Supplementary Information (SI) for Analytical Methods.
This journal is © The Royal Society of Chemistry 2024

Supporting Information

Early Colorectal Cancer Detection: A Serum Analysis Platform Combining SERS and Machine Learning

Miao Zhu[b, c, †], Yubin Han[a, †], Yitong Qiu[d, †], Yang Shen[a], Qingcheng Xu[e], Ya Huang[f], Tiantian Li[g], Mei Sun[b, c, *], Weiyu Pu[a, *]

[a]Department of Ultrasound intervention, Changshu Hospital Affiliated to Nanjing University of Chinese Medicine, Changshu 215500, China

[b]Department of Hematology, Northern Jiangsu People's Hospital Affiliated to Yangzhou University, Yangzhou 225001, China

[c]Yangzhou Institute of Hematology, Yangzhou 225001, China

[d]Math and applied math, Ocean University of China, Qingdao 266100, China

[e]Department of Digestive Diseases, Northern Jiangsu People's Hospital Affiliated to Yangzhou University, Yangzhou 225001, China

[f]Department of Oncology, Northern Jiangsu People's Hospital Affiliated to Yangzhou University, Yangzhou 225001, China

[g]Department of Ultrasound, Northern Jiangsu People's Hospital Affiliated to Yangzhou University, Yangzhou 225001, China

[†]Miao Zhu, Yubin Han and Yitong Qiu contributed equally to the work.

*Corresponding Authors Weiyu Pu: rainypwy@163.com

*Corresponding Authors Mei Sun: hematologysunmei@163.com

1. Materials and methods

1.1. Materials

Throughout the experiments, ultrapure water (resistivity >18 MΩ) from Shanghai Hitech Instruments Co., Ltd. was used. Chloroauric acid tetrahydrate ($HAuCl_4 \cdot 4H_2O$), anhydrous ethanol, and dopamine hydrochloride were purchased from Jiangsu Lianhua Technology Co., Ltd. 4-Mercaptobenzoic acid (4-MBA) and (3-aminopropyl) triethoxysilane (APTES) were provided by Jiangsu Subai Chemical Co., Ltd. HCT-116 cells were purchased from Procell Life Science & Technology Co., Ltd. (Wuhan, China). All glassware was cleaned with aqua regia and ultrapure water before use.

1.2. Synthesis of AuNCs

The synthesis of AuNCs followed these steps: first, 0.3 mL of $HAuCl_4$ solution (30 mM) was mixed with 15 mL of ultrapure water under vigorous stirring. Next, 1.8 mL of dopamine hydrochloride solution (60 mM) was added. Finally, the mixture was slowly heated to 55°C and maintained at that temperature for approximately 40 min to synthesize the AuNCs.

1.3. Preparation of AuNCs substrate

First, the silicon wafer was cleaned with aqua regia and then rinsed four times with ultrapure water. Next, the silicon wafer was cleaned twice with anhydrous ethanol and dried at 90°C in an air oven for 1.5 hours. After drying, the silicon wafer was immersed in a 2% (v/v) APTES solution of anhydrous ethanol for 20 hours to form a self-assembled layer, followed by thorough rinsing with ethanol and air drying. Subsequently, the APTES-functionalized silicon wafer was placed horizontally in an AuNCs colloidal solution for 10 hours. AuNCs were deposited onto the surface of the APTES-functionalized silicon wafer, forming an AuNCs layer. The thiol moiety of APTES covalently bonded to the substrate while its pendant primary amine group covalently bonded to the nanoparticles. Ultimately, the AuNCs substrate was successfully prepared.

1.4. Establishment of CRC mice model

The establishment of the CRC mice model followed these steps: first, HCT-116 cells were cultured. HCT-116 cells were purchased from Procell Life Science&Technology Co., Ltd. (Wuhan, China) and identified by STR. No mycoplasma was detected in the cells, HCT-116 cells were cultured in a special culture medium. The special culture medium for HCT-116 cells was purchased from Procell Life Science & Technology Co., Ltd. (Wuhan, China), and the culture dishes were placed in an incubator at a constant temperature of 37 ℃ and 5% carbon dioxide. The cells were digested and subcultured with 0.25% trypsin (Beyotime Biotechnology, China, Cat.No.C0201). Next, the xenotransplantation model in nude mice was established. Experimental nude mice (Comparative Medical Center of Yang zhou University) were raised to 5-6 weeks old, and HCT-116 cells (red fluorescent label) in logarithmic growth period were selected to make cell suspension. Under aseptic conditions, the cell suspension was slowly injected into the right armpit of experimental mice at an angle of 15 degrees, and tumor formation was observed every three days. Finally, blood collection and in vivo animal imaging were performed. The experimental mice were put into an animal living imaging instrument (animal living imaging technology Perkin Elmer, USA) after general anesthesia with an animal anesthesia instrument, and the fluorescence expression of tumor-bearing mice was observed. After imaging, pinched the mice neck, held a capillary tube and pierced the inner canthus of the right or left eye at a 45-degree angle, collected blood, and then compressed the mice with a dry cotton ball to stop bleeding. The frequency of detection was once a week up to three weeks, then the mice were sacrificed and the tumor was taken out and fixed with paraformaldehyde. Finally, tumor tissue sections was HE staining. The tissue was embedded in paraffin and sliced, and the dewaxed tissue was washed with PBST 2-3 times for 5 minutes each time. Slices are dyed in hematoxylin for 3-5 minutes, washed and dyed in eosin dye solution for 1-3 minutes, and then dehydrated and sealed. Images were obtained by Orthographic microscope (Olympus, Japan), at 20x magnification.

1.5. SERS measurement

Blood were collected from tumor-bearing mice model, with approximately 100 mL of blood drawn from each mice. The serum was separated by centrifugation at 3000 rpm for 10 minutes to remove cells and impurities. Then, about 10 μL of the serum sample was directly added to the AuNCs substrate, ensuring even distribution on the substrate surface. Next, the AuNCs substrate was placed on the sample stage of the Raman spectrometer. The Raman spectrometer parameters were set as follows: laser wavelength at 785 nm, laser power at 60 mW, and integration time of 20 seconds. Each serum sample was measured five times to ensure data accuracy. The laser focus was adjusted to align with the sample before performing the spectral measurement. After the measurement, the raw SERS spectral data were recorded and analyzed.

1.6. Instrumentation

Raman spectra were recorded with an inVia Raman spectrometer from Renishaw (UK). SEM images were obtained using a JSM-7610F scanning electron microscope from JEOL (Japan). Absorption spectra were measured with a Cary 60 UV-Vis spectrometer from Agilent Technologies (USA). Blood was separated using a TG16-WS high-speed centrifuge from Xiangyi (China). Silicon wafer was dried in a DZF-6050 drying oven from Shanghai Yiheng (China). The special culture medium for HCT-116 cells was purchased from Procell Life Science & Technology Co., Ltd. (Wuhan, China),

1.7. Data preprocessing and multivariate analysis

All raw SERS spectra were preprocessed using OriginPro 2022 software (OriginLab Corporation, USA). First, the spectra were baseline-corrected to remove any background noise and fluorescence interference. A polynomial order of 6 was used for the baseline correction. Next, the spectra were smoothed using a Savitzky-Golay filter with a window size of 11 points and a polynomial order of 3 to reduce noise without distorting the signal. Finally, the corrected and smoothed spectra were normalized to the maximum intensity to ensure comparability across different spectra. After all SERS spectra normalization, the average spectra for different stages were calculated to compare Raman peak differences. Then, Principal Component Analysis (PCA) was performed on all normalized SERS spectra using MATLAB R2022a

(MathWorks, USA). The PCA score plot, scree plot, eigenvalue plot, and loadings plot were all generated in MATLAB. The scree plot showed the contribution of each principal component (PC) to the total variance, while the score plot displayed the distribution of SERS spectra in the PCA space. The loadings plot revealed the Key characteristic peaks influencing PC. Subsequently, the PCA-DWNN model was applied to classify the SERS spectra. The first PC to the first fourteen PCs were sequentially used as input features in the DWNN. DWNN was configured with a choice of K values ranging from 1 to 10, and the optimal value was identified through Leave-One-Out Cross Validation (LOOCV) to balance classification accuracy and model complexity. For DWNN, In the weight strategy, the learning rate η was set between 0.01 and 0.1 to ensure weight updates were neither too rapid, causing oscillations, nor too slow, hindering learning progress. Weight updates employed an inverse distance weighting method, where ε was set at 0.001. During the training process, neighbor weights were dynamically adjusted based on each neighbor's contribution to the classification outcome.

2. Results and discussion

2.1. Average SERS spectra of serum of CRC mice at different stages

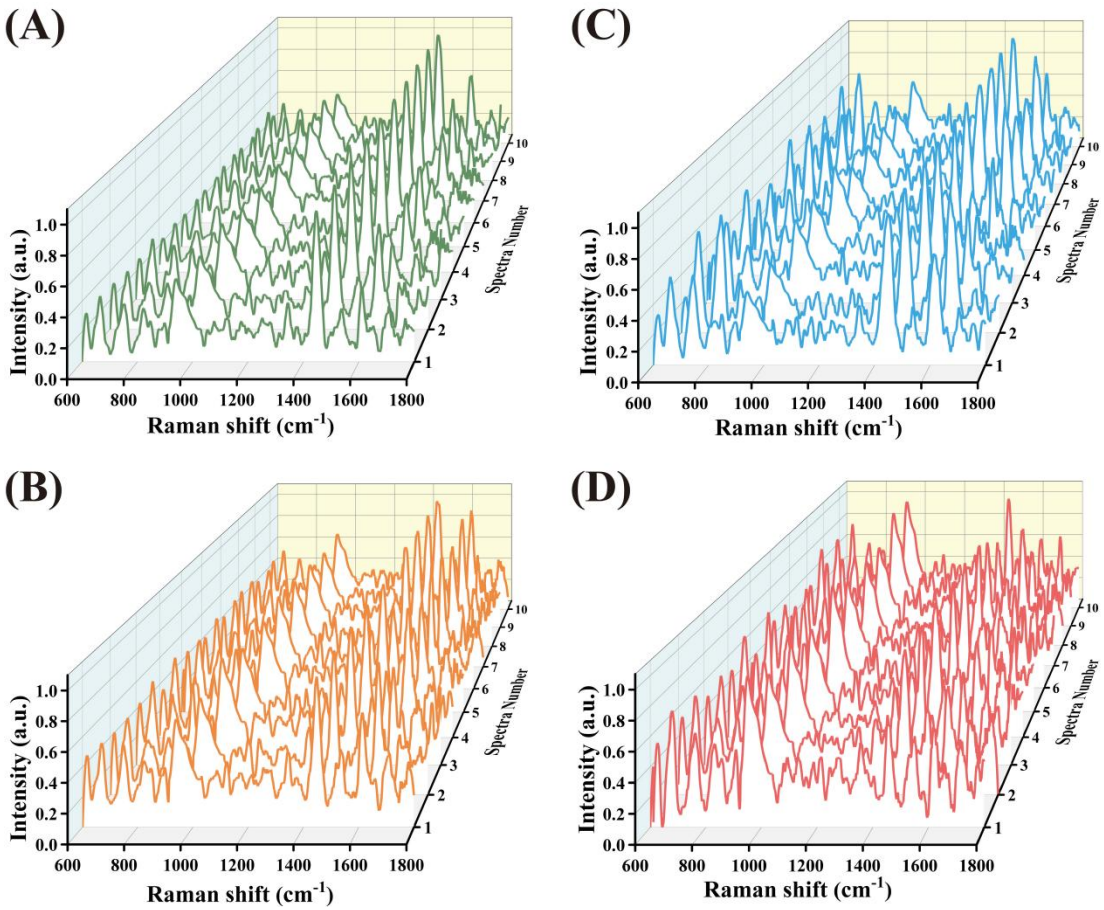

Fig. S1 The SERS spectra of the serum from 10 mice at (A) 0 Day, (B) 7 Days, (C) 14 Days, and (D) 21 Days.

Table S1 The detailed assignments of characteristic peaks

Raman shift (cm^{-1})	Assignment	Reference
617	C-C twisting (protein)	1
667	C-S stretching mode of cystine (collagen type I)	2, 3
747	T (ring breathing mode of DNA/RNA bases)	1
823	Phosphodiester	4
846	Monosaccharides (α-glucose), (C-O-C) skeletal mode	5
898	Monosaccharides (β-glucose), (C-O-C) skeletal mode	5
939	Proline, hydroxyproline, ν (C-C) skeletal of collagen backbone	3
1031	δ (C-H), phenylalanine	6
1123	C-C stretching mode of lipids & protein, C-N stretch	2, 7
1176	C-H bending tyrosine (proteins)	8

1201	Amide III (proteins)	8
1273	Amide III (proteins)	8
1343	CH3, CH2 wagging (collagen assignment)	9
1446	CH_2 bending mode of proteins & lipids	2, 7
1534	Amide carbonyl group vibrations and aromatic hydrogens	10
1613	Tyrosine	3
1710	One of absorption positions for the C==O stretching vibrations of cortisone	11
1752	C==O (lipid)	12

2.2 Multivariate analysis

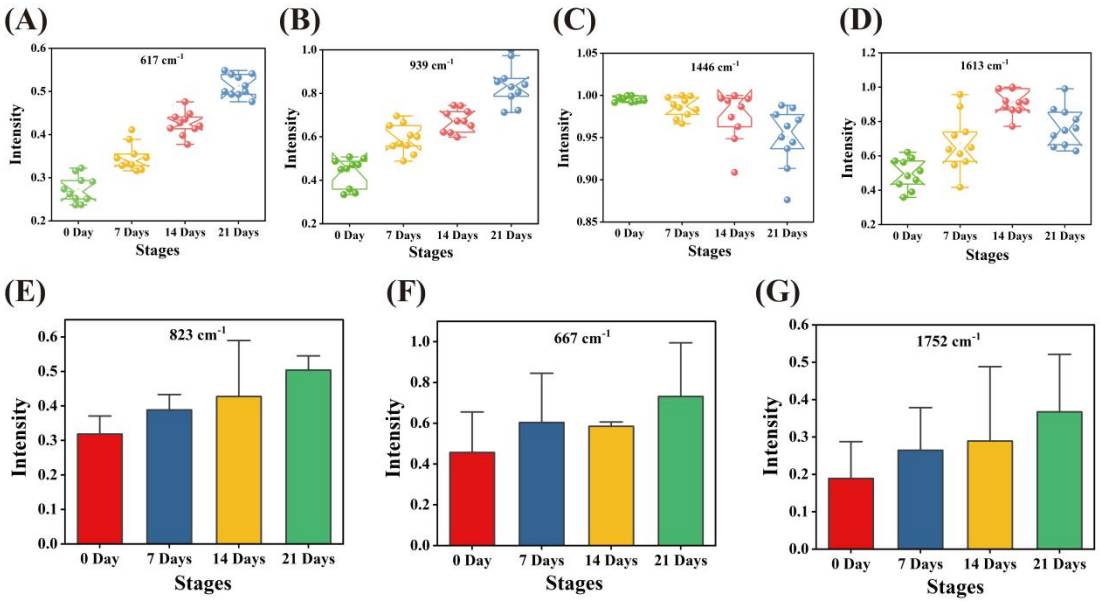

Fig. S2 The scatter plots of characteristic peaks at (A) 617, (B) 939, (C) 1446, and (D) 1613 cm^{-1}. The bar charts of intensity changes at (E) 823, (F) 667, and (G) 1752 cm^{-1}.

This principle diagram demonstrates the process of the Dynamically Weighted Nearest Neighbors (DWNN) algorithm for SERS spectra classification (Fig. S3). In the diagram, the query point (Query x) is marked with a red star, while the two classes of sample points (class w_1 and class w_2) are represented by magenta crosses and cyan crosses, respectively. The blue circles indicate the nearest neighbors (y_1, y_2, y_3, and y_4) selected, with arrows pointing to the query point to show their relationship. DWNN should be performed according to the following steps:

First, the spectral data is standardized to ensure all features are on the same scale:

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{1}$$

where μ and σ are the mean and standard deviation of the features, respectively.

Next, through dynamic weight calculation, the weights are adjusted based on the classification error rate for each sample, so that neighbors with higher similarity have a greater influence on the classification result. The initial weight is w_i, and the loss function L is defined as the classification error rate:

$$L = \frac{1}{N}\sum_{i=1}^{N} \Pi\ (\hat{y}_i \neq y_i) \tag{2}$$

where $\hat{y}_i$ is the predicted label, y_i is the true label, and Π is the indicator function that takes the value 1 if the predicted label does not match the true label, and 0 otherwise. The weight update formula is:

$$w_i^{t+1} = w_i^t - \eta \frac{\partial L}{\partial w_i^t} \tag{3}$$

where η is the learning rate. In practice, a simplified weight adjustment method can be used, such as inverse distance weighting:

$$w_i = \frac{1}{d(x, y_i) + \epsilon} \tag{4}$$

where ϵ is a small positive constant to prevent division by zero.

In the preprocessed feature space, the Euclidean distance between the query sample and the training samples is calculated:

$$d(x, y_i) = \sqrt{\sum_{j=1}^{N} (x_j - y_{ij})^2} \tag{5}$$

The K nearest neighbors are selected based on the shortest distances. Using the dynamically calculated weights, a weighted vote is performed on the neighbors' classes, and the class with the highest vote count is chosen as the predicted result:

$$\hat{y} = arg\ max_{c \in C} \sum_{i=1}^{K} w_i \cdot \prod(y_i = c) \tag{6}$$

where C is the set of all classes, and $\prod(y_i = c)$ is the indicator function that takes the value 1 if the neighbor y_i belongs to class c, and 0 otherwise.

This method improves the accuracy and efficiency of spectral data classification by dynamically adjusting neighbor weights. The principle diagram clearly shows the entire process, from input data to output classification results.

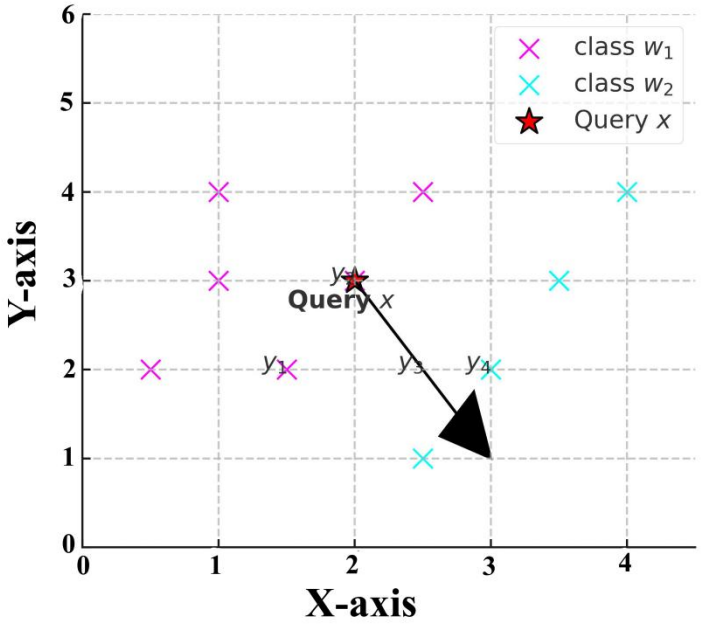

Fig. S3 The classification principle diagram of DWNN in two-dimensional space.

Fig. S4 illustrated the confusion matrices of different models for the classification performance on SERS spectra of CRC mice. The PCA-DWNN confusion matrix (Fig. S4A) showed that nearly all SERS spectra were correctly classified, with only one 14 Days spectrum misclassified as 7 Days spectrum, resulting in an accuracy of 97.5%. The PCA-KNN confusion matrix (Fig. S4B) performed well on 0 Day and 14 Days, but had misclassifications on 7 Days and 21 Days, with two 7 Days and one 21 Days spectra misclassified as 14 Days, resulting in an accuracy of 92.5%. The PCA-SVM confusion matrix (Fig. S4C) performed well on 0 Day, but had one misclassification at each of the other stages, resulting in an accuracy of 92.5%. The SVM confusion matrix (Fig. S4D) performed well at 0 Day, but had more misclassifications at 7 Days,

with three 7 Days spectra misclassified as 14 Days, resulting in an accuracy of 87.5%. The KNN confusion matrix (Fig. S4E) performed well at 21 days, but had more misclassifications at 14 Days, with two 14 Days spectra misclassified as 7 Days and 21 Days, respectively, resulting in an accuracy of 90.0%. The comparison of these confusion matrices showed that the PCA-DWNN model performed the best, while the PCA combined with KNN and SVM models also performed well at most stages. The PCA-LDA confusion matrix (Fig. S4F) performed well on 0 Day and 21 Days, but had misclassifications on 7 Days and 21 Days, with one 7 Days and one 14 Days spectra misclassified as 14 Days and 7 Days, resulting in an accuracy of 95.0%.

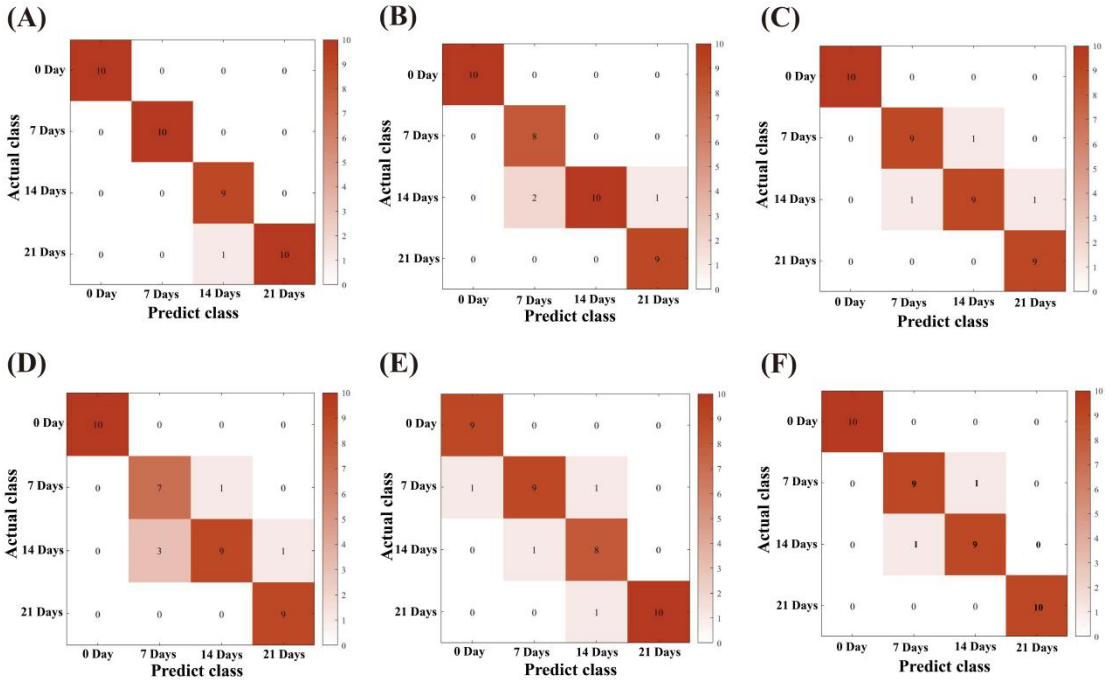

Fig. S4 (A) PCA-DWNN Confusion Matrix. (B) PCA-KNN Confusion Matrix. (C) PCA-SVM Confusion Matrix. (D) SVM Confusion Matrix. (E) KNN Confusion Matrix. (F) PCA-LDA Confusion Matrix.

References

1. J. W. Chan et al., *Biophys. J.*, 2006, **90**, 648-656.
2. N. Stone et al., *J. Raman Spectrosc.*, 2002, **33**, 564-573.
3. W. T. Cheng et al., *Microsc. Res. Tech.*, 2005, **68**, 75-79.
4. A. J. Ruiz-Chica et al., *J. Raman Spectrosc.*, 2004, **35**, 93-100.
5. G. Shetty et al., *Br. J. Cancer*, 2006, **94**, 1460-1464.

6. Z. Huang et al., *Int. J. Cancer*, 2003, **107**, 1047-1052.
7. N. Stone et al., *Faraday Discuss.,* 2004, **126**, 141-157.
8. I. Notingher et al., *J. R. Soc. Interface*, 2004, **1**, 79-90.
9. C. J. Frank et al., *Anal. Chem.*, 1995, **67**, 777-783.
10. B. R. Wood et al., *Biospectroscopy*, 1998, **4**, 75-91.
11. R. A. Shaw and H. H. Mantsch, *J. Mol. Struct.,* 1999, **480-481**, 1-13.
12. R. J. Lakshimi et al., *Radiat. Res.*, 2002, **157**, 175-182.