RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2024, **14**, 3599

Complete characterization of RNA biomarker fingerprints using a multi-modal ATR-FTIR and SERS approach for label-free early breast cancer diagnosis†

Shuyan Zhang, ‡[a] Steve Qing Yang Wu,[a] Melissa Hum,[b] Jayakumar Perumal,§[a] Ern Yu Tan,[cd] Ann Siew Gek Lee,*[bef] Jinghua Teng, *[a] U. S. Dinish §*[a] and Malini Olivo§*[a]

Breast cancer is a prevalent form of cancer worldwide, and the current standard screening method, mammography, often requires invasive biopsy procedures for further assessment. Recent research has explored microRNAs (miRNAs) in circulating blood as potential biomarkers for early breast cancer diagnosis. In this study, we employed a multi-modal spectroscopy approach, combining attenuated total reflection Fourier transform infrared (ATR-FTIR) and surface-enhanced Raman scattering (SERS) to comprehensively characterize the full-spectrum fingerprints of RNA biomarkers in the blood serum of breast cancer patients. The sensitivity of conventional FTIR and Raman spectroscopy was enhanced by ATR-FTIR and SERS through the utilization of a diamond ATR crystal and silver-coated silicon nanopillars, respectively. Moreover, a wider measurement wavelength range was achieved with the multi-modal approach than with a single spectroscopic method alone. We have shown the results on 91 clinical samples, which comprised 44 malignant and 47 benign cases. Principal component analysis (PCA) was performed on the ATR-FTIR, SERS, and multi-modal data. From the peak analysis, we gained insights into biomolecular absorption and scattering-related features, which aid in the differentiation of malignant and benign samples. Applying 32 machine learning algorithms to the PCA results, we identified key molecular fingerprints and demonstrated that the multi-modal approach outperforms individual techniques, achieving higher average validation accuracy (95.1%), blind test accuracy (91.6%), specificity (94.7%), sensitivity (95.5%), and *F*-score (94.8%). The support vector machine (SVM) model showed the best area under the curve (AUC) characterization value of 0.9979, indicating excellent performance. These findings highlight the potential of the multi-modal spectroscopy approach as an accurate, reliable, and rapid method for distinguishing between malignant and benign breast tumors in women. Such a label-free approach holds promise for improving early breast cancer diagnosis and patient outcomes.

Received 22nd August 2023
Accepted 17th November 2023

DOI: 10.1039/d3ra05723b

rsc.li/rsc-advances

1. Introduction

Breast cancer is a significant global health concern and remains the most commonly diagnosed cancer in women worldwide. In 2020 alone, approximately 2.3 million new cases were reported, with a total of 7.8 million women living with breast cancer diagnosed over the past five years.[1] Timely detection and treatment are crucial for improving survival rates. Although mammography, an X-ray imaging technique, serves as the current gold standard for breast cancer screening, it has

[a]*Institute of Materials Research and Engineering (IMRE), Agency for Science, Technology and Research (A*STAR), 2 Fusionopolis Way, Innovis #08-03, Singapore, 138634, Republic of Singapore. E-mail: jh-teng@imre.a-star.edu.sg; dinish@asrl.a-star.edu.sg; malini_olivo@asrl.a-star.edu.sg*

[b]*Division of Cellular and Molecular Research, Humphrey Oei Institute of Cancer Research, National Cancer Centre Singapore (NCCS), 30 Hospital Boulevard, Singapore, 168583, Republic of Singapore. E-mail: dmslsg@nccs.com.sg*

[c]*Breast & Endocrine Surgery, Tan Tock Seng Hospital (TTSH), 11 Jln Tan Tock Seng, Singapore, 308433, Republic of Singapore*

[d]*Lee Kong Chian School of Medicine, Nanyang Technological University, 50 Nanyang Avenue, Singapore, 639798, Republic of Singapore*

[e]*SingHealth Duke-NUS Oncology Academic Clinical Programme (ONCO ACP), Duke-NUS Medical School, Singapore 169857, Republic of Singapore*

[f]*Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117593, Republic of Singapore*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3ra05723b

‡ Current address: Timbre Technologies, Tokyo Electron America, 2859 Bayview Drive, Fremont, CA 94538, USA.

§ Current address: A*STAR Skin Research Labs (A*SRL), Agency for Science, Technology and Research (A*STAR), 31 Biopolis Way, #07-01 Nanos, Singapore, 138669, Republic of Singapore.

limitations, with approximately 20% of breast cancer cases going undetected.[2,3] Furthermore, the current mainstream diagnostic tools, including mammography, magnetic resonance imaging (MRI), and ultrasonography, often yield false positives, leading to undue stress for patients and additional diagnostic procedures.[3–6] Therefore, there is an unmet clinical need for a rapid, accurate, and reliable test for breast cancer screening. One potential solution lies in the detection of biomarkers, specifically circulating microRNAs (miRNAs), which have shown promise in differentiating between individuals with and without cancer, particularly for those with abnormal mammograms.[7–9]

MiRNAs are small, non-coding RNA molecules, approximately 22 nucleotides in length. They have emerged as promising biomarkers for cancer detection due to their stability and abundance in body fluids such as serum and plasma.[10,11] Unlike other RNA molecules, miRNAs possess specific structures that render them resistant to degradation by nucleases. This unique characteristic makes them attractive candidates for early cancer detection, as miRNA expression patterns have been found to be deregulated in cancer patients. Moreover, miRNAs exhibit wide distribution in various organs, indicating their potential utility in personalized medicine. Although existing detection techniques such as quantitative reverse transcriptase polymerase chain reaction (RT-qPCR) and next-generation sequencing (NGS) demonstrate high sensitivity and specificity,[12–14] their utilization can be expensive and time-consuming due to the need for chemical labeling. Hence, there is a demand for faster and more affordable methods for miRNA detection.

Fourier transform infrared (FTIR) spectroscopy is a powerful tool for analyzing the chemical composition and molecular structure of biological samples. This technique measures the absorption of light by the sample, providing a molecular fingerprint that can detect changes associated with disease progression. The attenuated total reflection FTIR (ATR-FTIR) spectroscopy utilizes a high refractive index crystal. When infrared light is incident on the crystal, it creates an evanescent wave due to differences in refractive indices between the crystal and the sample. This means that only the molecules in close proximity to the crystal surface interact with the evanescent wave, leading to a stronger signal for a thin layer of the sample compared to traditional FTIR. ATR-FTIR spectroscopy has the potential for rapid and accurate detection of miRNAs for early cancer diagnosis and personalized medicine.[15–19] Raman spectroscopy provides molecular information and can have sensitivity enhanced by surface-enhanced Raman scattering (SERS) to detect low-concentration samples.[20,21] SERS utilizes nanoroughened surfaces coated with metal (like copper, silver, or gold), called planar SERS substrates, or metal colloidal nanoparticles to enhance the Raman signal, enabling the detection of miRNA fingerprints at very low concentrations. Both ATR-FTIR and SERS techniques are label-free techniques that have been used for the detection of biomolecules in various biological samples.[22–25]

Previous studies have demonstrated the potential of FTIR and SERS techniques for sensitive and accurate detection and analysis of nucleic acids.[26–28] For instance, D. Li *et al.* and Y. Li *et al.* utilized SERS to detect miRNA and RNA bases, respectively, achieving improved sensitivity.[29,30] Rios *et al.* employed FTIR spectroscopy to detect DNA polymorphisms with high accuracy using machine learning algorithms.[31] Geinguenaud *et al.* utilized FTIR spectroscopy to study RNA structures, identifying key vibrational modes associated with RNA sugar puckering, backbone vibrations, phosphate stretching, and protein secondary structures.[32] These studies underscore the potential of using spectroscopy techniques for sensitive and accurate detection and analysis of nucleic acids.

Concurrently, the integration of machine learning and chemometrics with spectroscopy has gained interest not just for medical diagnostics,[33–35] but also for applications such as food quality control, detection of chloramphenicol in food products,[36] and the comparative study of chemometric challenges in food analysis.[37] The energy sector is similarly evolving with these methodologies. Progress in dye-sensitized solar cells is attributed to insights into interfacial effects in solid–liquid electrolytes,[38] the effect of polymer electrolytes at the nanoscale,[39] and the tuning of properties in carbazole photosensitizers.[40] Supercapacitors, another essential energy storage technology, have also benefited from machine learning, as seen in the work on laser-induced graphene-based capacitors.[41]

In this paper, we present a novel multi-modal spectroscopy approach for early breast cancer diagnosis using combined ATR-FTIR and SERS data. Our study involved the measurement of 91 clinical samples with malignant and benign diagnoses previously confirmed through histopathology analysis. We explored a total of 32 machine learning models, each with varying training, validation, and blind test ratios, for the classification task using ATR-FTIR alone, SERS alone, and the combined multi-modal data. The results showed that the multi-modal approach achieved the best performance, with a validation accuracy of 95.1% and a test accuracy of 91.6%. Among the machine learning models, the support vector machine (SVM) outperformed others, demonstrating an impressive area under the curve (AUC) value of 0.9979. This outcome demonstrates that multi-modal spectroscopy provides complementary information and improves the accuracy of miRNA detection. Our label-free and rapid testing method, assisted by machine learning, offers a comprehensive characterization of the molecular fingerprints of biomarker molecules and high accuracy in early breast cancer diagnosis.

2. Materials and methodology

2.1. Samples

The sample collection and processing procedures are similar to our previous study.[19] Serum samples for the analysis of microRNAs (miRNAs) were obtained from peripheral blood samples collected at the National Cancer Centre Singapore (Singapore) and Tan Tock Seng (Singapore) prior to biopsy and surgery. Additional serum samples were obtained from the SingHealth Tissue Repository (Singapore). These samples were not purchased or donated. The study followed the principles of the Declaration of Helsinki with approval from the Centralized Institutional Review Board of SingHealth (CIRB Ref: 2018/2874). Written informed consent was obtained from all participants.

A total of 91 samples were included in this study, with 44 diagnosed as malignant and 47 as benign based on histopathology analysis. To minimize the impact of confounding factors and technical biases in data analysis, pre-analytical factors, including sample collection, handling, processing, and storage, were standardized.[10] Blood samples were collected and promptly processed within 50–60 minutes of venipuncture to separate serum from whole blood. The serum samples were aliquoted and stored at −80 °C to prevent freeze–thaw cycles, with only non-hemolyzed samples used in this study. Subsequently, total RNA was isolated from 200 μL of serum using the miRNeasy Serum/Plasma Advanced Kit (Qiagen, N.V.), following the manufacturer protocol. An additional step involving the addition of bacteriophage MS2 RNA to the sample lysis buffer (1 μg mL^{-1} of QIAzol) was included to enhance the RNA yield. Total RNA extraction was performed using the same reagents and procedures for all 91 samples.

2.2. Experimental setup

The study employed ATR-FTIR and SERS techniques to analyze miRNA samples for early breast cancer diagnosis. ATR-FTIR spectroscopy, as illustrated in Fig. 1(a), utilizes an incident beam from a globar source that enters an ATR crystal with a high refractive index. Through total internal reflection, the beam is reflected at the crystal–sample interface, creating an evanescent wave that penetrates the sample. During this interaction, specific frequencies of light in the infrared range are absorbed by the sample, resulting in characteristic absorption bands. The reflected beam carries the spectral information of the absorbed frequencies and is directed toward the FTIR detector. ATR-FTIR spectroscopy provides valuable insights into the molecular composition and interactions within the sample, making it a powerful analytical technique for various applications. In this study, an ATR-FTIR system (Vertex 80v with ATR diamond crystal accessory, Bruker) was used to obtain spectra from 10 μL of miRNA samples under a vacuum condition. Each clinical sample was subjected to 20 measurements sequentially without changing the sample. For each measurement, an average was taken based on 64 scans at a resolution of 4 cm^{-1}. All these measurement results were then used for subsequent analysis. The vacuum condition ensured that the collected data was free from interference by water vapour,[19] as shown in the ESI, Fig. S1.†

As depicted in Fig. 1(b), SERS involves the illumination of the sample with laser light and the detection of the enhanced inelastically scattered photons through the plasmonic effect. Enhancement of Raman signal is achieved by depositing the sample on nano-roughened metal-coated surfaces called SERS substrates. Here, SERS substrates were fabricated on silicon wafers, and nanostructures were in the form of nanopillars, which were formed using the inductively coupled plasma-based blanket etching method. The size of nanopillars was typically ~200 nm in height, and it was coated with a 150 nm layer of silver.[42] When the laser light interacts with the sample, due to the localized electric field enhancement generated by the silver-coated nanopillars, resulting in amplifying the Raman signal of the molecules in the proximity. This enhanced Raman scattering provides detailed molecular information, enabling sensitive and selective detection of the sample. SERS offers immense potential for various applications, including chemical analysis and biosensing.[43] SERS measurements were conducted using a Raman microscope system (Invia, Renishaw) integrated with a Leica microscope. The laser light (785 nm) was coupled through a long working distance objective lens (50×, 0.5 NA) to excite the sample and collect the scattered Raman signal. The clinical miRNA samples (10 μL) were pipetted onto the bare SERS substrates, and enhanced Raman signals were collected in backscattering geometry. Multiple measurements were taken at 20 different locations (~20 μm apart) on the substrate, and averaged spectra were used for analysis. The spectral measurements were performed with a laser power of ~450 μW.

2.3. Data processing workflow

The workflow of the sample preparation, data collection, and data analysis is illustrated in Fig. 2. The raw data underwent pre-processing steps before machine learning analysis, which

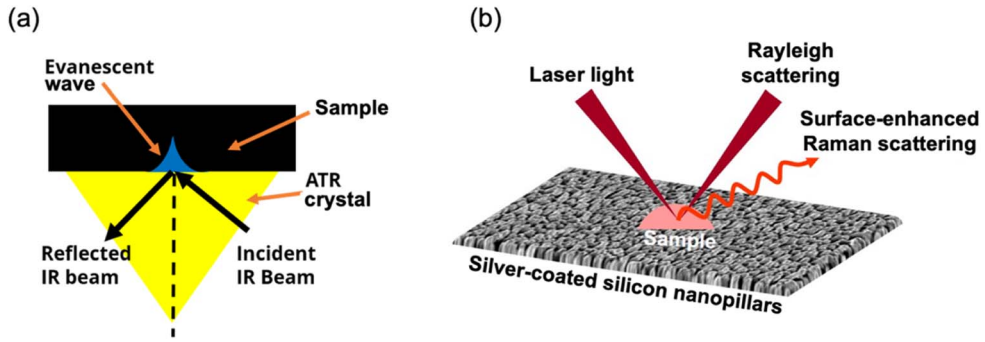


Fig. 1 (a) ATR-FTIR spectroscopy uses a beam from a globar source entering an ATR crystal. Through internal reflection, an evanescent wave interacts with the sample, absorbing specific infrared frequencies. The reflected beam, carrying this information, is then directed to the FTIR detector. (b) SERS uses laser light to detect enhanced scattered photons *via* plasmonic effects. The sample is deposited on nano-roughened, metal-coated surfaces (SERS substrates). Interaction with laser light amplifies the Raman signal due to the electric field from silver-coated nanopillars, offering detailed molecular insights for precise sample detection.

included baseline correction, Savitzky–Golay smoothing, removal of noisy and atmospheric peaks, and normalization. The processed ATR-FTIR and SERS data were combined based on the wavelength. ATR-FTIR wavenumber was converted to the wavelength using eqn (1), and the SERS Raman shift was converted to the wavelength using eqn (2), where $\lambda_{ex} = 785$ nm. After the conversion, the ATR-FTIR data ranged from 2 to 20 μm, and SERS data ranged from 0.8 to 0.9 μm. Consequently, the combined multi-modal data spanned from 0.8 to 20 μm with a gap of 1.1 μm from 0.9 to 2 μm.

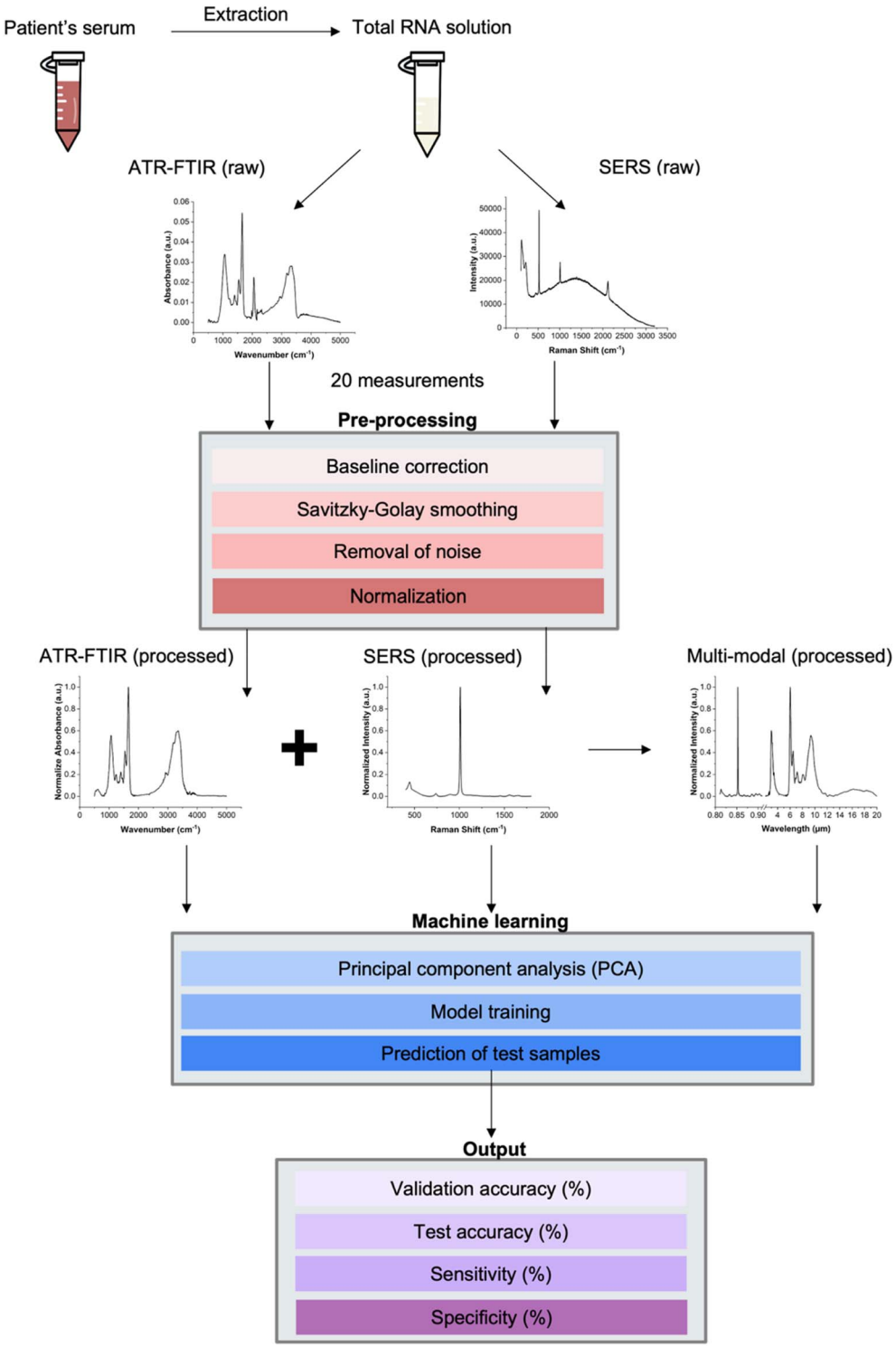


Fig. 2 Workflow illustrating the process of sample preparation; data collection using ATR-FTIR and SERS techniques; data processing for ATR-FTIR alone, SERS alone, and multi-modal; machine learning, and final output including validation accuracy, test accuracy, sensitivity, and specificity.

$$\lambda(\mu m) = \frac{10^4}{\text{wavenumber}(cm^{-1})} \quad (1)$$

$$\lambda(\mu m) = \frac{10^{-3}}{\frac{1}{\lambda_{ex}(nm)} - \frac{\text{Raman shift}(cm^{-1})}{10^7}} \quad (2)$$

Machine learning algorithms were applied to the processed ATR-FTIR data, SERS data, and multi-modal data separately. The steps included principal component analysis (PCA), model training, and prediction of test results. The outcomes were evaluated using five parameters: validation accuracy, test accuracy, sensitivity, specificity, and *F*-score.

2.4. Machine learning methods

In this study, a total of 32 different machine learning models were developed and trained using MATLAB (R2022a, MathWorks). During the model training process, PCA and cross-validation methods were implemented to enhance the accuracy and robustness of the models.

2.4.1. Data preparation. The dataset consisting of spectroscopic measurements of 91 samples was divided into training and test datasets. To assess the model performance, 6 different sets of blind test samples (*i.e.*, not overlapping with the training and validation datasets) were selected, including 5, 10, 15, 20, 25, and 30 test samples. The remaining samples were utilized for training and validation purposes to construct the machine learning models. Table 1 provides a breakdown of the sample splitting, indicating the ratio of test samples to training + validation samples. To eliminate potential biases in the test dataset, each ratio was run three times, with each run employing a randomly selected test sample set.

Ten-fold cross-validation and PCA were employed for training the models. Ten-fold cross-validation involved partitioning the dataset into ten sets of data, with one set used for validation and the other sets utilized for training. This methodology ensured that the models were trained on different datasets, promoting greater generalization and robustness. As the spectral data used in this study had high dimensionality, PCA was employed to reduce the computational requirements. The Origin software (2022a, OriginLab) was utilized to perform PCA by generating a scree plot and identifying the elbow point to determine the optimal number of principal components (PCs). PCA was conducted for each data method (ATR-FTIR, SERS, and Multi-modal).

2.4.2. Types of models. The machine learning algorithms used in this study encompassed decision trees, discriminant analysis, logistic regression, naïve Bayes, SVM, *k*-nearest neighbors (KNN), ensemble models, neural networks, and kernel approximations. A list of the models is provided in the ESI, Table S1.† The selection of these models allowed for a comparison of their performance on different datasets. Decision trees utilize conditions to make decisions and branch into different branches based on predictor values and trained weights. Discriminant analysis classifies data based on Gaussian distributions, while logistic regression employs a sigmoid curve as a decision boundary. Naïve Bayes classifiers utilize the Bayes theorem to calculate the probability of a sample belonging to a particular class. SVMs utilize separating hyperplanes to distinguish data points, and KNN models classify samples based on the classes of their nearest neighbors. Ensemble models combine weaker techniques such as bagging and boosting to create a more robust ensemble model. Neural networks consist of layers of neurons with weights that are trained during model training, while kernel approximations transform lower-dimensional data into higher-dimensional

Table 1 Breakdown of sample splitting for machine learning datasets with different test *vs.* training + validation ratios. The total number of samples is 91, with 44 being malignant and 47 being benign samples

		Training + validation samples			Test samples			Total samples		
Ratio	Run	Malignant	Benign	Total	Malignant	Benign	Total	Malignant	Benign	Total
0.058	1	41	45	86	3	2	5	44	47	91
	2	43	43		1	4				
	3	42	44		2	3				
0.123	1	39	42	81	5	5	10			
	2	41	40		3	7				
	3	40	41		4	6				
0.197	1	36	40	76	8	7	15			
	2	37	39		7	8				
	3	38	38		6	9				
0.282	1	33	38	71	11	9	20			
	2	34	37		10	10				
	3	34	37		10	10				
0.379	1	31	35	66	13	12	25			
	2	31	35		13	12				
	3	32	34		12	13				
0.492	1	28	33	61	16	14	30			
	2	29	32		15	15				
	3	29	32		15	15				

data using kernel functions, enabling linear classifiers such as planes to separate data points belonging to different classes.

2.4.3. Model selection. The performance of the models was evaluated using various metrics, including validation and test accuracy, the discrepancy between the validation and test accuracy, specificity, sensitivity, and *F*-score. Models with a discrepancy exceeding 15% were excluded to prevent overfitting. These metrics were calculated using the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values obtained from the confusion matrix, as shown in eqn (3)–(5).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$F\text{-score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{5}$$

3. Results and discussion

3.1. Molecular fingerprints

Fig. 3 presents the average of the measurement results obtained using both ATR-FTIR and SERS techniques. Fig. 3(a) displays the normalized absorption spectra of a malignant sample (red) and a benign sample (blue) as measured by ATR-FTIR spectroscopy. Two distinct fingerprint regions are observed: one ranging from 500 to 2000 cm^{-1} and the other from 2500 to 3500 cm^{-1}. This characteristic is consistent across all samples. To validate our measurement accuracy and reproducibility, we also measured synthetic miRNA samples, observing similar features as shown in the ESI, Fig. S2.† Additionally, it can be noted that the peak wavenumbers are nearly identical for both sample types, but the relative peak intensities differ. For instance, the differences in peak intensities at 1066 cm^{-1}, 1541 cm^{-1}, and 3340 cm^{-1} are smaller for the malignant samples compared to the benign samples. Moreover, the width of the broad peak from 2500 to 3500 cm^{-1} is larger for the malignant samples than for the benign samples. Fig. 3(b) illustrates the peak wavenumbers and their corresponding chemical bonds and vibrational groups for DNA and RNA molecules, as documented in the literature.[32,44,45] The most prominent peak wavenumber in both malignant and benign spectra is observed at 1657 cm^{-1}, corresponding to C2═O2 stretching in cytosine or guanine. The second notable peaks are located at 3188 cm^{-1} and 3340 cm^{-1}, corresponding to O–H stretching and N–H stretching, respectively. It is worth mentioning that the peak intensity at 1066 cm^{-1} is more pronounced in malignant samples than in benign samples, corresponding to PO_2^- symmetric stretching.

On the other hand, the SERS spectra in Fig. 3(c) reveal limited molecular fingerprints. The most prominent peak is observed at 1010 cm^{-1}, accompanied by a small peak at 446 cm^{-1}. The functional groups associated with these peaks are depicted in Fig. 3(d), with 1010 cm^{-1} representing CC aromatic ring chain vibrations and 446 cm^{-1} indicating CC aliphatic chains.

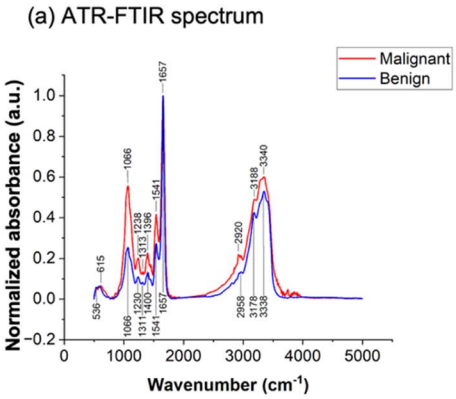
Fig. 3(e) showcases the multi-modal spectra. The smaller wavelength region represents the SERS spectra, while the larger wavelength region represents the ATR-FTIR spectra. Notably, after the wavelength conversion, the ATR-FTIR spectra were horizontally flipped. It is evident that the number of SERS peaks is considerably lower than that of the ATR-FTIR peaks.

3.2. Visual peak analysis

PCA is a powerful approach for reducing and interpreting large multivariate datasets with linear structures, enabling the discovery of previously unsuspected relationships. In this study, PCA was applied to the ATR-FTIR, SERS, and multi-modal data, as depicted in Fig. 4. By utilizing PCA, we were able to investigate the relationship between the light absorption and scattering intensities of biomolecules and their respective wavelengths, while also determining the optimal number of PCs to retain. A scree plot, serving as a visual aid, was employed to identify the appropriate number of PCs. The number is determined by locating the "elbow" point where the remaining eigenvalues become relatively small and of comparable size.

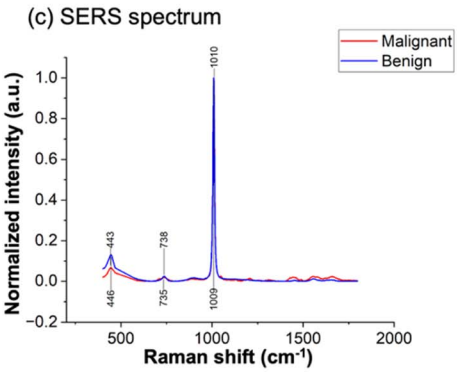
In Fig. 4(a), the scree plot for the ATR-FTIR data is presented. Although the elbow point is not distinctly apparent, we consider the third point as the elbow point. Fig. 4(b) illustrates the loading with reference wavenumber plot for the ATR-FTIR data, showcasing the loading patterns of PC1, PC2, and PC3. These PCs collectively account for 87.2% of the total variance, with PC1 contributing 64.1%, PC2 contributing 11.8%, and PC3 contributing 11.3%. The vertical lines on the plot indicate the important wavenumbers for each PC. Notably, PC1 is associated with significant wavenumbers at 1061 cm^{-1} and 3423 cm^{-1}, with respective loading values of −0.09 and 0.02. For PC2, the influential wavenumbers include 1011 cm^{-1}, 3192 cm^{-1}, and 3367 cm^{-1}, with corresponding loading values of 0.05, −0.05, and −0.07. In PC3, the crucial wavenumbers are 1061 cm^{-1} and 3367 cm^{-1}, with respective loading values of −0.07 and −0.04. These findings align with the spectra presented in Fig. 3(a), where 1061 cm^{-1} corresponds to PO_2^- symmetric stretching and 3367 cm^{-1} corresponds to N–H stretching.[46,47] Notably, the important wavenumbers for each PC correspond to specific chemical bonds or functional groups that are significant in differentiating between malignant and benign samples. These chemical bonds or functional groups play a vital role in DNA and RNA structures, and their variation can provide insights into the differences between malignant and benign DNA/RNA solutions.[48]

For the SERS data, Fig. 4(c) showcases the scree plot, indicating the sixth point as the elbow point. Fig. 4(d) presents the loading with reference Raman shift plot for the SERS data, illustrating the loadings of the first three PCs. PC1 accounts for 37.0% of the total variance, PC2 for 21.8%, and PC3 for 7.8%. The vertical lines on the plot correspond to important Raman shifts for each PC. PC1 is characterized by significant Raman shifts at 441 cm^{-1}, 738 cm^{-1}, and 1003 cm^{-1}, with respective loading values of 0.13, 0.07, and 0.08. In PC2, the influential



(b) ATR-FTIR chemical bonds

Wavenumber (cm^{-1})	Functional group/vibration
788	Free cytosine vibration
798	Free adenine vibration
813, 866	Main n-type sugar marker
915	Ribose ring vibration
966	C-C and C-O ribose stretching
995	Ribose-phosphate main chain vibration
1082	PO_2^- symmetric stretching
1110	C-O ribose ring stretching
1240	PO_2^- anti-symmetric stretching
1475	Adenine and guanine ring vibration
1492	Cytosine in-plane vibration
1600	C=N ring vibration of guanine
1650	C2=O2 stretching in cytosine
1685	C2=O2 stretching in guanine
2950	C-H stretching
3210	O-H stretching
3340	N-H stretching



(d) SERS chemical bonds

Raman shift (cm^{-1})	Functional group/vibration
400	CC aliphatic chains
600 - 1300	CC alicyclic, aliphatic chain vibrations
1000	CC aromatic ring chain vibrations
1600 - 1800	C=O stretching

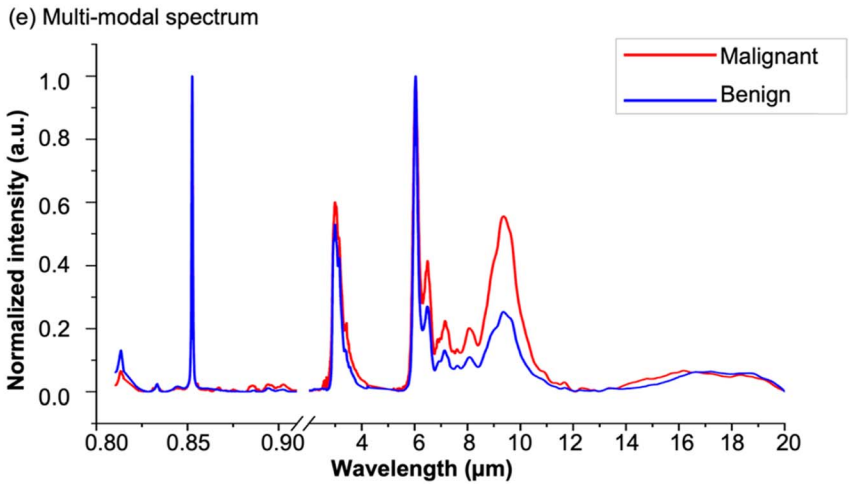


Fig. 3 Molecular fingerprint measurements of malignant (red curves) and benign (blue curves) samples. ATR-FTIR: (a) normalized average spectrum with labeled peak wavenumbers and (b) corresponding chemical bonds. Two distinct fingerprint regions are observed: one ranging from 500 to 2000 cm^{-1} and the other from 2500 to 3500 cm^{-1}. SERS: (c) normalized average spectrum with labeled peak wavenumbers and (d) corresponding chemical bonds. Multi-modal: (e) average spectrum of ATR-FTIR and SERS where the ATR-FTIR wavenumber units and SERS Raman shift units were converted to wavelength units based on eqn (1) and (2).

features include 441 cm^{-1} and 738 cm^{-1}, with loading values of −0.12 and −0.04, respectively. PC3 is characterized by the prominent Raman shifts at 1003 cm^{-1} and 1012 cm^{-1}, with respective loading values of 0.29 and −0.27. Notably, the Raman shifts at 441 cm^{-1} and 738 cm^{-1} are important in both PC1 and PC2, while 1003 cm^{-1} exhibits more influence in PC1 and PC3. These findings are consistent with the spectra depicted in Fig. 3(c), where 441 cm^{-1} corresponds to CC aliphatic chains, 738 cm^{-1} is likely due to CC alicyclic and aliphatic chain vibrations, and 1003 cm^{-1} may be associated with aromatic ring chain vibrations. These molecular features are relevant to DNA and RNA structures and exhibit variations that contribute to the distinction between malignant and benign samples.[49]

Fig. 4(e) displays the scree plot for the multi-modal data, with the third point identified as the elbow point. Fig. 4(f) illustrates the loading with reference wavelength for the multi-

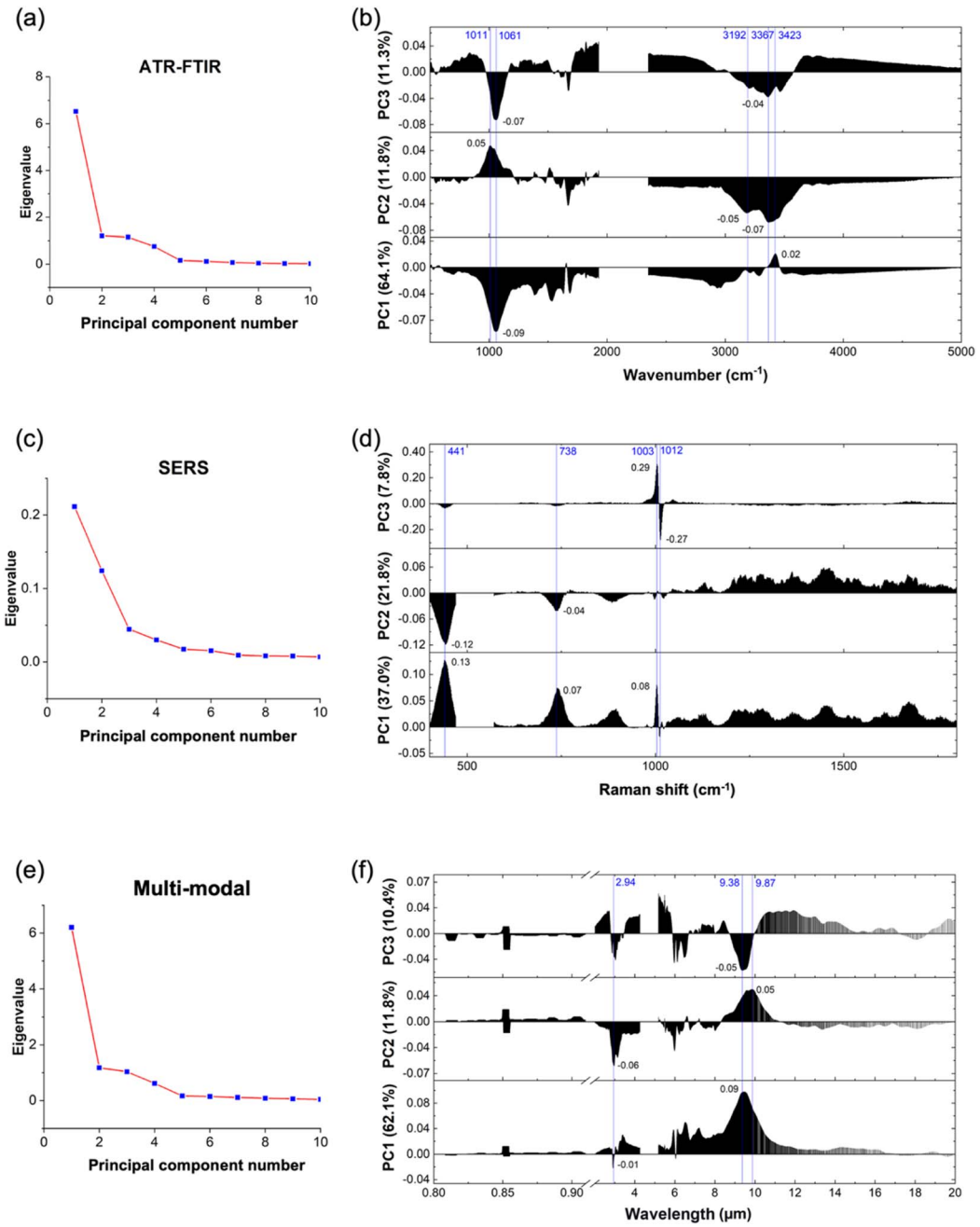


Fig. 4 Results of principal component analysis. ATR-FTIR data: (a) scree plot indicating the third point is the elbow point and (b) loading with reference wavenumber plot showing PC1, PC2, and PC3 characteristic wavenumbers (marked in blue) and their corresponding loading values (marked in black). SERS data: (c) scree plot indicating the sixth point is the elbow point and (d) loading with reference wavenumber plot showing PC1, PC2, and PC3 characteristic wavenumbers (marked in blue) and their corresponding loading values (marked in black). Multi-modal data: (e) scree plot indicating the third point is the elbow point and (f) loading with reference wavenumber plot showing PC1, PC2, and PC3 characteristic wavenumbers (marked in blue) and their corresponding loading values (marked in black). It is shown that the ATR-FTIR data dominate in the characteristics than the SERS data in the multi-modal approach.

modal data, highlighting the contributions of the first three PCs. PC1 accounts for 62.1% of the total variance, PC2 for 11.8%, and PC3 for 10.4%. The vertical lines on the plot denote the important wavelengths for each PC. Notably, 2.94 μm is a significant wavelength in both PC1 and PC2, while 9.38 μm exhibits more influence in PC1 and PC3. These findings align with the spectra depicted in Fig. 3(e). Importantly, it is worth noting that all the significant features originate from the ATR-FTIR data region. This observation suggests that the ATR-FTIR technique is notably more efficient than the SERS technique in classifying malignant and benign breast cancer miRNA biomarkers. More advanced SERS techniques may be explored to improve its detection efficiency, such as introducing an interfacial agent or aggregating agent.[29,30]

In summary, the application of PCA to the ATR-FTIR, SERS, and multi-modal data provides valuable insights into the relationships between biomolecular absorption or scattering intensities and their corresponding wavenumbers or wavelengths. By identifying important wavenumbers and wavelengths associated with specific chemical bonds or functional groups, PCA enables the differentiation between malignant and benign miRNA solutions, contributing to the classification of breast cancer biomarkers.

3.3. Machine learning results

In this section, we will discuss the results and analysis of the machine learning models developed for breast cancer diagnosis using spectral data from three different methods – ATR-FTIR,

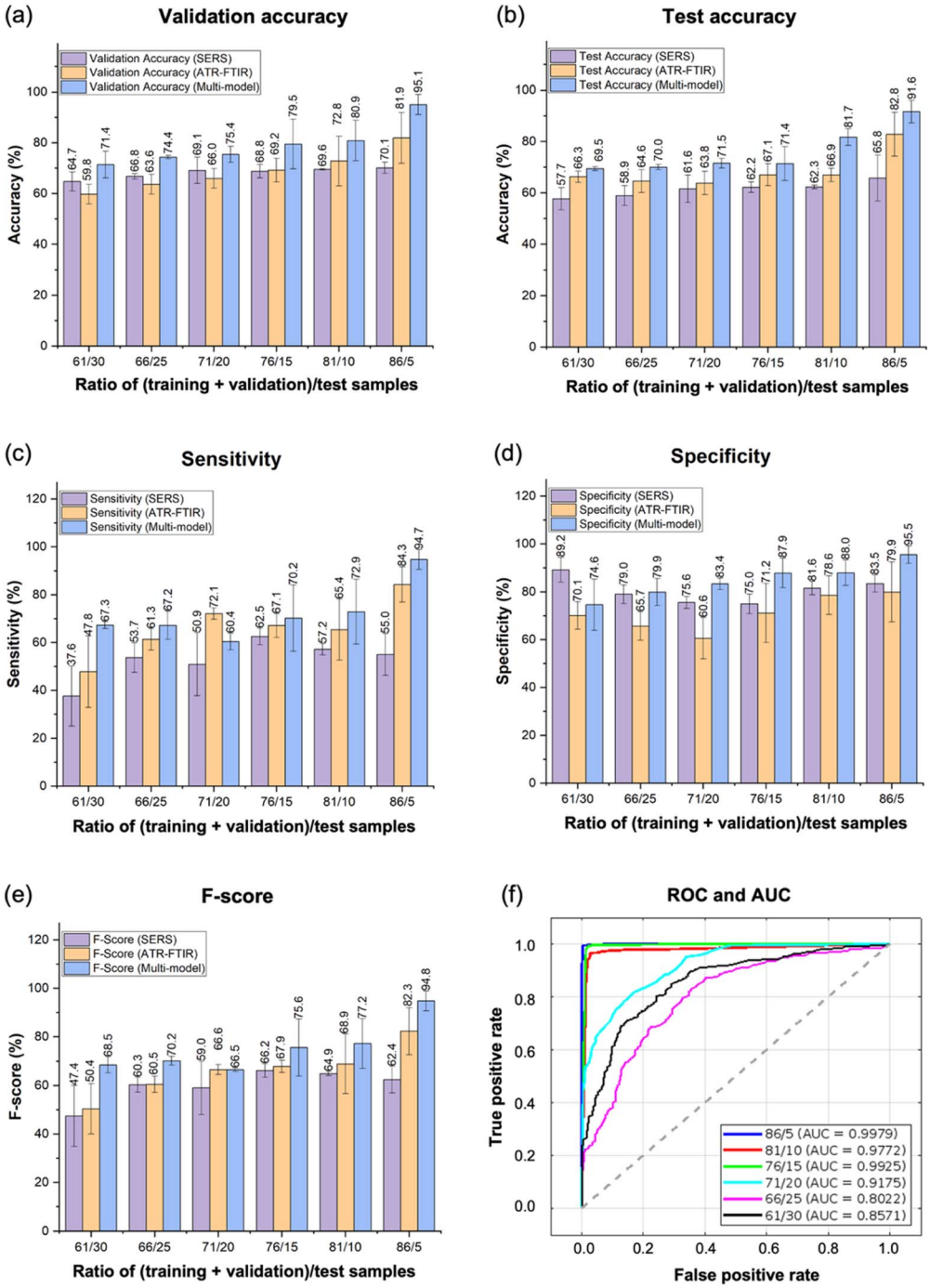


Fig. 5 Machine learning results for various ratios of (training + validation)/test samples. Plots of (a) validation accuracy, (b) test accuracy, (c) sensitivity, (d) specificity and (e) *F*-score where SERS data is shown in purple, ATR-FTIR data is shown in orange, and multi-modal data is shown in blue. The numbers on the bar plots indicate the average values of the three runs and the error bars indicate the standard deviation. (f) Plots of the ROC curves (solid lines) and AUC values (legend values) for the multi-modal data for each ratio.

SERS, and multi-modal spectroscopy. The selection criterion for choosing the best model for each dataset was based on high validation and test accuracies, small validation-test accuracy discrepancy, high sensitivity, specificity, and *F*-score. The average value and standard error were calculated across different runs for each ratio of (training + validation)/test samples of each measurement method, and these values were used for plotting, as shown in Fig. 5.

Fig. 5(a) and (b) depict the validation and test accuracy results, respectively. The average values of the three runs for each split ratio are represented on the bar plots, with standard deviations shown as error bars. The multi-modal data approach exhibits the highest validation accuracy, reaching an impressive 95.1%. With a validation accuracy of 95.1%, we can anticipate approximately 95 correct predictions out of every 100 samples tested. This accuracy level is comparable to the histopathology diagnosis with artificial intelligence assistance and surpasses the histopathology diagnosis alone, indicating its potential as a superior diagnostic tool.[50–52] Notably, as the ratio increases, the accuracy also demonstrates improvement. However, even with a low ratio, a consistently high test accuracy of 69.5% is maintained. It is important to highlight that the SERS accuracy exhibits greater variations from the expected increasing trend, which can be attributed to the relatively fewer features present in the SERS data compared to ATR-FTIR and the multi-modal data. Moreover, the SERS accuracy generally tends to be lower than the ATR-FTIR accuracy, while the multi-modal accuracy surpasses both individual accuracies. This disparity can be explained by the additional information provided by the multi-modal spectroscopy data, which enhances the accuracy of the diagnostic predictions.

Fig. 5(c)–(e) present the results of the sensitivity, specificity, and *F*-score analyses. The multi-modal approach outperforms the ATR-FTIR and SERS data methods individually, achieving the highest sensitivity, specificity, and *F*-score, all at an impressive value of around 95%. This signifies the model ability to accurately classify 95 out of 100 true positive and true negative samples. Moreover, an increase in the ratio leads to improved sensitivity, specificity, and *F*-score. Notably, even at a low ratio, a consistently high sensitivity, specificity, and *F*-score of approximately 70% are maintained. It is important to note that the SERS data exhibits a less discernible trend in sensitivity, specificity, and *F*-score values. This behavior can be attributed to the relatively fewer features available in the SERS spectra, potentially limiting the model ability to capture the differential features required for distinguishing between malignant and benign classes. In addition, we have identified that the best models are the SVM, KNN, and SVM for ATR-FTIR, SERS, and multi-modal data methods, respectively.

Fig. 5(f) displays the receiver operating characteristic (ROC) curves and corresponding AUC values for the multi-modal data at each ratio. The color code is indicated in the legend. A perfect classifier would exhibit a true positive rate (sensitivity) of 1.0 and a false positive rate (1-specificity) of 0.0, while a random classifier is represented by the dashed line. The AUC value ranges from 0.0 to 1.0, with 1.0 indicating a perfect model. Our best AUC value of 0.9979 is achieved at the (training + validation)/test ratio of 86/5 using SVM, and the value generally decreases as the number of test samples increases, with the exception of the 76/15 ratio. Notably, even at the 61/30 ratio, our results demonstrate a relatively high AUC of 0.8571. These findings suggest promising discrimination capabilities in distinguishing between malignant and benign samples.

4. Conclusions

In conclusion, this study highlights the potential of utilizing the multi-modal spectroscopy approach for the detection of miRNA biomarkers in early breast cancer diagnosis. By combining the highly sensitive ATR-FTIR and SERS techniques, complete fingerprint profiles of the biomarkers were obtained. Notably, the ATR-FTIR technique provided a broader range of fingerprint profiles across a wider wavelength range compared to SERS. Machine learning analysis demonstrated the highest accuracy (95.1%) in classifying malignant and benign cases when utilizing the multi-modal approach. These findings indicate the effectiveness of the proposed approach for accurate and reliable label-free breast cancer diagnosis. Furthermore, the approach can be generalized to other biomarker types, including proteins and lipids, thereby expanding its potential applications in various areas of biomedical research. Overall, this study contributes to the development of a robust and versatile spectroscopy-based approach for early cancer detection and holds promise for future advancements in the field.

Author contributions

Conception: S. Zhang, A. S. G. Lee, J. Teng, D. U. S., and M. Olivo. Clinical samples: M. Hum, E. Y. Tan, and A. S. G. Lee. Experiment: S. Zhang, Q. Y. S. Wu, and J. Perumal. Data analysis: S. Zhang. Manuscript preparation: all authors. Supervision: A. S. G. Lee, J. Teng, D. U. S., and M. Olivo.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors would like to acknowledge the funding support from: Agency for Science, Technology and Research, Singapore: IAF-PP Grant H19H6a0025 and BMRC UIBR Grant; National Medical Research Council, Singapore: NMRC/CBRG/0087/2015. The authors would also like to thank Ghayathri Balasundaram for insightful discussions; Chi Lok Wong, Grace Chong, Casey Ang, and Krithika Rajesh for helping with the data collection; Casey Ang, Sonia Tan, and Yingxin Liu for helping with data analysis and figures preparation.

References

1 World Health Organization, *Breast Cancer*, 2023.
2 American Cancer Society, *Limitations of Mammograms*, 2022.

3 J. E. Joy, E. E. Penhoet, D. B. Petitti and National Cancer Policy Board (U.S.), Committee on New Approaches to Early Detection and Diagnosis of Breast Cancer, *Saving Women's Lives*, National Academies Press, Washington, D.C., 2005.
4 H. Thornton, *Br. Med. J.*, 2003, **327**, 101–103.
5 C. van den Ende, A. M. Oordt-Speets, H. Vroling and H. M. E. van Agt, *Int. J. Cancer*, 2017, **141**, 1295–1306.
6 H. D. Nelson, M. Pappas, A. Cantor, J. Griffin, M. Daeges and L. Humphrey, *Ann. Intern. Med.*, 2016, **164**, 256.
7 J. Wang, J. Chen and S. Sen, *J. Cell. Physiol.*, 2016, **231**, 25–30.
8 L. Sempere, C. Graveel, H. Calderone, J. Westerhuis and M. Winn, *Breast Cancer: Targets and Therapy*, 2015, pp. 59–79.
9 C. E. Condrat, D. C. Thompson, M. G. Barbu, O. L. Bugnar, A. Boboc, D. Cretoiu, N. Suciu, S. M. Cretoiu and S. C. Voinea, *Cells*, 2020, **9**, 276.
10 S. Y. Loke, P. Munusamy, G. L. Koh, C. H. T. Chan, P. Madhukumar, J. L. Thung, K. T. B. Tan, K. W. Ong, W. S. Yong, Y. Sim, C. L. Oey, S. Z. Lim, M. Y. P. Chan, T. S. J. Ho, B. K. J. Khoo, S. L. J. Wong, C. H. Thng, B. K. Chong, E. Y. Tan, V. K. M. Tan and A. S. G. Lee, *Cancers*, 2019, **11**, 1872.
11 R. Zou, S. Y. Loke, V. K.-M. Tan, S. T. Quek, P. Jagmohan, Y. C. Tang, P. Madhukumar, B. K.-T. Tan, W. S. Yong, Y. Sim, S. Z. Lim, E. Png, S. Y. S. Lee, M. Y. P. Chan, T. S. J. Ho, B. K. J. Khoo, S. L. J. Wong, C. H. Thng, B. K. Chong, Y. Y. Teo, H.-P. Too, M. Hartman, N. C. Tan, E. Y. Tan, S. C. Lee, L. Zhou and A. S. G. Lee, *Cancers*, 2021, **13**, 2130.
12 S. A. Bustin, *A-Z of Quantitative PCR*, International University Line, 2004.
13 S. C. Schuster, *Nat. Methods*, 2008, **5**, 16–18.
14 J. Shendure and H. Ji, *Nat. Biotechnol.*, 2008, **26**, 1135–1145.
15 I. C. C. Ferreira, E. M. G. Aguiar, A. T. F. Silva, L. L. D. Santos, L. Cardoso-Sousa, T. G. Araújo, D. W. Santos, L. R. Goulart, R. Sabino-Silva and Y. C. P. Maia, *J. Oncol.*, 2020, **2020**, 1–11.
16 V. E. Sitnikova, M. A. Kotkova, T. N. Nosenko, T. N. Kotkova, D. M. Martynova and M. V. Uspenskaya, *Talanta*, 2020, **214**, 120857.
17 N. Simsek Ozek, S. Tuna, A. E. Erson-Bensan and F. Severcan, *Analyst*, 2010, **135**, 3094.
18 J. M. Cameron, C. Rinaldi, H. J. Butler, M. G. Hegarty, P. M. Brennan, M. D. Jenkinson, K. Syed, K. M. Ashton, T. P. Dawson, D. S. Palmer and M. J. Baker, *Cancers*, 2020, **12**, 1710.
19 S. Zhang, Q. Y. S. Wu, Y. F. Chen, M. Hum, D. C. L. Wong, E. Y. Tan, A. S. G. Lee, J. Teng, U. S. Dinish and M. Olivo, *Nanoscale*, 2023, **15**, 10057–10066.
20 J. Perumal, A. Mahyuddin, G. Balasundaram, D. Goh, C. Y. Fu, A. Kazakeviciute, U. Dinish, M. Choolani and M. Olivo, *Cancer Manage. Res.*, 2019, **11**, 1115–1124.
21 K. J. I. Ember, M. A. Hoeve, S. L. McAughtrie, M. S. Bergholt, B. J. Dwyer, M. M. Stevens, K. Faulds, S. J. Forbes and C. J. Campbell, *npj Regener. Med.*, 2017, **2**, 12.
22 S. Zhang, Y. Qi, S. P. H. Tan, R. Bi and M. Olivo, *Biosensors*, 2023, **13**, 557.
23 C. Krafft and V. Sergo, *Spectroscopy*, 2006, **20**, 195–218.
24 S. Zhang, C. L. Wong, S. Zeng, R. Bi, K. Tai, K. Dholakia and M. Olivo, *Nanophotonics*, 2020, **10**, 259–293.
25 M. Turino, N. Pazos-Perez, L. Guerrini and R. A. Alvarez-Puebla, *RSC Adv.*, 2022, **12**, 845–859.
26 Z. Yao, Q. Zhang, W. Zhu, M. Galluzzi, W. Zhou, J. Li, A. V. Zayats and X. F. Yu, *Nanoscale*, 2021, **13**, 10133–10142.
27 G. I. Dovbeshko, V. I. Chegel, N. Y. Gridina, O. P. Repnytska, Y. M. Shirshov, V. P. Tryndiak, I. M. Todor and G. I. Solyanik, *Biopolymers*, 2002, **67**, 470–486.
28 F. Ni, R. Sheng and T. M. Cotton, *Anal. Chem.*, 1990, **62**, 1958–1963.
29 D. Li, L. Xia, Q. Zhou, L. Wang, D. Chen, X. Gao and Y. Li, *Anal. Chem.*, 2020, **92**, 12769–12773.
30 Y. Li, T. Gao, G. Xu, X. Xiang, B. Zhao, X. X. Han and X. Guo, *Anal. Chem.*, 2019, **91**, 7980–7984.
31 T. Gomes Rios, G. Larios, B. Marangoni, S. L. Oliveira, C. Cena and C. Alberto do Nascimento Ramos, *Spectrochim. Acta, Part A*, 2021, **261**, 120036.
32 F. Geinguenaud, V. Militello and V. Arluison, Application of FTIR Spectroscopy to Analyze RNA Structure, in *RNA Spectroscopy, Methods in Molecular Biology*, ed. V. Arluison and F. Wien, Humana, New York, NY, 2020, vol. 2113, pp. 119–133.
33 M. Khanmohammadi, K. Ghasemi and A. B. Garmarudi, *RSC Adv.*, 2014, **4**, 41484–41490.
34 N. M. Ralbovsky and I. K. Lednev, *Chem. Soc. Rev.*, 2020, **49**, 7428–7453.
35 C. A. Meza Ramirez, M. Greenop, Y. A. Almoshawah, P. L. Martin Hirsch and I. U. Rehman, *Expert Rev. Mol. Diagn.*, 2023, **23**, 375–390.
36 H. Li, W. Geng, Z. Zheng, S. A. Haruna and Q. Chen, *Food Chem.*, 2023, **418**, 135998.
37 W. F. d. C. Rocha, C. B. do Prado and N. Blonder, *Molecules*, 2020, **25**, 3025.
38 F. Bella, J. Popovic, A. Lamberti, E. Tresso, C. Gerbaldi and J. Maier, *ACS Appl. Mater. Interfaces*, 2017, **9**, 37797–37803.
39 F. Bella, A. Sacco, G. Massaglia, A. Chiodoni, C. F. Pirri and M. Quaglio, *Nanoscale*, 2015, **7**, 12010–12017.
40 A. Carella, R. Centore, F. Borbone, M. Toscanesi, M. Trifuoggi, F. Bella, C. Gerbaldi, S. Galliano, E. Schiavo, A. Massaro, A. B. Muñoz-García and M. Pavone, *Electrochim. Acta*, 2018, **292**, 805–816.
41 M. Reina, A. Scalia, G. Auxilia, M. Fontana, F. Bella, S. Ferrero and A. Lamberti, *Adv. Sustainable Syst.*, 2022, **6**, 2100228.
42 J. Perumal, U. Dinish, A. Bendt, A. Kazakeviciute, C. Y. Fu, I. L. H. Ong and M. Olivo, *Int. J. Nanomed.*, 2018, **13**, 6029–6038.
43 J. Perumal, Y. Wang, A. B. E. Attia, U. S. Dinish and M. Olivo, *Nanoscale*, 2021, **13**, 553–580.
44 M. Banyay, M. Sarkar and A. Gräslund, *Biophys. Chem.*, 2003, **104**, 477–488.
45 Merck, *IR Spectrum Table & Chart*, 2023.
46 J. M. Berg and J. L. Tymoczko, *Biochemistry*, W. H. Freeman, 8th edn, 2015.

47 D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, W. H. Freeman and Company, New York, 7th edn, 2016.
48 M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner and F. L. Martin, *Nat. Protoc.*, 2014, **9**, 1771–1791.
49 K. Kong, C. Kendall, N. Stone and I. Notingher, *Adv. Drug Delivery Rev.*, 2015, **89**, 121–134.
50 K. S. Wang, G. Yu, C. Xu, X. H. Meng, J. Zhou, C. Zheng, Z. Deng, L. Shang, R. Liu, S. Su, X. Zhou, Q. Li, J. Li, J. Wang, K. Ma, J. Qi, Z. Hu, P. Tang, J. Deng, X. Qiu, B. Y. Li, W. D. Shen, R. P. Quan, J. T. Yang, L. Y. Huang, Y. Xiao, Z. C. Yang, Z. Li, S. C. Wang, H. Ren, C. Liang, W. Guo, Y. Li, H. Xiao, Y. Gu, J. P. Yun, D. Huang, Z. Song, X. Fan, L. Chen, X. Yan, Z. Li, Z. C. Huang, J. Huang, J. Luttrell, C. Y. Zhang, W. Zhou, K. Zhang, C. Yi, C. Wu, H. Shen, Y. P. Wang, H. M. Xiao and H. W. Deng, *BMC Med.*, 2021, **19**, 76.
51 Y. Tolkach, T. Dohmgörgen, M. Toma and G. Kristiansen, *Nat. Mach. Intell.*, 2020, **2**, 411–418.
52 E. E. Siddig, N. A. Mhmoud, S. M. Bakhiet, O. B. Abdallah, S. O. Mekki, N. I. El Dawi, W. Van de Sande and A. H. Fahal, *PLoS Neglected Trop. Dis.*, 2019, **13**, e0007056.