci. Rep.*, 2018, **8**(1), 4094.
42 H. Wang, C. Liu and L. Deng, Enhanced Prediction of Hot Spots at Protein-Protein Interfaces Using Extreme Gradient Boosting, *Sci. Rep.*, 2018, **8**(1), 14285.
43 A. H. Agbaria, G. B. Rosen, I. Lapidot, D. H. Rich, S. Mordechai, J. Kapelushnik, M. Huleihel and A. Salman, Rapid diagnosis of infection etiology in febrile pediatric oncology patients using infrared spectroscopy of leukocytes, *J. Biophotonics*, 2020, **13**(2), e201900215.
44 U. Moorthy and U. D. Gandhi, A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization, *J. Ambient Intell. Humaniz. Comput.*, 2021, **12**(3), 3527–3538.
45 Y. Liu, Z. Liu, X. Luo and H. Zhao, Diagnosis of Parkinson's disease based on SHAP value feature selection, *Biocybern. Biomed. Eng.*, 2022, **42**(3), 856–869.
46 H. Liu and H. Motoda, *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*, Chapman & Hall/CRC, 2007.
47 Y. Saeys, I. Inza and P. Larranaga, A review of feature selection techniques in bioinformatics, *Bioinformatics*, 2007, **23**(19), 2507–2517.
48 U. Sharaha, Y. D. Eshel, D. Bykhovsky, J. Mazar, I. Lapidot, M. Huleihel, S. Mordechai, A. Salman and J. Kapelushnik, Augmentation of Infrared Microscopy of White Blood Cells and Medical Measures for Rapid and Accurate Diagnosis of Bacterial or Viral Infections in Febrile Pediatric Oncology Patients: An Expert System-Based Study, *Anal. Chem.*, 2025, DOI: **10.1021/acs.analchem.5c01728**.
49 C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, Springer, 2006, vol. 4.
50 M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood and K. A. Heys, Using Fourier transform IR spectroscopy to analyze biological materials, *Nat. Protoc.*, 2014, **9**(8), 1771–1791.
51 L. Rieppo, S. Saarakkala, T. Närhi, H. J. Helminen, J. S. Jurvelin and J. Rieppo, Application of second derivative spectroscopy for increasing molecular specificity of fourier transform infrared spectroscopic imaging of articular cartilage, *Osteoarthritis Cartilage*, 2012, **20**(5), 451–459.
52 I. Omurlu, M. Ture and A. Kurum, Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease, *Expert Syst. Appl.*, 2008, **34**, 366–374.
53 P. Jeganathan, On the Asymptotic Theory of Estimation When the Limit of the Log-Likelihood Ratios Is Mixed Normal, *Sankhya, Ser. A*, 1982, **44**(2), 173–212.
54 G. Liang, W. Fu and K. Wang, Analysis of *t*-test misuses and SPSS operations in medical research papers, *Burns Trauma*, 2019, **7**, s41038-019-0170-3.
55 R. L. Wasserstein and N. A. Lazar, The ASA Statement on *p*-Values: Context, Process, and Purpose, *Am. Stat.*, 2016, **70**(2), 129–133.
56 C. Yu, E. Gestl, K. Eckert, D. Allara and J. Irudayaraj, Characterization of human breast epithelial cells by confocal Raman microspectroscopy, *Cancer Detect. Prev.*, 2006, **30**(6), 515–522.
57 L. Mikoliunaite, R. D. Rodriguez, E. Sheremet, V. Kolchuzhin, J. Mehner, A. Ramanavicius and D. R. Zahn, The substrate matters in the Raman spectroscopy analysis of cells, *Sci. Rep.*, 2015, **5**(1), 1–10.
58 C. Aksoy and F. Severcan, Role of vibrational spectroscopy in stem cell research, *Spectroscopy*, 2012, **27**(3), 167–184.
59 R. Gautam, S. Vanga, F. Ariese and S. Umapathy, Review of multidimensional data processing approaches for Raman and infrared spectroscopy, *EPJ Tech. Instrum.*, 2015, **2**(1), 1–38.
60 C. Kallaway, L. M. Almond, H. Barr, J. Wood, J. Hutchings, C. Kendall and N. Stone, Advances in the clinical application of Raman spectroscopy for cancer diagnostics, *Photodiagn. Photodyn. Ther.*, 2013, **10**(3), 207–219.
61 K. Gajjar, L. D. Heppenstall, W. Pang, K. M. Ashton, J. Trevisan, I. I. Patel, V. Llabjani, H. F. Stringfellow, P. L. Martin-Hirsch, T. Dawson and F. L. Martin, Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis, *Anal. Methods*, 2012, **5**, 89–102.
62 N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr and J. M. O'Sullivan, A review of feature selection methods for machine learning-based disease risk prediction, *Front. Bioinf.*, 2022, **2**, 927312.
63 D. Cashman, M. Keller, H. Jeon, B. C. Kwon and Q. Wang, A Critical Analysis of the Usage of Dimensionality Reduction in Four Domains, *arXiv*, 2025, preprint, *arXiv*:2503.08836, DOI: **10.1109/TVCG.2025.3567989**.
64 R. O. Duda, P. E. Hart, P. E. Hart, P. E. Hart, D. G. Stork, E. Library and J. Wiley, *Sons, Pattern Classification*, Wiley, 2001.
65 A. Salman, I. Lapidot, A. Pomerantz, L. Tsror, E. Shufan, R. Moreh, S. Mordechai and M. Huleihel, Identification of fungal phytopathogens using Fourier transform infrared-attenuated total reflection spectroscopy and advanced statistical methods, *J. Biomed. Opt.*, 2012, **17**(1), 017002.
66 F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao and J. Zhu, Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges, in *Natural Language Processing*

and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II, Springer-Verlag, Dunhuang, China, 2019, pp. 563–574.

67 Z. Movasaghi, S. Rehman and I. U. Rehman, Raman Spectroscopy of Biological Tissues, *Appl. Spectrosc. Rev.*, 2007, **42**(5), 493–541.

68 Y. Fu, T. Zou, X. Shen, P. J. Nelson, J. Li, C. Wu, J. Yang, Y. Zheng, C. Bruns and Y. Zhao, Lipid metabolism in cancer progression and therapeutic strategies, *MedComm*, 2021, **2**(1), 27–59.

69 M. Ho Zhi Guang, E. L. Kavanagh, L. P. Dunne, P. Dowling, L. Zhang, S. Lindsay, D. Bazou, C. Y. Goh, C. Hanley and G. Bianchi, Targeting proteotoxic stress in cancer: a review of the role that protein quality control pathways play in oncogenesis, *Cancers*, 2019, **11**(1), 66.

70 R. J. DeBerardinis and N. S. Chandel, Fundamentals of cancer metabolism, *Sci. Adv.*, 2016, **2**(5), e1600200.

71 G. Gentric, V. Mieulet and F. Mechta-Grigoriou, Heterogeneity in Cancer Metabolism: New Concepts in an Old Field, *Antioxid. Redox Signaling*, 2016, **26**(9), 462–485.

72 L. P. Fernández, M. Gomez de Cedron and A. Ramirez de Molina, Alterations of lipid metabolism in cancer: implications in prognosis and treatment, *Front. Oncol.*, 2020, **10**, 577420.

73 Q. Zhong, X. Xiao, Y. Qiu, Z. Xu, C. Chen, B. Chong, X. Zhao, S. Hai, S. Li and Z. An, Protein posttranslational modifications in health and diseases: Functions, regulatory mechanisms, and therapeutic implications, *MedComm*, 2023, **4**(3), e261.

74 J. Wang, M. Wang, X. Zeng, Y. Li, L. Lei, C. Chen, X. Lin, P. Fang, Y. Guo and X. Jiang, Targeting membrane contact sites to mediate lipid dynamics: innovative cancer therapies, *Cell Commun. Signaling*, 2025, **23**(1), 1–26.

75 S. Zhang, N. Zhu, H. F. Li, J. Gu, C. J. Zhang, D. F. Liao and L. Qin, The lipid rafts in cancer stem cell: a target to eradicate cancer, *Stem Cell Res. Ther.*, 2022, **13**(1), 432.

76 R. Hosea, S. Hillary, S. Naqvi, S. Wu and V. Kasim, The two sides of chromosomal instability: drivers and brakes in cancer, *Signal Transduction Targeted Ther.*, 2024, **9**(1), 75.

77 F. Khan, D. Elsori, M. Verma, S. Pandey, S. Obaidur Rab, S. Siddiqui, N. M. Alabdallah, M. Saeed and P. Pandey, Unraveling the intricate relationship between lipid metabolism and oncogenic signaling pathways, *Front. Cell Dev. Biol.*, 2024, **12**, 1399065.

78 J. Y. Hahm, J. Park, E.-S. Jang and S. W. Chi, 8-Oxoguanine: from oxidative damage to epigenetic and epitranscriptional modification, *Exp. Mol. Med.*, 2022, **54**(10), 1626–1642.

79 R. Kanwal and S. Gupta, Epigenetic modifications in cancer, *Clin. Genet.*, 2012, **81**(4), 303–311.

80 S. Robbins, R. Cotran and V. Kumar, *Robbins pathological basis of disease*, WB Saunders, Philadelphia, 1994.

81 R. W. Ruddon, *Cancer biology*, Oxford University Press, 2007.

82 M. Li, Q. Sun and X. Wang, Transcriptional landscape of human cancers, *Oncotarget*, 2017, **8**(21), 34534–34551.

83 A. Ertel, A. Verghese, S. W. Byers, M. Ochs and A. Tozeren, Pathway-specific differences between tumor cell lines and normal and tumor tissue cells, *Mol. Cancer*, 2006, **5**(1), 1–13.

84 A. Salman, E. Shufan, L. Zeiri and M. Huleihel, Detection and identification of cancerous murine fibroblasts, transformed by murine sarcoma virus in culture, using Raman spectroscopy and advanced statistical methods, *Biochim. Biophys. Acta, Gen. Subj.*, 2013, **1830**(3), 2720–2727.

85 P. Branco, L. Torgo and R. P. Ribeiro, A Survey of Predictive Modeling on Imbalanced Domains, *ACM Comput. Surv.*, 2016, **49**(2), 31.

86 G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue and G. Bing, Learning from class-imbalanced data: Review of methods and applications, *Expert Syst. Appl.*, 2017, **73**, 220–239.

87 M. Buda, A. Maki and M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, *Neural Networks*, 2018, **106**, 249–259.