Noninvasive prostate cancer screening based on serum surface-enhanced Raman spectroscopy and support vector machine
Shaoxin Li, Yanjiao Zhang, Junfa Xu, Linfang Li, Qiuyao Zeng, Lin Lin, Zhouyi Guo, Zhiming Liu, Honglian Xiong , and Songhao Liu

Citation: Applied Physics Letters 105, 091104 (2014); doi: 10.1063/1.4892667
View online: http://dx.doi.org/10.1063/1.4892667
View Table of Contents: http://scitation.aip.org/content/aip/journal/apl/105/9?ver=pdfcov
Published by the AIP Publishing

Articles you may be interested in
Saliva analysis combining membrane protein purification with surface-enhanced Raman spectroscopy for nasopharyngeal cancer detection

Minimum classification error-based weighted support vector machine kernels for speaker verification

Esophageal cancer detection based on tissue surface-enhanced Raman spectroscopy and multivariate analysis
Appl. Phys. Lett. 102, 043702 (2013); 10.1063/1.4789996

Discrimination of gastric cancer from normal by serum RNA based on surface-enhanced Raman spectroscopy (SERS) and multivariate analysis
Med. Phys. 39, 5664 (2012); 10.1118/1.4747269

A GMM-based breast cancer risk stratification using a resonance-frequency electrical impedance spectroscopy
Med. Phys. 38, 1649 (2011); 10.1118/1.3555300
Noninvasive prostate cancer screening based on serum surface-enhanced Raman spectroscopy and support vector machine

Shaoxin Li,1,2,a) Yanjiao Zhang,3,a) Junfa Xu,2 Linfang Li,4,b) Qiuyao Zeng,4 Lin Lin,1 Zhouyi Guo,5 Zhiming Liu,5 Honglian Xiong,5 and Songhao Liu5

1Biomedical Engineering Laboratory, Guangdong Medical College, Dongguan 523808, China
2Guangdong Provincial Key Laboratory of Medical Molecular Diagnostics, No. 1 Xincheng Road, Dongguan 523808, China
3School of Basic Medicine, Guangdong Medical College, Dongguan 523808, China
4State Key Laboratory of Oncology in South China and Department of Clinical Laboratory, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
5MOE Key Laboratory of Laser Life Science & SATCM Third Grade Laboratory of Chinese Medicine and Photonics Technology, College of Biophotonics, South China Normal University, Guangzhou 510631, China

(Received 25 June 2014; accepted 29 July 2014; published online 2 September 2014)

This study aims to present a noninvasive prostate cancer screening methods using serum surface-enhanced Raman scattering (SERS) and support vector machine (SVM) techniques through peripheral blood sample. SERS measurements are performed using serum samples from 93 prostate cancer patients and 68 healthy volunteers by silver nanoparticles. Three types of kernel functions including linear, polynomial, and Gaussian radial basis function (RBF) are employed to build SVM diagnostic models for classifying measured SERS spectra. For comparably evaluating the performance of SVM classification models, the standard multivariate statistic analysis method of principal component analysis (PCA) is also applied to classify the same datasets. The study results show that for the RBF kernel SVM diagnostic model, the diagnostic accuracy of 98.1% is acquired, which is superior to the results of 91.3% obtained from PCA methods. The receiver operating characteristic curve of diagnostic models further confirm above research results. This study demonstrates that label-free serum SERS analysis technique combined with SVM diagnostic algorithm has great potential for noninvasive prostate cancer screening. © 2014 AIP Publishing LLC.

[http://dx.doi.org/10.1063/1.4892667]

Prostate cancer is one of the most common malignancies of older male worldwide with the sixth leading cause of cancer-related death.1 Clinically there are several detecting means include B-mode ultrasound, biopsy, and tumor markers screening, but these techniques have some disadvantages. For example, B-mode ultrasound can only image the formed solid tumor, which suggests that patients are not in early stage of cancer. Biopsy is the gold standard of cancer examination, but it is invasive and impractical for a high-risk patient with multiple suspicious lesions. Tumor markers screening such as prostate specific antigen (PSA) detection has significantly improved the early diagnosis, but the PSA test also has some limitations in sensitivity and specificity.

In recent years, Raman spectroscopy has drawn considerable attention due to its great potential for improving clinical diagnosis.2–6 However, the weak Raman signal of biological samples hinders the clinical applications of this technology. Fortunately, with the discovery of surface enhanced Raman scattering (SERS), Raman spectroscopy technique acquired considerable development.7 SERS has been applied to probe pesticide residues, protein, nucleic acid, DNA, and other biomacromolecules.7,8 The most recent reports show that SERS has been used for target detection of tumor markers in the blood or cell surface by immunoassay approaches.9

Blood samples are ideal disease screening materials for non-invasive cancer diagnosis, which can be executed conveniently and even repeatedly for high risk patients. At the early stages of cancer, the content of biomolecules such as protein, fat, and DNA contained in human blood will undergo subtle alterations which can be revealed by Raman spectroscopy. However, the differences of Raman spectroscopy between normal and pathologic tissue samples are usually tiny that it is difficult to distinguish with direct methods. The robust and powerful spectral data processing technique are much needed to extract effectively diagnostic information. Principal component analysis (PCA), linear discriminant analysis (LDA), and other multivariate statistical techniques have been successfully applied to the diagnosis and classification of various tissues Raman spectroscopy.10 These methods improve the precision and reliability of Raman spectroscopy analysis; however, there is still some space to improve diagnostic efficiency.

Support vector machine (SVM), a relatively young multivariate technique, is considered to be superior over traditional linear approaches due to its capability of processing binary classification problem with nonlinear boundary by mapping sample dataset into a higher dimensional space.11

The fundamental idea involves that SVM algorithm separate classes with optimal hyperplane, which maximizes the margin of separation between the hyperplane and the closest
data points on both sides of the hyperplane. SVM has been successfully applied in the field of face recognition, text categorization, gene selection, and so on.\textsuperscript{12,13} For the datasets that are not separable in the input space, SVM techniques map the datasets into a higher dimensional space by kernel functions. The three frequently used kernel functions are

\[
\text{Linear: } K(x_i, x_j) = x_i \cdot x_j + 1, \quad (1)
\]

\[
\text{Polynomial kernel: } K(x_i, x_j) = (x_i \cdot x_j + 1)^d, \quad (2)
\]

\[
\text{Gaussian radial basis function (RBF): } K(x_i, x_j) = \exp \left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \quad (3)
\]

where \(x_i\) and \(x_j\) are the two generic sample data vectors.

Once the data are mapped into the feature space an infinite number of separating hyperplanes may exist, creating the risk of overfitting the hyperplanes to the given data points. To overcome this problem, a penalty factor \(C\) is introduced to allow some training data to be misclassified, the higher \(C\) the lower misclassification rate.

To develop a simple, noninvasive method for prostate cancer diagnosis, this study employs Ag-based SERS technique combined with SVM to classify serum SERS spectra between prostate cancer patients and normal subjects. Three kernel functions including linear, polynomial, and Gaussian radial basis function (RBF) are used to build diagnostic models. Each diagnostic model is evaluated with leave-one-sample-out cross-validation method, which involves using one sample held out from dataset as the validation data, and the remain spectra as the training data. This process is repeated such that each sample is used once as the validation data. To compare the performance of SVM algorithms, PCA-LDA techniques are employed to classify the same Raman datasets. The LIBSVM toolbox 3.1 created by Chang and Lin is used for SVM classifications. All procedure is implemented with MATLAB language.

Ag nano-particles (AgNPs) were synthesized with silver nitrate reduction method.\textsuperscript{14} The obtained AgNPs have mean diameter about 50 nm. Serum samples were collected from 68 healthy volunteers and 93 prostate cancer patients, who were confirmed clinically with histopathology (45 cases mid-cancers and 48 cases advanced cancers). After 12 h of overnight fasting, a single 3 ml peripheral blood samples were collected from the study subjects between 7:00 and 8:00 A.M. After blood samples were centrifuged at 3500 rpm for 5 min, serum was obtained by removing supernatant. All patients came from Sun Yat-sen university cancer center and the ethical approval was signed to permit collection of blood sample prior to research.

Before SERS measurement, 20 \(\mu\)l silver colloidal nanoparticles were homogenously mixed with 20 \(\mu\)l serum. After incubating for 1 h at room temperature, a drop of resulting mixture was transferred onto an aluminum plate and naturally dried for 1 h for SERS measurement. A Renishaw Raman microscopy (inVia, UK) was used for the collection of SERS spectra with a spectral resolution about 1 cm\(^{-1}\). The 785 nm diode laser was focused on the sample surface with excitation power about 0.5 MW. The spectral data acquisition time was 10 s by a Leica DM2500 microscope equipped with objective 20\(\times\).

We have measured the regular Raman spectra and SERS spectra of serum sample come from the same subject in order to assess the silver colloid enhancement effects on the Raman scattering of serum samples. Fig. 1(a) displays the serum SERS spectra from a prostate patient, and Fig. 1(b) is the regular Raman spectra of serum sample from the same patient without silver colloid. It is clearly shown that there is a considerable enhanced for the serum SERS spectra with Ag colloid, while there is no Raman peak observed for the regular Raman spectra of serum sample without silver colloid. The dramatically increase in many dominant vibration bands suggests strong interactions between AgNPs and serum. Fig. 1(c) is the Raman spectroscopy of silver colloid without serum sample. No interference signal is observed in the interested spectral range.

A total of 161 serum SERS spectra were obtained, in which 93 SERS spectra were from prostate patients and 68 from normal subjects. The measured spectra were processed by smoothing, baseline correction, and area normalized under the curve.\textsuperscript{15} Fig. 2 displays normalized mean SERS spectra \(\pm 1\) standard deviations of prostate cancer and normal serum in the range of 400–1800 cm\(^{-1}\). The solid lines indicate mean spectra and the shaded lines represent one standard deviation. It is clearly displayed from Figs. 2(a) and 2(b) that the cancer and normal serum SERS spectra are very similar in the profile. Therefore, a powerful data analysis algorithm is needed to acquire the effective information for differentiating the normal and cancer serum SERS spectra. Fig. 2(c) is the difference spectra between prostate cancer and normal serum. The distinct differences imply that there is enormous potential to diagnosis prostate cancer using serum SERS spectra. Primary Raman bands are observed in normal and cancer serum SERS spectra at the following peak positions with tentative biochemical assignments:\textsuperscript{10,16,17} 481 cm\(^{-1}\) (glycogen), 516 cm\(^{-1}\) (S–S disulfide stretch in proteins), 650 cm\(^{-1}\) (C–C twist, tyrosine), 685 cm\(^{-1}\) (C–S twist), 728 cm\(^{-1}\) (C–H bending adenine, coenzyme), 830 cm\(^{-1}\) (Out-of-plane ring breathing, tyrosine), 859 cm\(^{-1}\) (C–C stretch of proline ring, ring breathing of tyrosine), 911 cm\(^{-1}\) (C–C stretch of proline ring, glucose), 1025 cm\(^{-1}\) (C–H stretch of phenylalanine), 1134 cm\(^{-1}\) (C–N stretch, D-mannos), 1217 cm\(^{-1}\) (C–C\(_2\)H\(_5\) phenylalanine, tryptophan), 1316 cm\(^{-1}\)
This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 113.118.136.90 On: Mon, 15 Sep 2014 00:55:10

(\text{CH}_3\text{CH}_2\text{ twisting collagen/lipids}), 1347\text{ cm}^{-1} (\text{CH}_3\text{CH}_2\text{ wagging, tryptophan, adenine, guanine}), 1445\text{ cm}^{-1} (\text{CH}_2\text{ bending, collagen/lipids}), \text{ and } 1586\text{ cm}^{-1} (\text{C=C bending, phenylalanine, acetoacetate, riboflavin}). \text{ The difference in vigorous metabolism consuming a lot of fat.}^{18} \text{ The Raman components. The likely reason may attribute to the tumor's gen/lipids) in cancer serum suggests a reduction in the per-

A grid search is performed by trying various pairs of parameter C and polynomial order d. The range of C is set from \(2^{-10}\) to \(2^{10}\) and polynomial order d from 1 to 10. Fig. 4 is the 3D map of diagnostic accuracy as a function of parameter C and polynomial order d. The maximum diagnostic accuracy of 98.2% is obtained at the order of 2 with the parameter C from \(2^{-10}\) to \(2^{10}\). With the \(d = 2\) and \(C = 1\), the diagnostic accuracy of 97.5%, sensitivity of 95.7%, and specificity of 100% are obtained in Table I.

For the RBF kernel SVM algorithms, the penalty parameter C and Gaussian width \(\sigma\) are optimized with grid search. The range of C is set from \(2^{-20}\) to \(2^{10}\) and Gaussian width \(\sigma\) from \(2^5\) to \(2^{15}\), increasing in steps of the power of two. Fig. 5 shows the 3D map of diagnostic accuracy as a function of parameter C and Gaussian radial width \(\sigma\). It is clearly exhibits that the diagnostic accuracy changes with the parameter C and Gaussian radial width \(\sigma\). The largest diagnostic accuracy of 98.8% achieved at \(C = 2^{-10}\) and \(\sigma = 2^{10}\). In fact, there are several pairs of parameter C and \(\sigma\) that can give the largest diagnostic accuracy of 98.8%. Based on the optimal values of \(C = 2^{-10}\) and \(\sigma = 2^{10}\), the cross validation results of 98.1% for the diagnostic sensitivity and 100% for the specificity are achieved in Table I for differentiation serum SERS spectra between normal volunteers and prostate cancer patients.

In order to comparably assess the performance of SVM algorithms, PCA-LDA techniques are employed to analyze the same SERS dataset to distinguish prostate cancer from normal subjects. PCA is a standard tool in multivariate data analysis to reduce the number of dimensions, while retaining as much as possible of the data’s variation. In this study, PCA is used to process the serum SERS spectra which consist of 1270 variables in the range of 400–1800 cm\(^{-1}\).
the combination of LDA, the first 20 PCs that account for 94.4% of the total variance are employed to classify the SERS spectra. The classification accuracy of 91.3%, sensitivity of 84.9% and specificity of 100% are obtained with the leave-one sample-out cross-validation methods. The results are lower than these of SVM algorithms.

In order to further confirm the performance of diagnostic models developed by SVM algorithms, the receiver operating characteristic curve (ROC) of RBF kernel SVM and PCA-LDA algorithms are generated in Fig. 6. The ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The integration area under the ROC (AUC) is a quantitative indicator to represent classifier performances. The larger AUC value means the greater forecast accuracy for the classifier. In this study, the AUC of RBF kernel SVM and PCA-LDA classifier models are 0.998 and 0.991, respectively. These results further confirm that RBF kernel SVM algorithms yield a better diagnostic accuracy than the PCA-LDA algorithms. Though the ROC of SVM is not strictly ROC but we think it can demonstrate the performance of SVM on some extent.

In this study, we have measured serum SERS spectra of prostate cancer patients and normal volunteers with silver colloid. The huge increase in many dominant vibration bands and the distinct differences between normal and cancerous serum SERS spectra imply that there is enormous potential to label-free diagnose prostate cancer using serum SERS spectra through a peripheral blood sample. Three SVM classifier models including linear, polynomial, and RBF kernel functions are built and evaluated with measured serum SERS spectra. The diagnostic accuracy of three SVM diagnostic models are higher than that of PCA-LDA algorithms, and the maximum diagnostic accuracy of 98.8% is acquired with RBF kernel SVM classifier model. These results prove that the diagnostic performance of SVM is superior to that of PCA algorithms. The ROC of SVM and PCA algorithms further confirms the difference of their diagnostic performance. The cause maybe attributed to two sides: one hand, the PCA technique lost a part of effective diagnostic information during processing the SERS spectra; on the other hand, between normal and prostate cancer serum SERS spectra, there is nonlinear boundary that cannot be distinguished with linear algorithms such as PCA-LDA. However, there are some flaws for this study. For instance, the sample size is not enough, and the time variability of sample is absent. These factors need to be further studied in our future work. In conclusion, this preliminary research demonstrates that serum SERS techniques associating with SVM algorithms have great potential to non-invasive screen prostate cancer through a peripheral blood sample.

The authors would like to acknowledge the financial support of the National Natural Science Foundation of China (61275187 and 60778047), Specialized Research Fund for the Doctoral Program of Higher Education of China (20114407110001), the Natural Science Foundation of Guangdong Province (9251063101000009) and the Cooperation Project in Industry, Education, and Research of Guangdong province and Ministry of Education of China (2011A090200011).

![FIG. 5. 3D map of diagnostic accuracy as a function of parameter C and Gaussian radial width $\sigma$ using the RBF kernel SVM algorithm.](image)

![FIG. 6. ROC of RBF kernel SVM and PCA-LDA algorithms. The integration area under the ROC for PCA-LDA and SVM are 0.991 and 0.998, respectively.](image)