Official Journal of CIOMP
elight.springeropen.com

LETTER

Open Access

Early cancer detection by serum biomolecular fingerprinting spectroscopy with machine learning

Shilian Dong[1†], Dong He[1†], Qian Zhang[2†], Chaoning Huang[1], Zhiheng Hu[3], Chenyang Zhang[1], Lei Nie[4], Kun Wang[4], Wei Luo[5], Jing Yu[6], Bin Tian[7], Wei Wu[7], Xu Chen[3], Fubing Wang[2,10*], Jing Hu[8,9*] and Xiangheng Xiao[1,10*]

Abstract

Label-free surface-enhanced Raman scattering (SERS) technique with ultra-sensitivity becomes more and more desirable in biomedical analysis, which is yet hindered by inefficient follow-up data analysis. Here we report an integrative method based on SERS and Artificial Intelligence for Cancer Screening (SERS-AICS) for liquid biopsy such as serum via silver nanowires, combining molecular vibrational signals processing with large-scale data mining algorithm. According to 382 healthy controls and 1582 patients from two independent cohorts, SERS-AICS not only distinguishes pan-cancer patients from health controls with 95.81% overall accuracy and 95.87% sensitivity at 95.40% specificity, but also screens out those samples at early cancer stage. The supereminent efficiency potentiates SERS-AICS a promising tool for detecting cancer with broader types at earlier stage, accompanying with the establishment of a data platform for further deep analysis.

[†]Shilian Dong, Dong He and Qian Zhang contributed equally to this work.

*Correspondence:
Fubing Wang
wfb20042002@sina.com
Jing Hu
jinghu_somed@uestc.edu.cn
Xiangheng Xiao
xxh@whu.edu.cn
Full list of author information is available at the end of the article

Cancer is a major public health problem and the second leading cause of death worldwide, and China alone accounts for approximately 4,820,000 new cancer cases and 3,210,000 cancer deaths in 2022 [1]. Increasing survival rates of cancer patients mainly rely on early intervention, which is largely determined by earlier screening and diagnosis of lesions. Compared to imaging or histopathology, laboratory test particularly using blood, urine or other liquid biopsy, is a low-cost, non-invasive and easily repeated manner for early cancer prediction by detecting specific cancerous biomarkers such as circulating tumor DNA, proteins, cancer metabolites, and even cell-derived exosomes and circulating tumor cells [2–4]. But challenges still remain, which include: 1) no valid and plentiful tumor biomarkers for diverse cancer types; 2) no widely viable approaches for cancer detection especially at asymptomatic stages; 3) no comprehensive analytical platform for big datasets to differentiate healthy and cancerous populations [5]. In order to tackle these problems, surface-enhanced Raman scattering (SERS) technology has been recently introduced towards structurally amplified fingerprinting of low-concentration molecules and can distinguish subtle changes between healthy people and cancer patients [6]. Despite of its advantages in sensitivity and selectivity, these reported SERS-based cancer detection methods either focus on single or a few biomarkers tested in limited cancer types with insufficient samples [7], or stick on the preliminary stage lacking friendly data interpretation based on more efficiently high-throughput analysis [8]. To overcome the above bottlenecks, a few pioneering attempts have been made for optimization of either SERS technique or data analysis. For example, in SERS probe construction, researchers have proposed biocompatible top-down or bottom-up SERS platforms for cellular analysis for monitoring multiple biomolecules [9–12]. But the bio-specific binding currently relied upon for these new methods makes testing for large volumes of samples expensive. And the coincubation process of nanomaterials with biological samples which new technologies exist, results in very time-consuming pre-processing, all of which limit the effective roll-out of universal screening for pan-cancer. In addition, Huang et. al proposed to use various machine learning models for dimensionality reduction, feature extraction and analysis of Raman data [13–16]. For example, combining principal component analysis with support vector machines (SVM) is able to achieve accurate identification of breast cancer subtypes [13]. But it is difficult for operators to obtain large sets of independent SERS data. In most cases, large sets of data were obtained by dependent and repeated measures. And previous effective dimension screening mechanism used in the algorithm model is not comprehensive enough. Often only less than 10 of the thousands of Raman dimensions could be selected for operation, which might output unsatisfactory results on another dataset due to overfitting. Therefore, a highly accurate SERS-based strategy with simple sample preparation and efficient fitting method for large-volume clinical samples is in critical needs for practical pan-cancer screening.

Here we analyzed a tremendous data set of 382 healthy controls and 1582 patients from two independent cohorts, and reported a label-free SERS-Artificial Intelligence combined method for Cancer Screening (SERS-AICS), attempting to consolidate the detection advantages of canonical SERS system with the calculation strengths in the most updated big data analytic tools (Fig. 1). We utilized as low as 15 μl of patient serum sample each for lung, colorectal, hepatic, gastric and esophageal cancers, which are among the most lethal cancer types in either gender in China [17–19], and examined it by liquid phase SERS appliance in a mixture with diameter around 70–120 nm Ag nanowires (Ag NWs) [20], which is a solider and more viable one than other plasmonic nano-substrates (Additional file 1: Figs S1, S2A). By comparing the SERS spectroscopy of same serum under different materials and morphologies of SERS probe (Au nanoparticles and Au bipyramid nanoparticles) enhancement, we found that Ag NWs nanowires still exhibit the best SERS enhancement even when compared to other SERS probe particles with salt-induced aggregation, shown in Additional file 1: Fig S1. Generally, the Raman characteristic peaks of organic compounds in biological samples are located in the range of 600–1800 cm^{-1}, which is also the fingerprint region of Raman spectra. Therefore, the testing range from 600 cm^{-1} to 1800 cm^{-1} was selected. The Raman peak at 957 cm^{-1} corresponds to cholesterol, which is generated by the C–C backbone vibration. The peaks at 1004 cm^{-1} and 1156 cm^{-1} are due to CC- bond stretching and the deformation of C-CH_3. The peaks from 1230 cm^{-1} to 1282 cm^{-1} correspond to amide III, which are due to the bending vibration of the peptide bond caused by the combination of N–H and Cα-H bonds. The peak of phospholipids (1447 cm^{-1}) is formed by the CH_3-CH_2 bending of phospholipids and the protein side chains. The peak at 1515 cm^{-1} is due to the vibration of cytosine [21–24]. Additionally, we tested the system using Ag NWs and found that it didn't introduce background noise (Additional file 1: Fig S2B) or experience dramatic changes in spectral peaks within 36 h processing (Additional file 1: Fig S2C). After acquiring the dichotomous raw data sets of SERS spectra from a large quantity of clinical serum samples including cancerous and healthy sources, we firstly performed pre-processing for dimensionality reduction and feature picking by using covariance matrices [25]. The trimmed data with

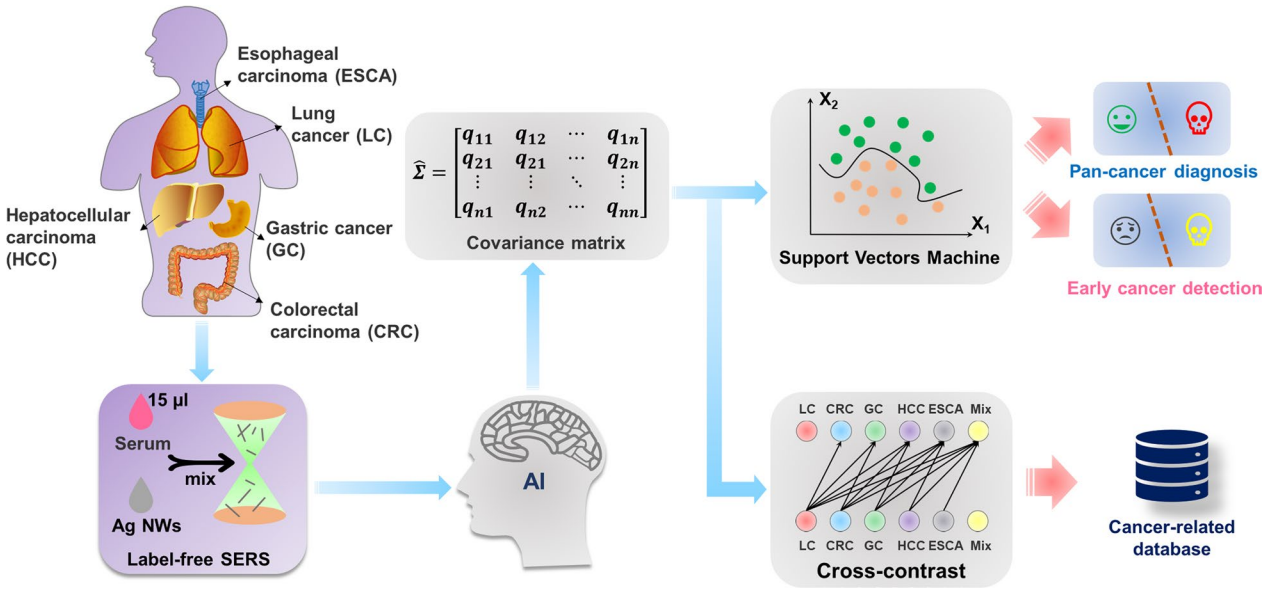

Fig. 1 The workflow diagram of SERS-AICS. A drop of 15 μl serum was used for label-free SERS spectrum detection by Ag NWs. After dimensionality reduction of covariance matrix of high-dimensional serum spectral data, cancer related dimension database of bond level can be obtained. The subsequent support vector machine algorithm not only carries out accurate pan-cancer screening for five kinds of cancers with high mortality, but also achieves the identification between early-stage cancers and common diseases

minimized features were further analyzed with the SVM model [26–29] for screening out patients of top-5 leading cancers from healthy people even at very early stage (Fig. 1). SVM is a classic binary classification machine learning model that has significantly lower computational requirements compared to those of deep learning models. A large number of past studies have shown that SVM performs exceptionally well in classifying tabular data, especially for non-linear dataset when using Gaussian kernel function. And it has also been reported in some cancer detection applications [13], so SVM was chosen as an appropriate tool for our SERS non-linear dataset from serum samples. Moreover, comparison of output features among different types of cancer via covariance matrix calculation, can not only screen out up to 50 Raman feature dimensions for classification and recognition of the algorithm model, but also continue to search out some potential shared dimensions for cancer lesion according to our accumulating database platform (Fig. 1).

The SERS-AICS assay was established and then tested for pan-cancer screening from internal cohort. The internal cohort composed of 1375 individuals that include 324 healthy controls (HC) and 1051 patients diagnosed with stage I–IV cancers according to American Joint Commission on Cancer (AJCC) [30]: lung cancer (LC, n = 244), colorectal carcinoma (CRC, n = 216), gastric cancer (GC, n = 195), hepatocellular carcinoma (HCC, n = 203), esophageal carcinoma (ESCA, n = 193). Each of these cancer types was further split into a training group and an internal validation group with a ratio of 8:2. And the allopatric cohort was collected for external validation, which consists of 58 healthy controls and 237 cancer patients. The full information with details of both the patients and controls is summarized in Additional file 1: Table S1. In this way, we guaranteed the quantities and diversity of clinical samples for following data collection and analysis based on SERS-AICS. Based on Ag NWs SERS probe which is suitable for serum analysis, we performed Raman tests and obtained SERS average spectrum for all serum samples from the above five cancers and healthy controls. We found no significant differences in the main Raman peaks for all samples, with only minor differences in some Raman peaks that were difficult to detect with the naked eye (Additional file 1: Fig S2D). Actually, one of the biggest challenges for the current label-free SERS technique is that the Raman spectra data, especially from serum with complex composition, are difficult to distinguish the subtle differences embedded in cancerous group from healthy controls. For Raman spectral data with large dimensionality, the dimensionality reduction step of the original algorithm data is crucial, which can largely reduce the difficulty of model construction. We firstly analyzed the differences in all the effective dimensions of 5 cancer samples and healthy control samples, and selected the top 10 effective dimensions with the largest differences to plot a violin diagram for each cancer or the mix group compared to healthy control (Fig. 2), which features a kernel density estimator

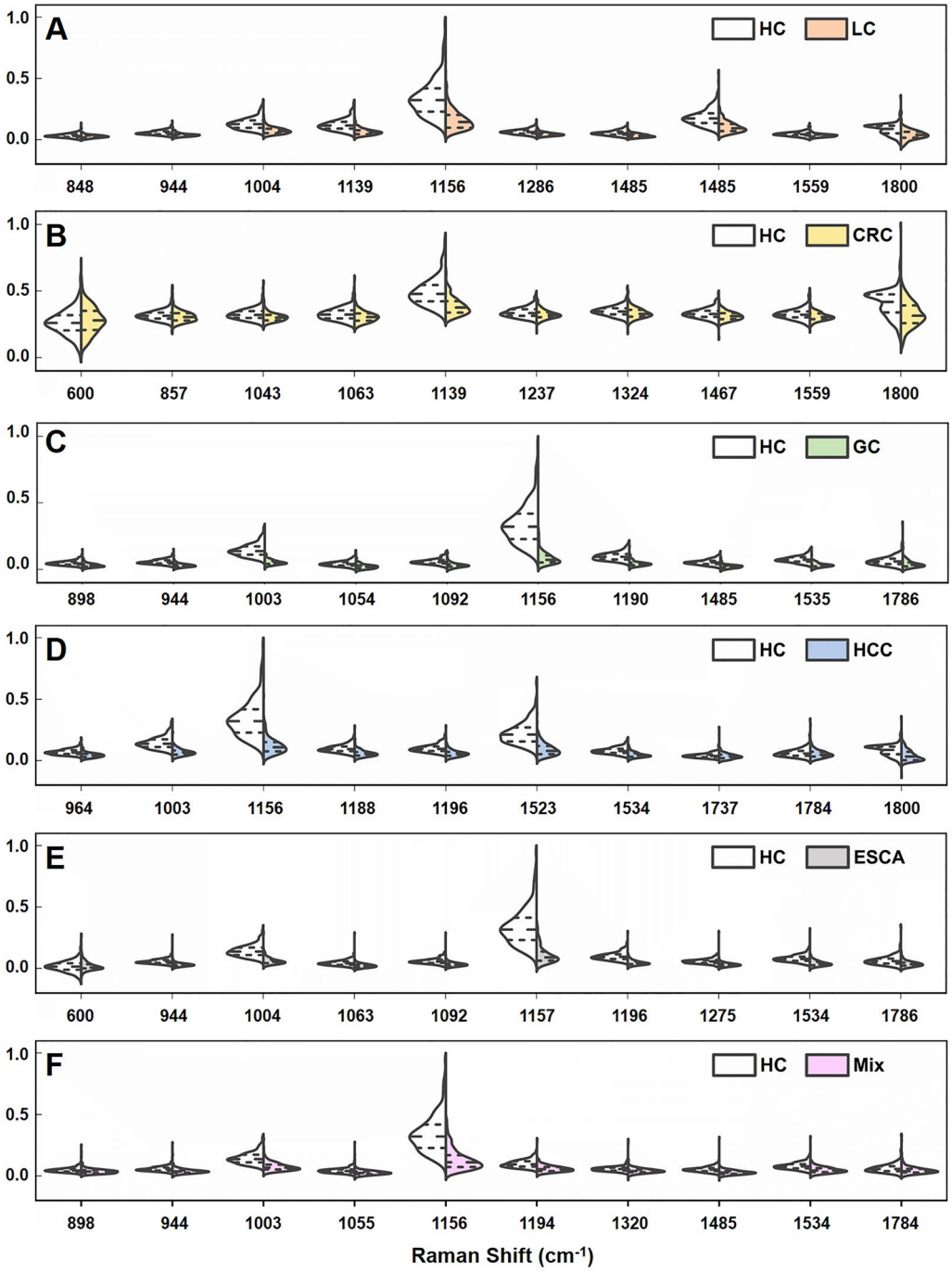

Fig. 2 Dimension discrimination of Raman spectral data for cancer groups versus healthy controls. Violin plots show intensity distributions at 10 typical Raman peaks of **A** LC, **B** CRC, **C** GC, **D** HCC, **E** ESCA or **F** mixture cancer group vs healthy control group selected by SERS-AICS. The middle dash lines indicate median value of the Raman intensities of each corresponding peak while the upper and lower dotted lines indicate intensity values of first quartile and third quartile. White represents healthy control group while the other colors represent different cancer type or mixture group

of the underlying distribution for the overall probability density. The cancer group showed a significant difference in intensity parameter from the healthy control at the 10 representative Raman shift peaks, which were preferential for the following SVM model. Interestingly, for most cancers, we found that the cancer/HC showed the

greatest difference in intensity at the peak of 1156 cm^{-1}, which overlapped the telescopic vibration of the C=O and implicated the predictive potentials of components harboring C=O in serum.

For pan-cancer detection, we firstly started from raw data of training cohort for LC, the largest sample group among five cancer types, to perform AICS-based analysis. Our analysis owns highly advances in coupling of covariance matrix calculation and the SVM modeling for more delicate characterization of slight alterations in spectra. In order to simplify the effective dimensions for the SVM model, we pretreated the raw SERS data by calculating the covariance of 195 LC samples together with 259 HC ones. Considering to reduce computational complexity and the impact of noise peaks, a larger interval would be better. However, an interval that is too large would result in significant information loss after dimensionality reduction, so we have chosen a moderate interval of 60 dimensions. All 1465 real dimensions ranged from 600 cm^{-1} to 1800 cm^{-1} in spectra were divided with an interval of 60 dimensions, and then the two features with the least correlation at all intervals with repetition were selected until all features were selected from the cancer group. The final representative 50 dimensions were then undergone by SVM calculation for finding nonlinear binary classification between cancerous and normal data. In order to evaluate the performance of the SERS-AICS system based on training data, we further analyzed the internal cohort of LC, which includes 49 patients and 65 healthy controls, for dimension reduction to 60 dimensions with the greatest differences in Raman spectrum for lung cancer (Additional file 1: Fig S3 and Additional file 1: Table S2). The dichotomies SVM classifier separated the 244 LC samples and 324 HC samples from both training and internal cohorts. As shown in Receiver Operating Characteristic (ROC) curve [31, 32], the area under the curve (AUC) [33] of the LC/HC model was 0.90 for discriminating lung cancer patients against healthy controls (Fig. 3A).

To test whether the SERS-AICS system applied for the other four cancers (CRC, GC, HCC and ESCA) with high death rates, we repeated the whole workflow for lung cancer and screened out the representative 50 dimensions with the greatest differences in serum-based Raman spectrum between patients with each cancer type and healthy controls (Additional file 1: Fig S4–S7 and Additional file 1: Tables S3–S6). The SVM classifier further generated AUC of each cancer model compared with the corresponding healthy control, which resulted in a value of 0.84 for CRC (Fig. 3B) and 0.99 for GC, 0.97 for HCC and 0.96 for ESCA (Fig. 3C–E). In order to mimic the real-life analysis of serum sample that is not known ahead of time as cancerous or healthy one, we also randomly selected 80 samples from each of the above five cancer serum samples to obtain a 400 mixed cancer sample set and found that our strategy still works well to pick up cancerous data unbiasedly. Consistently, we found that the AUC of the fusion model was 0.89 for 400 Mix /324 HC set, showing high identification results to separate pan-cancerous patients from healthy people (Fig. 3F, Additional file 1: Fig S8 and Additional file 1: Table S7).

Sensitivity and specificity are two fundamental measures for judging cancer detection, which implicates the ability of the test to correctly identify those individuals with or without cancer, separately. Thus, we further calculated the sensitivity, specificity and accuracy of 5 individual cancer groups and the mixed dataset. All data sets exhibited satisfying distinguishment from any cancerous samples to normal samples, showing the confusion matrix with an overall accuracy of 95.81% where the individual accuracy varied between 94.10% for LC and 98.25% for CRC (Fig. 3G). Similarly, the overall sensitivity of the confusion matrix was as high as 95.87%, exhibiting the individual sensitivity ranging from 91.84% for LC and 98.75% for mixed cancer among different cancer types, mixed cancer group and control (Fig. 3H). The overall specificity of the confusion matrix was as high as 95.40%, exhibiting the individual sensitivity ranging from 90.63% for mixed cancer and 98.57% for CRC among different cancer types, mixed cancer group and control (Fig. 3I). Besides, we further validated our analytic strategy in external datasets from allopatric cohort, and also acquired high separation between samples from 5 individual or mixed cancer groups and healthy controls (Additional file 1: Fig S9–S15 and Additional file 1: Table S8–S13). Notwithstanding SERS spectroscopy data with massive dimensions, our results showed that artificial intelligence algorithm can precisely capture the subtle differences between different spectral data sets of cancer samples and health control, and establish the classification of fusion models to distinguish them effectively with high accuracy and sensitivity at high specificity as well, suggesting its high potential for pan-cancer screening in clinical practice.

Failing to predict cancer before developing to later stage or discriminate against other diseases is another bottleneck in the field for reaching effective treatment of cancer and alleviating patients suffering. How to accurately cull samples at early stage out of common disease samples is very critical for the following diagnosis and treatment plan (Fig. 4A). To test whether our SERS-AICS method could achieve this goal, we collected non-cancerous disease-related clinical serum data sets from four types of disease and compared to their corresponding cancerous samples at AJCC stage I and II from internal cohort. Similar to pan-cancer characterization outputs,

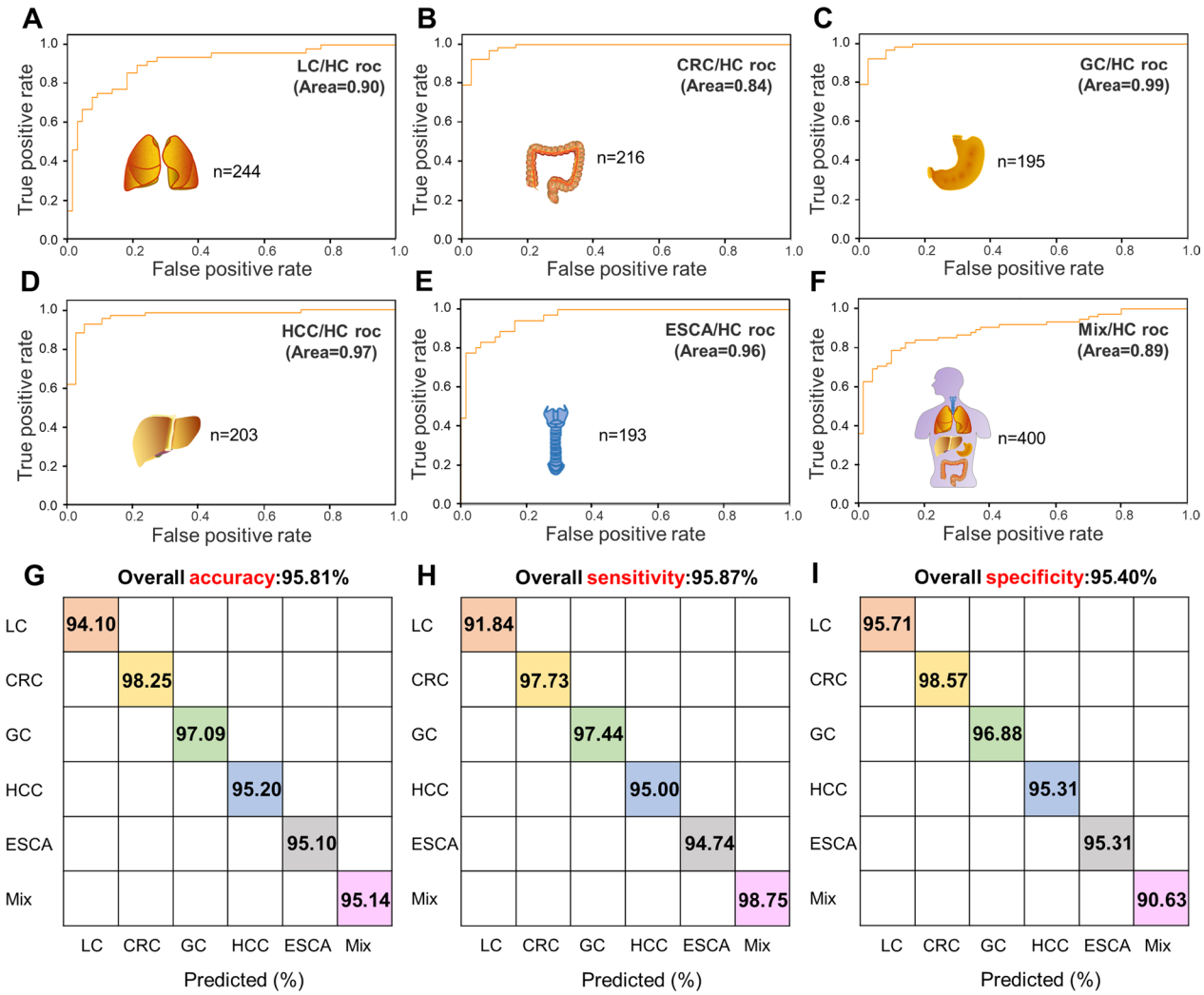

Fig. 3 SERS-AICS characterization of five cancers with high mortality. ROC curves with covariance matrices-assisted SVM model for distinguishing **A** 244 lung cancer patients, **B** 216 colorectal carcinoma patients, **C** 195 gastric cancer patients, **D** 203 hepatocellular carcinoma patients, **E** 193 esophageal carcinoma patients, **F** 400 mixture cancer patients from 324 healthy controls in the internal cohort. The **G** accuracy, **H** sensitivity and **I** specificity of single or multiple cancers/healthy control, the overall accuracy, sensitivity and specificity of all cancers could reach at 95.81%, 95.87%, 95.40%. The 400 mixed cancer patients were obtained by randomly selecting 80 samples from the five types of cancer each

our SERS-AICS strategy could also separate precancerous samples from those with other diseases, as the results showed an AUC value of 0.81 for 45 disease/35 early cancer set in lung, 0.94 for 42 disease/32 early cancer set in colorectum, 0.89 for 39 disease/36 early cancer set in gastric, and 0.93 for 33 disease/32 early cancer set in liver (Fig. 4B–E, Additional file 1: Figs S16–S19 and Additional file 1: Tables S14–S17). We determined the accuracy, sensitivity and specificity of four early cancers identification and found the best separation for gastric group sets with an accuracy of 93.33% and a sensitivity of 100% at 85.71% specificity, and hepatocellular group with an accuracy of 92.31% and a sensitivity of 85.71% at 100% specificity (Fig. 4F). For lung and colorectal datasets, we could still obtain overall satisfying prediction for early cancer patients out of population with non-cancerous diseases, except those with a slightly lower specificity of 77.78% and sensitivity of 66.67%, respectively (Fig. 4F), which may be due to variation from small number of samples for analysis but still provide a potential tool when there aren't good non-invasive screening methods for these two types of cancer at early stage. Thus, the SERS-AICS system described here might become an option for screening asymptomatic population before they develop cancer to advanced stages.

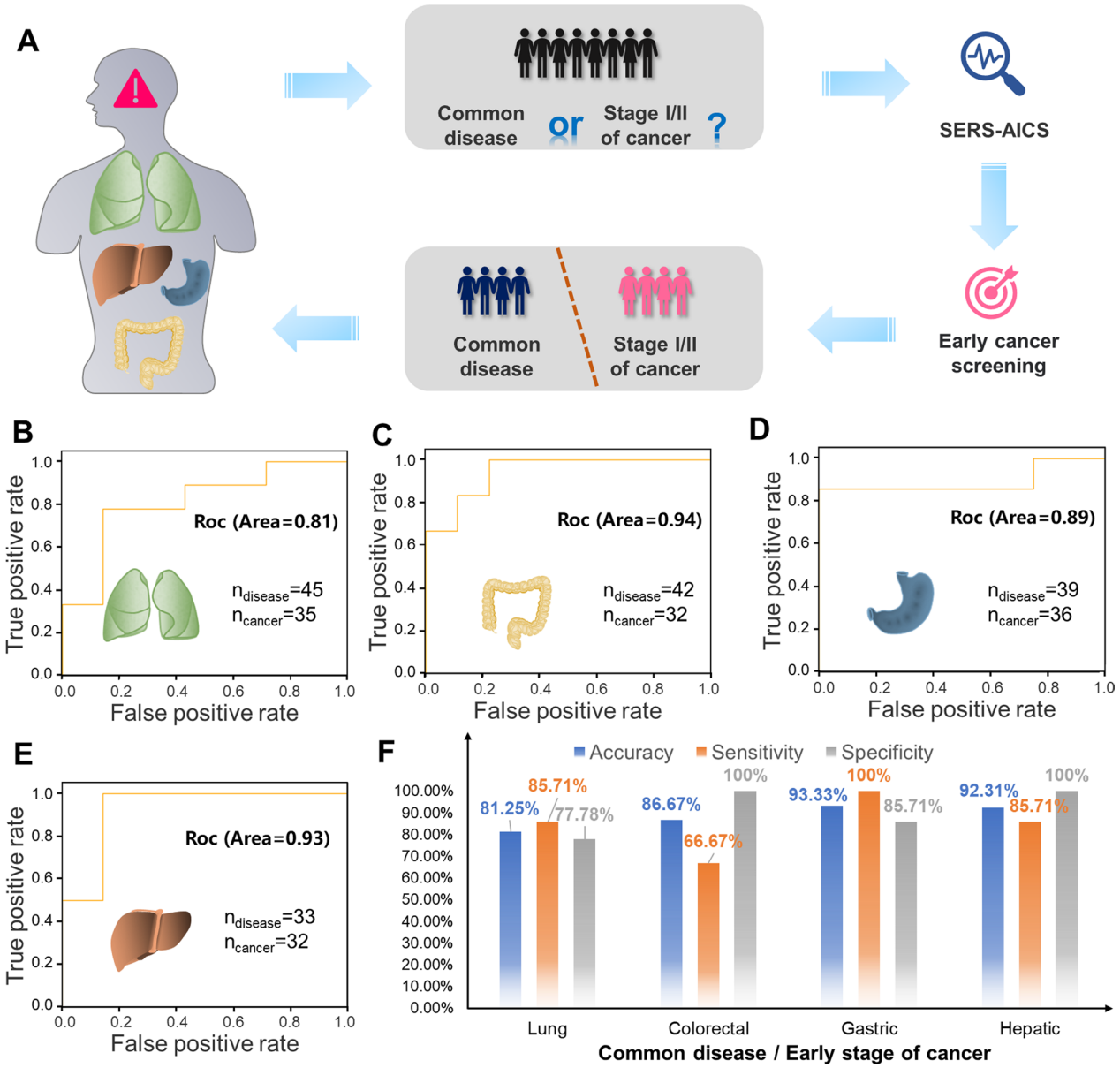

Fig. 4 Early screening for four representative cancers by SERS-AICS. **A** The SERS-AICS method could also effectively distinguish common diseases from early cancers in stage I and II with high accuracy. **B** ROC curves with covariance matrices-assisted SVM model for distinguishing 45 common disease patients from 35 early-stage cancer patients about lung. **C** ROC curves with covariance matrices-assisted SVM model for distinguishing 42 common disease patients from 32 early-stage cancer patients about colorectum. **D** ROC curves with covariance matrices-assisted SVM model for distinguishing 39 common disease patients from 36 early-stage of cancer patients about gastric. **E** ROC curves with covariance matrices-assisted SVM model for distinguishing 33 common disease patients from 32 early-stage cancer patients about liver. **F** The accuracy, sensitivity and specificity of different common disease/early stage of cancer

Another limitation impeding development of cancer screening is the lack of database to store, construct and trace the massive profiling of individual cancer patients, permitting ongoing deep analyses like exploring shared characteristic features as new cancer biomarkers. According to our data, the SERS-AICS detection/analysis system has not only collected and processed 1964 serum samples in total, but also exhibited high accuracy, sensitivity and specificity for identification of the top five most lethal cancers out of healthy controls. More importantly, our proposed covariance-assisted SVM classification strategy has unique advantages for the analysis of serum-based Raman data with a large sample pool approaching two thousand cases. This

allows for more reliable spectral-omics data across five representative cancers, providing an important guideline for future universal cancer screening (Fig. 5A). Taking lung cancer sample dataset as an example for dimension reduction, the 30 true dimensions of the first group were distributed between 600 cm^{-1} and 623.77051 cm^{-1}, two of which with a correlation closer to 0 representing more variability between dimensions [34], and the dimension at 600 cm^{-1} and 618.03276 cm^{-1} showed the minimal correlation in the representative heatmap [35–38] of the covariance matrix for lung cancer and healthy controls (Fig. 5B). Based on the attribution statistical analysis of the above cancer-related specificity dimensions, we can calculate the shared identical Raman peak positions of all the 5 types of cancers, any combination of different types of cancer, and/or comparison between precancerous condition and corresponding specific regular diseases (Fig. 5C and Additional file 1: Fig S20–S22), which are probably utilized as common features for cancer characterization. Compared to current individual molecular biomarkers, these distinct Raman peaks may implicate the full capacity of molecular vibrational spectrum for cancer screening [39–42], suggesting a potential measurement standard to correlate the spectral and biomolecular identity in future exploration.

Early detection of cancer, which is critical to improve survival rates, always comes with challenges including

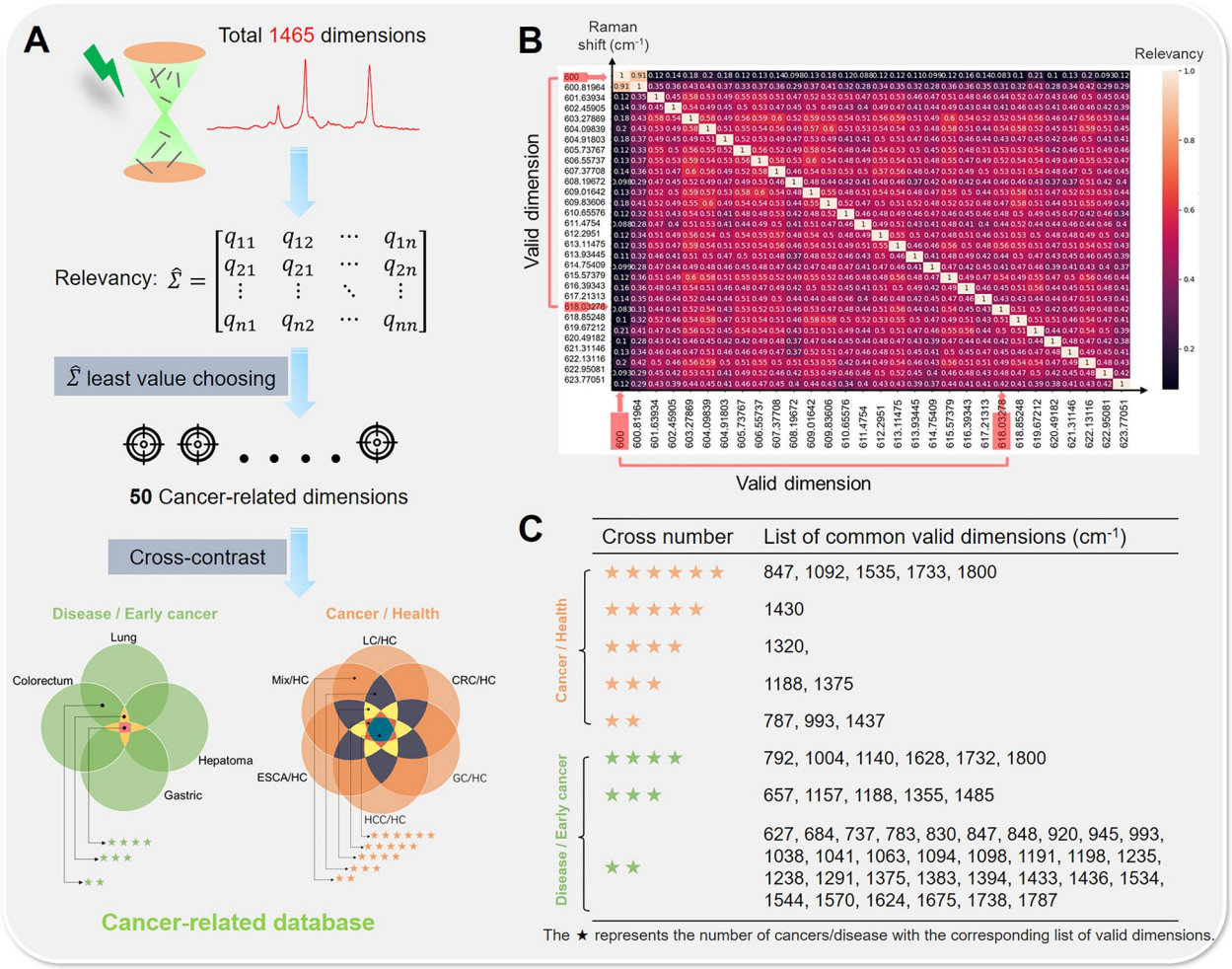

Fig. 5 Construction of cancer-related database at bond level by SERS-AICS. **A** The original spectral data of 1465 dimensions of serum with no obvious specificity can obtain 50 valid dimensions with the best specificity after the correlation between dimensions is judged by the covariance matrix, which related to the molecular bond energy information in serum associated with cancer or disease. **B** Heatmap of the covariance matrix formed on 30 true dimensions between 600 and 623.77051 cm^{-1} using peak data for lung cancer and healthy controls, the dimension at 600 cm^{-1} and 618.03276 cm^{-1} showed the minimal correlation. **C** List of common valid dimensions of different cancers compared with normal control group or common disease

the possibility of accurate prediction. In comparison to the currently developed cancer screening assays, our SERS-AICS technology provides a more sensitive signal capture based on Raman scattering technology, and a more reliable artificial intelligence algorithms coupling with unbiased dimension selection by covariance matrix and precise classification by SVM. The integrated SERS-AICS detection/analysis approach is able to not only distinguish major types of cancer from normal samples effectively, but also pick up samples at precancerous status out of the ones with non-cancerous diseases with high accuracies. Based on the multiplexing analysis described above, our assay can also construct a systematic cancer-related database simultaneously, providing a platform for information collection and management for future exploration. Moreover, unlike the generally applicable biomedical methods such as mass spectrometry and high throughput sequencing for focusing on specific protein/nucleic acid biomarkers, our SERS-AICS technique are specially designed to capture all unique vibrational spectra information of the whole molecules within the serum samples and rebuild an analytical model by all-inclusive profiling for better discrimination of subtle dimensional differences among samples from cancer patients and healthy people, providing a "panorama" view for cancer identification at molecular energy level. In addition, our non-probe method could reflect subsistent features of the sample to avoid possible false positives and also harbor the advantages of fast spectral acquisition, low cost and high throughout, implicating a solid and affordable approach for clinical translation. In summary, the SERS-AICS technique provides a promising comprehensive tool for real-world cancer detection in conjunction with routine physical exams for both ordinary population at risk and cancer patients. And it could also serve as a feasible preceding test before imaging tests for a definitive diagnosis. The SERS-AICS tool aims to further expand the application to a variety of cancer types, extend the monitoring throughout the management of cancer patients especially at very early stage, and establish a modeling system for large-scale data recording, restoring and researching.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43593-023-00051-5.

Additional file 1: Figure S1. Comparison of SERS spectra from the same serum sample based on different SERS probes. **Figure S2.** The morphology and SERS detection performance characterization of Ag nanowire. **Figure S3.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for internal lung cancer and healthy controls. **Figure S4.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for internal colorectal carcinoma and healthy controls. **Figure S5.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for internal gastric cancer and healthy controls. **Figure S6.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for internal hepatocellular carcinoma and healthy controls. **Figure S7.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for internal esophageal carcinoma and healthy controls. **Figure S8.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for internal mixture cancer and healthy controls. **Figure S9.** SERS-AICS system-based identification and characterization of five external cancers with high mortality. **Figure S10.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for external lung cancer and healthy controls. **Figure S11.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for external colorectal carcinoma and healthy controls. **Figure S12.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for external gastric cancer and healthy controls. **Figure S13.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for external hepatocellular carcinoma and healthy controls. **Figure S14.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for external esophageal carcinoma and healthy controls. **Figure S15.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for external mixture cancer and healthy controls. **Figure S16.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for common lung disease compared to the stage I/II of lung cancer. **Figure S17.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for common colorectal disease compared to the stage I/II of colorectal carcinoma. **Figure S18.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for common gastric disease compared to the stage I/II of gastric cancer. **Figure S19.** Heatmap of the covariance matrix formed on 60 true dimensions between 600-648.3607 cm^{-1} and 30 true dimensions 600-623.7705 cm^{-1} using peak data for common hepatopathy compared to the stage I/II of hepatoma. **Figure S20.** Establishment of internal cancers-related database. **Figure S21.** Establishment of external cancers-related database. **Figure S22.** Establishment of early stage of cancers related database. **Table S1.** Summary of cancer patient and healthy control clinical characteristic. **Table S2.** List of characteristic peak locations obtained by covariance identification correlation analysis for internal lung cancer compared to the healthy control group. **Table S3.** List of characteristic peak locations obtained by covariance identification correlation analysis for internal colorectal carcinoma compared to the healthy control group. **Table S4.** List of characteristic peak locations obtained by covariance identification correlation analysis for internal gastric cancer compared to the healthy control group. **Table S5.** List of characteristic peak locations obtained by covariance identification correlation analysis for internal hepatocellular carcinoma compared to the healthy control group. **Table S6.** List of characteristic peak locations obtained by covariance identification correlation analysis for internal esophageal carcinoma compared to the healthy control group. **Table S7.** List of characteristic peak locations obtained by covariance identification correlation analysis for internal mixture cancer compared to the healthy control group. **Table S8.** List of characteristic peak locations obtained by covariance identification correlation analysis for external lung cancer compared

to the healthy control group. **Table S9.** List of characteristic peak locations obtained by covariance identification correlation analysis for external colorectal carcinoma compared to the healthy control group. **Table S10.** List of characteristic peak locations obtained by covariance identification correlation analysis for external gastric cancer compared to the healthy control group. **Table S11.** List of characteristic peak locations obtained by covariance identification correlation analysis for external hepatocellular carcinoma compared to the healthy control group. **Table S12.** List of characteristic peak locations obtained by covariance identification correlation analysis for external esophageal carcinoma compared to the healthy control group. **Table S13.** List of characteristic peak locations obtained by covariance identification correlation analysis for external mixture cancer compared to the healthy control group. **Table S14.** List of characteristic peak locations obtained by covariance identification correlation analysis for common lung disease compared to the stage I/II of lung cancer. **Table S15.** List of characteristic peak locations obtained by covariance identification correlation analysis for common colorectal disease compared to the stage I/II of colorectal carcinoma. **Table S16.** List of characteristic peak locations obtained by covariance identification correlation analysis for common gastric disease compared to the stage I/II of gastric cancer. **Table S17.** List of characteristic peak locations obtained by covariance identification correlation analysis for common hepatopathy compared to the stage I/II of hepatoma.

Acknowledgements
Not applicable.

Author contributions
SD and XX designed and planned the study, and developed experimental protocols. SD, DH, QZ, FW, CH, CZ, BT, WW, JH and XX optimized the experimental protocols. QZ, LN, KW, WL, JY and FW collect the information of sample and clinical data. ZH, XC and CH analyzed and interpreted data. SD, JH and XX wrote the manuscript and incorporated feedback from all authors. SD, DH, and QZ contributed equally to this study.

Funding
This work was supported by the National Natural Science Foundation of China (12025503, 12102086), Science Fund for Creative Research Groups of the Natural Science Foundation of Hubei Province (No. 2022CFA005), Experimental Technology project of Wuhan University (WHU-2021-SYJS-06), Sichuan Science and Technology Program (2021YJ0182). This work was also supported by the Fundamental Research Funds for the Central Universities (No. 2042021kf0227, 2042022kf1181) and medical Sci-Tech innovation platform of Zhongnan Hospital (PTXM2021001).

Availability of data and materials
All data needed to evaluate the conclusions in the study are present in the main text or the supplementary materials. Source data can be found at the Wuhan University Repository.

Declarations

Ethics approval and consent to participate
The use of samples in this study were performed with waiver statement and approved by the Medical Ethics Committee of Zhongnan Hospital of Wuhan University, Hubei Cancer Hospital, Wuhan Hospital of Traditional Chinese and Western Medicine and Tianjin Medical University General Hospital under the approval number No. 2020104, No. LLHBCH2022YN-038, No. 2022-47 and No. IRB2022-WZ-122.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Physics, National Demonstration Center for Experimental Physics Education, Wuhan University, Wuhan 430072, China. [2]Department of Laboratory Medicine, Zhongnan Hospital of Wuhan University, Wuhan 430071, China. [3]School of Computer Science, Wuhan University, Wuhan 430072, China. [4]Department of Hepatobiliary Pancreatic Surgery, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430079, China. [5]Department of Clinical Laboratory, Tianjin Medical University General Hospital, Tianjin 300052, China. [6]Department of Blood Transfusion, Wuhan Hospital of Traditional Chinese and Western Medicine, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China. [7]Laboratory of Printable Functional Materials and Printed Electronics, Research Center for Graphic Communication, Printing and Packaging, Wuhan University, Wuhan 430072, China. [8]Sichuan Provincial Key Laboratory for Human Disease Gene Study and the Center for Medical Genetics, Department of Laboratory Medicine, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, University of Electronic Science and Technology, Chengdu 611731, China. [9]School of Medicine, University of Electronic Science and Technology of China, Chengdu 611731, China. [10]Wuhan Research Center for Infectious Diseases and Cancer, Chinese Academy of Medical Sciences, Wuhan 430071, China.

Received: 15 March 2023 Revised: 28 May 2023 Accepted: 6 June 2023
Published online: 24 July 2023

References

1. H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **71**, 209–249 (2021)
2. J.C. Wan, C. Massie, J. Garcia-Corbacho, F. Mouliere, J.D. Brenton, C. Caldas, S. Pacey, R. Baird, N. Rosenfeld, Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat. Rev. Cancer **17**, 223–238 (2017)
3. J. Phallen, M. Sausen, V. Adleff, A. Leal, C. Hruban, J. White, V. Anagnostou, J. Fiksel, S. Cristiano, E. Papp, S. Speir, T. Reinert, M.B.W. Orntoft, B.D. Woodward, D. Murphy, S. Parpart-Li, D. Riley, M. Nesselbush, N. Sengamalay, A. Georgiadis, Q.K. Li, M.R. Madsen, F.V. Mortensen, J. Huiskens, C. Punt, N. van Grieken, R. Fijneman, G. Meijer, H. Husain, R.B. Scharpf, L.A. Diaz Jr., S. Jones, S. Angiuoli, T. Ørntoft, H.J. Nielsen, C.L. Andersen, V.E. Velculescu, Direct detection of early-stage cancers using circulating tumor DNA. Sci. Transl. Med. **9**, eaan2415 (2017)
4. A.M. Newman, S.V. Bratman, J. To, J.F. Wynne, N.C. Eclov, L.A. Modlin, C.L. Liu, J.W. Neal, H.A. Wakelee, R.E. Merritt, J.B. Shrager, B.W. Loo Jr., M. Diehn, An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat. Med. **20**, 548–554 (2014)
5. D. Crosby, S. Bhatia, K.M. Brindle, L.M. Coussens, C. Dive, M. Emberton, S. Esener, R.C. Fitzgerald, S.S. Gambhir, P. Kuhn, T.R. Rebbeck, S. Balasubramanian, Early detection of cancer. Science **375**, 1244 (2022)
6. P.D. Howes, R. Chandrawati, M.M. Stevens, Colloidal nanoparticles as advanced biological sensors. Science **346**, 1247390 (2014)
7. L. Guerrini, E. Garcia-Rico, A. O'Loghlen, V. Giannini, R.A. Alvarez-Puebla, Surface-enhanced Raman scattering (SERS) spectroscopy for sensing and characterization of exosomes in cancer diagnosis. Cancers **13**, 2179 (2021)
8. J. Kondo, T. Ekawa, H. Endo, K. Yamazaki, N. Tanaka, Y. Kukita, H. Okuyama, J. Okami, F. Imamura, M. Ohue, K. Kato, T. Nomura, A. Kohara, S. Mori, S. Dan, M. Inoue, High-throughput screening in colorectal cancer tissue-originated spheroids. Cancer Sci. **110**, 345–355 (2019)
9. S. Abalde-Cela, R. Rebelo, L. Wu, A.I. Barbosa, L. Rodríguez-Lorenzo, K. Kant, R.L. Reis, V.M. Correlo, L. Diéguez, A SERS-based 3D nanobiosensor: towards cell metabolite monitoring. Mater. Adv. **1**, 1613–1621 (2020)
10. J. Ko, J. Ham, H. Lee, K. Lee, W.G. Koh, Integration of a fiber-based cell culture and biosensing system for monitoring of multiple protein markers secreted from stem cells. Biosens. Bioelectron. **193**, 113531 (2021)

11. W. Nam, X. Ren, S.A.S. Tali, P. Ghassemi, I. Kim, M. Agah, W. Zhou, Refractive-index-insensitive nanolaminated SERS substrates for label-free Raman profiling and classification of living cancer cells. Nano Lett. **19**, 7273–7281 (2019)
12. W. Nam, H. Chen, X. Ren, M. Agah, I. Kim, W. Zhou, Nanolaminate plasmonic substrates for high-throughput living cell SERS measurements and artificial neural network classification of cellular drug responses. ACS Appl. Nano Mater. **5**, 10358–10368 (2022)
13. L. Zhang, C. Li, D. Peng, X. Yi, S. He, F. Liu, X. Zheng, W.E. Huang, L. Zhao, X. Huang, Raman spectroscopy and machine learning for the classification of breast cancers. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. **264**, 120300 (2022)
14. Z. Wang, J. Ye, K. Zhang, L. Ding, T. Granzier-Nakajima, J.C. Ranasinghe, Y. Xue, S. Sharma, I. Biase, M. Terrones, S.H. Choi, C. Ran, R.E. Tanzi, S.X. Huang, C. Zhang, S. Huang, Rapid biomarker screening of Alzheimer's disease by interpretable machine learning and graphene-assisted Raman spectroscopy. ACS Nano **16**, 6426–6436 (2022)
15. K. Liu, B. Liu, Y. Zhang, Q. Wu, M. Zhong, L. Shang, Y. Wang, P. Liang, W. Wang, Q. Zhao, B. Li, Building an ensemble learning model for gastric cancer cell line classification via rapid raman spectroscopy. Comput. Struct. Biotechnol. J. **21**, 802–811 (2023)
16. J. Ye, Y.-T. Yeh, Y. Xue, Z. Wang, N. Zhang, H. Liu, K. Zhang, R. Ricker, Z. Yu, A. Roder, N. Perea Lopez, L. Organtini, W. Greene, S. Hafenstein, H. Lu, E. Ghedin, M. Terrones, S. Huang, S.X. Huang, Accurate virus identification with interpretable Raman signatures by machine learning. Proc. Natl. Acad. Sci. U. S. A. **119**, e2118836119 (2022)
17. F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **68**, 394–424 (2018)
18. J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D.M. Parkin, M. Piñeros, M. Piñeros, F. Bray, Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int. J. Cancer **144**, 1941–1953 (2019)
19. C. Cao, D. Wang, C. Chung, D. Tian, A. Rimner, J. Huang, D.R. Jones, A systematic review and meta-analysis of stereotactic body radiation therapy versus surgery for patients with non-small cell lung cancer. J. Thorac. Cardiovasc. Surg. **157**, 362-373.e368 (2019)
20. G. Wu, Y. Liu, B. He, Q. Bao, A. Duan, F.F. Jin, Thermal controls on the Asian summer monsoon. Sci. Rep. **2**, 1–7 (2012)
21. K. Czamara, K. Majzner, M.Z. Pacia, K. Kochan, A. Kaczor, M. Baranska, Raman spectroscopy of lipids: a review. J. Raman Spectrosc. **46**, 4–20 (2015)
22. C. Zheng, S. Qing, J. Wang, G. Lu, H. Li, X. Lu, C. Ma, J. Tang, X. Yue, Diagnosis of cervical squamous cell carcinoma and cervical adenocarcinoma based on Raman spectroscopy and support vector machine. Photodiagnosis Photodyn. Ther. **27**, 156–161 (2019)
23. L. Habartova, B. Bunganic, M. Tatarkovic, M. Zavoral, J. Vondrousova, K. Syslova, V. Setnicka, Chiroptical spectroscopy and metabolomics for blood-based sensing of pancreatic cancer. Chirality **30**, 581–591 (2018)
24. S.J. Lee, A.C. Noble, Characterization of odor-active compounds in Californian Chardonnay wines using GC-olfactometry and GC-mass spectrometry. J. Agric. Food Chem. **51**, 8036–8044 (2003)
25. C. Matrix, M.R. Reynolds Jr., G.Y. Cho, Multivariate control charts for monitoring the mean vector and covariance matrix. J. Qual. Technol. **38**, 230–253 (2006)
26. W.S. Noble, What is a support vector machine? Nat. biotechnol. **24**, 1565–1567 (2006)
27. M.M. Eid, Y.H. Elawady, Efficient pneumonia detection for chest radiography using ResNet-based SVM. J. Electr. Comput. Eng. **5**, 1–8 (2021)
28. J.C. Fu, S.K. Lee, S.T.C. Wong, J.Y. Yeh, A.H. Wang, H.K. Wu, Image segmentation feature selection and pattern classification for mammographic microcalcifications. Comput Med Imaging Graph **29**, 419–429 (2005)
29. J.P. Kandhasamy, S.J.P.C.S. Balamurali, Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci **47**, 45–51 (2015)
30. S.K. Kamarajah, W.R. Burns, T.L. Frankel, C.S. Cho, H. Nathan, Validation of the American Joint Commission on Cancer (AJCC) staging system for patients with pancreatic adenocarcinoma: a Surveillance, Epidemiology and End Results (SEER) analysis. Ann. Surg. Oncol. **24**, 2023–2030 (2017)
31. J.A. Hanley, Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging **29**, 307–335 (1989)
32. M.H. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. **39**, 561–577 (1993)
33. J.C. Pruessner, C. Kirschbaum, G. Meinlschmid, D.H. Hellhammer, Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change. Psychoneuroendocrinology **28**, 916–931 (2003)
34. W.G. Hopkins, Measures of reliability in sports medicine and science. Sports Med **30**, 1–15 (2000)
35. N.F. Fernandez, G.W. Gundersen, A. Rahman, M.L. Grimes, K. Rikova, P. Hornbeck, A. Ma'ayanClustergrammer, A web-based heatmap visualization and analysis tool for high-dimensional biological data. Sci. Data **4**, 1–12 (2017)
36. N. Kim, H. Park, N. He, H.Y. Lee, S. Yoon, QCanvas: an advanced tool for data clustering and visualization of genomics data. Genomics Inform. **10**, 263–265 (2012)
37. B.B. Khomtchouk, J.R. Hennessy, C. Wahlestedt, Shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics. PLoS ONE **12**, e0176334 (2017)
38. N. Gehlenborg, J. Dietzsch, K. Nieselt, A framework for visualization of microarray data and integrated meta information. Inf Vis **4**, 164–175 (2005)
39. S. Zong, L. Wang, C. Chen, J. Lu, D. Zhu, Y. Zhang, Z. Wang, Y. Cui, Facile detection of tumor-derived exosomes using magnetic nanobeads and SERS nanoprobes. Anal. Methods. **8**, 5001–5008 (2016)
40. Z. Wang, S. Zong, Y. Wang, N. Li, L. Li, J. Lu, Z. Wang, B. Chen, Y. Cui, Screening and multiple detection of cancer exosomes using an SERS-based method. Nanoscale **10**, 9053–9062 (2018)
41. E.A. Kwizera, R. O'Connor, V. Vinduska, M. Williams, E.R. Butch, S.E. Snyder, X. Chen, X. Huang, Molecular detection and analysis of exosomes using surface-enhanced Raman scattering gold nanorods and a miniaturized device. Theranostics **8**, 2722–2738 (2018)
42. T.D. Li, R. Zhang, H. Chen, Z.P. Huang, X. Ye, H. Wang, A.M. Deng, J.L. Kong, An ultrasensitive polydopamine bi-functionalized SERS immunoassay for exosome-based diagnosis and classification of pancreatic cancer. Chem. Sci. **9**, 5372–5382 (2018)