Journal of Biomedical Optics

SPIEDigitalLibrary.org/jbo

Detection of nasopharyngeal cancer using confocal Raman spectroscopy and genetic algorithm technique

Shao-Xin Li
Qiu-Yan Chen
Yan-Jiao Zhang
Zhi-Ming Liu
Hong-Lian Xiong
Zhou-Yi Guo
Hai-Qiang Mai
Song-Hao Liu

SPIE

Detection of nasopharyngeal cancer using confocal Raman spectroscopy and genetic algorithm technique

Shao-Xin Li,[a,b] Qiu-Yan Chen,[c] Yan-Jiao Zhang,[b] Zhi-Ming Liu,[a] Hong-Lian Xiong,[a] Zhou-Yi Guo,[a] Hai-Qiang Mai,[c] and Song-Hao Liu[a]
[a]South China Normal University, School of Information and Optoelectronic Science and Engineering, Guangzhou 510631, China
[b]School of Information Engineering, Guangdong Medical College, Dongguan 523808, China
[c]Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China and Department of Nasopharyngeal Carcinoma, Guangzhou 510060, China

Abstract. Raman spectroscopy (RS) and a genetic algorithm (GA) were applied to distinguish nasopharyngeal cancer (NPC) from normal nasopharyngeal tissue. A total of 225 Raman spectra are acquired from 120 tissue sites of 63 nasopharyngeal patients, 56 Raman spectra from normal tissue and 169 Raman spectra from NPC tissue. The GA integrated with linear discriminant analysis (LDA) is developed to differentiate NPC and normal tissue according to spectral variables in the selected regions of 792–805, 867–880, 996–1009, 1086–1099, 1288–1304, 1663–1670, and 1742–1752 cm^{-1} related to proteins, nucleic acids and lipids of tissue. The GA-LDA algorithms with the leave-one-out cross-validation method provide a sensitivity of 69.2% and specificity of 100%. The results are better than that of principal component analysis which is applied to the same Raman dataset of nasopharyngeal tissue with a sensitivity of 63.3% and specificity of 94.6%. This demonstrates that Raman spectroscopy associated with GA-LDA diagnostic algorithm has enormous potential to detect and diagnose nasopharyngeal cancer. [DOI: 10.1117/1.JBO.17.12.125003]

Keywords: nasopharyngeal carcinoma; Raman spectroscopy; genetic algorithm.

Paper 12493 received Aug. 2, 2012; revised manuscript received Oct. 9, 2012; accepted for publication Nov. 1, 2012; published online Dec. 3, 2012.

1 Introduction

Nasopharyngeal cancer (NPC) is a nonlymphomatous squamous cell carcinoma that occurs in the epithelial lining of the nasopharynx.[1] The incidence of NPC occurs with much greater frequency in southern China, northern Africa, and Alaska. It is reported that the disease is the third most common malignancy among men, with an incidence about 50 per 100,000 in the Guangdong Province of China.[2] At present, the clinical diagnosis of NPC mainly depends on white-light endoscope technique and excisional biopsy. Endoscope technique relies on the observation of gross morphological changes of tissues, thus, it is difficult to identify the NPC in its early stages. Excisional biopsy remains the gold standard approach for cancer diagnosis, but it involves a high level of operational requirements for doctors and causes a lot of discomfort or may even be impractical for high-risk patients.[3] Therefore, it is eminently desirable to develop a nondestructive technique to detect NPC at an early stage.

In recent years, the Raman spectroscopy technique has received a great deal of interest in cancer diagnosis.[4,5] Raman spectroscopy is a vibrational spectroscopic technique that can provide specific spectroscopic fingerprint information about the molecular composition, structure and content of constituents.[6] Compared with other optical spectroscopic techniques, such as the fluorescence spectroscopy and infrared absorption spectroscopy, Raman spectroscopy has significant advantages. For instance, there is no photobleaching in Raman scattering, and Raman spectral peaks are narrow.[7,8] Especially, when the near-infrared excitation light is used, water exhibits very low absorption and tissues reveal far less autofluorescence compared with visible light excitation. This makes it easier to detect biochemical components within a deeper layer of the tissue. Over the past decade, Raman spectroscopy has been applied to classify normal and malignant tissues of various body sites, including the breast, bladder, lung, prostate, cervix, skin, and so on.[4,9–15] These studies show that Raman spectral features could be used to correlate with the molecular and structural changes associated with carcinomatous transformations, demonstrating the feasibility of early cancer detection by Raman spectroscopy.

The differences of Raman spectra among different tissue types are usually subtle because of obvious spectral overlapping. Thus, powerful and robust spectral data processing and diagnostic algorithms are required to extract significant Raman spectral features associated with the histopathology. In recent years, multivariate statistical algorithms such as principal component analysis (PCA), linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLS-DA) have been widely applied for the classification of various tissues by Raman spectroscopy.[4,16] These statistical methods improve the precision and reliability of the procedure.[17]

PCA is a mathematical tool that reduces the dimensions of the dataset by orthogonally projecting data onto a lower dimensional linear space, such that the variance of the projected data is maximized. Combined with LDA and support vector machines, it has been applied to distinguish the biomedical Raman spectra.[18,19] PCA is efficient for data classification, but PCs have no

Address all correspondence to: Zhou-Yi Guo, South China Normal University, School of Information and Optoelectronic Science and Engineering, Guangzhou 510631, China. Tel: 86-020-85211428; Fax: 86-020-85216052; E-mail: ann@scnu.edu.cn; or Hai-Qiang Mai, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China and Department of Nasopharyngeal Carcinoma, Guangzhou 510060, China. Email : maihq@mail.sysu.edu.cn

0091-3286/2012/$25.00 © 2012 SPIE

definite physical meanings because they are extracted from a linear combination of original variables. Furthermore, PCA is generally calculated using the entire spectrum that contains much redundant data and noise, which have no contribution to the principal components and reduce the performance of the PCA.[20]

Feature selection is another data reduction technique that mines feature subsets from the original data space. The classification attributes of the subset are, as much as possible, to remain consistent with the raw data. The main advantages of feature selection are to decrease the number of variables to help understand the discovered pattern by eliminating irrelevant features from raw dataset, and to increase the accuracy of classification. To improve the ability of finding a meaningful low-dimensional data structure in high-dimensional sample space, several feature selection strategies such as filter, embedded and wrapper have been developed according to different evaluation criteria.[21,22] The genetic algorithm (GA), based on Darwinian evolution and Mendelian genetics, has emerged as an adaptive heuristic search algorithm for efficient features selection.[23] Compared with other search methods, GA is robust, parallel, and has global search superiority. GA has displayed excellent ability in feature selection in microarray analysis, mass spectrometry, sequence analysis, and so on.[24–27] This study introduces GA combined with LDA method to select characteristic spectra for classification from Raman spectroscopy of NPC tissues. Furthermore, to compare the performance of GA-LDA, PCA-LDA algorithms are used to analyze the same Raman spectra of NPC tissues. The main purpose of this study is to explore the feasibility of classifying and diagnosing Raman spectra of NPC tissues with GA-LDA algorithms, and to provide an intuitive reference for the clinical diagnosis of Raman spectroscopy.

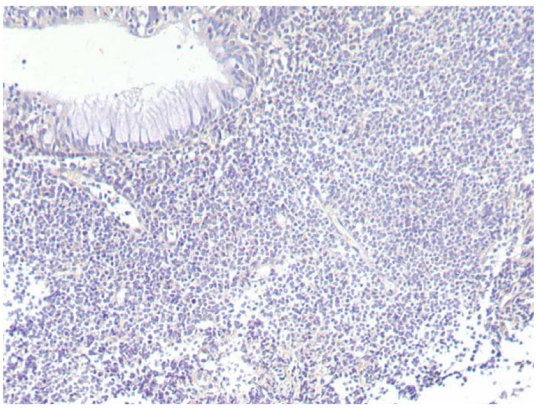
2 Materials and Methods

2.1 *Patients and Tissue Sample*

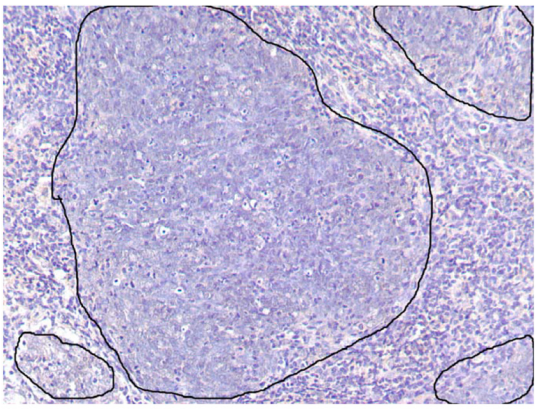
A total of 63 nasopharyngeal tissues were collected from 63 patients (17 women and 46 men with a median age of 44.6 years) who underwent endoscopic biopsies. Prior to research, all patients signed an informed consent to permit collection of excision specimens. After biopsies, the samples were divided into two parts, one for Raman measurements within 2 h and another for histopathologic examination conducted by a specialist nasopharyngeal pathologist after being fixed in 10% formalin solution. The results revealed that of the 63 nasopharyngeal tissue samples, 46 were dysplasia, and 17 were normal. Figure 1 shows the comparison of haematoxylin and eosin (H&E) stained tissue sections of normal and dysplastic nasopharyngeal tissues.

2.2 *Instrumentation*

The Raman spectroscopy was recorded with a confocal Raman microscopy (Renishaw, inVia, United Kingdom) in the range of 720 to 1800 cm^{-1} with a spectral resolution about 1 cm^{-1} under a 785 nm diode laser excitation. The spectra were collected in back-scattered geometry using a Leica DM2500 microscope equipped with objective 20×. The power of laser exposed on sample was about 1 mw with a spot diameter about 5 μm. The software package WIRE 3.2 (Renishaw) was employed for spectral acquisition and analysis. Each Raman spectra was acquired twice with an integration time of 10 s. All data was collected under the same conditions.



(a)



(b)

Fig. 1 (a) HE staining of normal nasopharyngeal tissue: a case showing nasopharyngeal pseudostratified ciliated columnar epithelium, no cell atypia. Magnification, 200×. (b) HE staining of nasopharyngeal carcinoma tissue: a case showing cancer cell nests (black line roped) with obvious nuclear atypia and abundant cytoplasm. Magnification, 200×.

2.3 *Data Preprocessing*

The Raman spectra acquired from nasopharyngeal tissues contained many autofluorescence and background noises.[28] A fifth-order polynomial was employed to fit the broad tissue autofluorescence background, and then this polynomial was subtracted from original spectra. In order to compare the changes of spectral shapes and relative peak intensities among different nasopharyngeal tissue samples, an area normalized of spectra was employed. Vancouver Raman algorithm was employed for spectra smoothing and baseline correction. It is an automated autofluorescence background subtraction algorithm based on modified multi-polynomial fitting by presetting of the relevant parameters such as size of Boxcar smooth, order of polynomial fit and stop criteria.[29]

2.4 *Genetic Algorithm*

GA is a cyclic process of iterative optimization.[30] The candidate solutions of the optimization problem are known as individuals that are presented with a variable sequence called chromosome or gene string. Chromosome is generally expressed as a simple string or numeric string, a process known as encoding. The first step of GA is to randomly generate a certain number of individuals to constitute a population. Then the individuals are evaluated by the fitness function, which is a particular type of

objective function that is used to measure the optimization variables to be solved.[31–33]

The following step is to produce individuals of the next generation and to form a population. This process is done by selecting and breeding which consists of crossover and mutation. Selection is carried out in accordance with the fitness value, but it does not mean that the individual with high fitness must be adopted. The principle of GA is that the individuals of higher fitness have greater probabilities of being accepted. A relatively optimal group can be assembled by initial data through this selection process. Then, the selected individuals go into the process of mating. Each two individuals produce two new individuals to replace the originals by crossover. Nonmating individuals remain unchanged. The chromosomes of mating parents exchange to produce two new chromosomes at a crossover point generated randomly on the chromosome. The next step is a mutation to produce fresh offspring of individuals. After this process (selection, crossover, and mutation) is finished, a new generation is created. Generation after generation, the overall fitness is increased. This process is constantly repeated: each individual is evaluated by the fitness, two individuals mate, mutate, and produce the next generation until the termination criteria is met.[31]

In this study, float point encoding is employed.[34,35] The entire spectrum which contained 1012 variables is divided into 253 segments, and each segment has four continuous data points. Each individual consists of seven fragments corresponding to 28 data points. The initial individual is made of seven randomly generated different integers less than 253. Parameters given for the GA are 100 generations of runs, 20 individuals each generation, mutation probability of 5%, and crossover rate of 70%. The individual is evaluated with LDA method. The standard of evaluation is the integration area under the receiver operating characteristic (ROC) curve.

LDA is a statistical technique used to group the features of objects through projecting an original feature space to a lower dimensional space with as little loss in discrimination as possible.[16] A method of leave-one-out cross-validation is employed to validate the classification performance. As such, one Raman spectra is left out, and the LDA model is rebuilt with the remaining Raman spectra to classify the withheld Raman spectra. This process is repeated until all Raman spectra retained are discriminated. A technique of roulette wheel selection is used to choose next-generation individuals. In this process, the fitness of all individuals is first normalized to 1 by dividing the fitness of each selection with the total fitness of all individuals. Then the normalized fitness is accumulated one by one to form probability space. The higher individual fitness equals a higher probability space on the 'wheel' and therefore a higher chance of selection. A random number less than 1 is generated when selection begins, and the individual whose probability space contains the random number is chosen. During the GA iterations, the most optimal individual with the best classification accuracy is completely preserved until the emergence of new individual with better fitness to replace it.

3 Results

Figure 2 shows the comparison of normalized average Raman spectra −1 standard deviations of NPC and normal tissues in the range from 720 to 1800 cm^{-1}. Primary Raman peaks are observed in both NPC and normal tissue at the following peak positions with tentative assignments:[4,36–43] 788 cm^{-1} [vs(O—P—O), cytosine, uracil ring breathing], 831 cm^{-1} [out of plane ring breathing tyrosine, O—P—O stretch], 855 cm^{-1} [ν(C—C) ring of proline, ring breathing of tyrosine], 940 cm^{-1} [ν(C—C) of proline, valine and protein backbone], 1004 cm^{-1} [νs(C—C) phenylalanine], 1033 cm^{-1} [C—H in plane bending of phenylalanine], 1095 cm^{-1} [vs (PO_{2^-}), ν(C—N) of protein], 1128 cm^{-1} [ν(C—C) of lipids/ν(C—N) of protein], 1158 cm^{-1} [C—C/C—N strething mode of proteins], 1208 cm^{-1} [ν(C—C_6H_5)) L-tryptophan and phenylalanine],[41] 1339 cm^{-1} [CH_3CH_2 wagging of collagen and polynucleotide chain], 1449 cm^{-1} [$\delta(CH_2)$ of proteins and lipids], 1660 cm^{-1} [amide I ν(C=O) collagen, α-Helix/ν(C=C) of lipid]. The strongest peaks are at 1004, 1339, 1449, and 1660 cm^{-1}. The spectral differences between NPC and normal tissues are clearly displayed from the difference spectra of Fig. 2(c), implying the enormous potential to diagnose NPC with Raman spectroscopy.

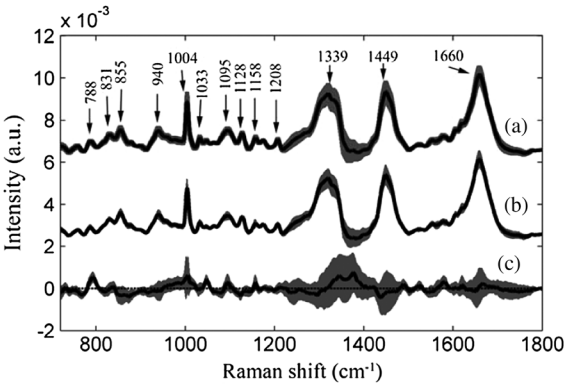


Fig. 2 Average Raman spectra of NPC tissue in the range from 720 to 1800 cm^{-1}. The solid lines indicate the average spectra and the shaded lines represent one standard deviation. (a) NPC tissue spectra, (b) normal tissue spectra, (c) cancer-normal difference spectra (the difference spectra intensity is enlarged five times for clear display).

The GA-LDA algorithms are developed to search for the significant Raman spectral features that are relevant to different nasopharyngeal tissue pathologies. The algorithm parameters are adjusted repeatedly to obtain the best subset of Raman variables for tissue differentiation. Figure 3 displays the best and mean ± 1SD area under ROC curve (AUC) of the individuals in 50 generations with GA-LDA algorithms. There is a significant increase of AUC by the GA-LDA algorithm. In order to

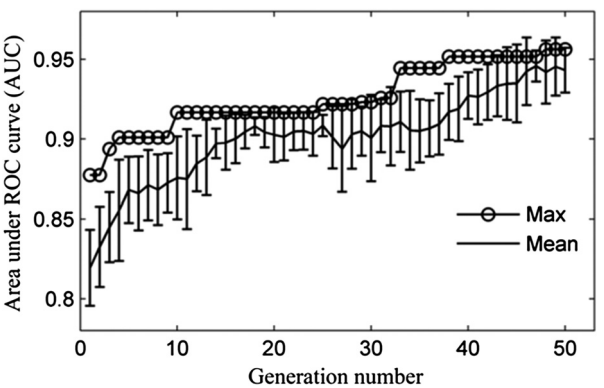


Fig. 3 The mean area under ROC curve (AUC) of the individual ±1 SD versus the best performance individuals in 50 generations with GA-LDA (20 individual, cross rate = 0.7, mutation rate = 0.05).

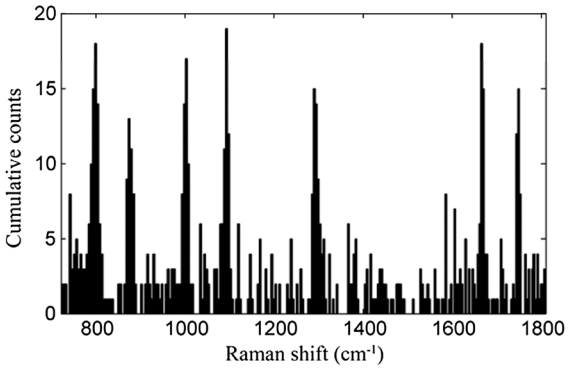


Fig. 4 The cumulative counts of Raman bands chosen with GA-LDA in 100 runs.

pick out diagnostic feature information contained in Raman spectra, the GA-LDA algorithms are repeated over 100 runs for selecting seven spectral bands. Figure 4 is the cumulative counts of Raman bands chosen with GA-LDA algorithms. It reveals the most useful spectral feature located in the regions of 792–805, 867–880, 996–1009, 1086–1099, 1288–1304, 1663–1670, and 1742–1752 cm^{-1}. The corresponding tentative assignments of significant Raman bands in nasopharyngeal tissue are listed in Table 1.[37,38,41–43] The intensity differences of these Raman features between normal and cancerous tissues are also verified to be significant ($p < 0.005$, unpaired Student's t-test, 2-sided, equal variances for classification and diagnosis of nasopharyngeal tissues.

The scatter plot of the linear discriminant scores of normal and NPC tissue using GA-LDA diagnostic model based on the seven significant Raman bands together with the leave-one-out cross-validation is displayed in Fig. 5. The dividing line produces diagnostic sensitivity of 69.2%, specificity of 100.0% and overall accuracy of 76.9% for discriminating NPC from normal nasopharyngeal tissue. Hence, the results of classification suggest that the GA-LDA algorithms develop a novel way to diagnose NPC from normal nasopharyngeal tissues by searching significant Raman features to build the discriminant model.

To further evaluate the performance of the diagnostic model developed by GA-LDA algorithms, the ROC curve is generated (Fig. 6) from the scatter plot in Fig. 5. The ROC curve is an intuitive method of reflecting the relationship of sensitivity and specificity. It is obtained by calculating the sensitivity and specificity of different diagnostic thresholds. The integration area under the ROC curves is positively correlated with the diagnostic accuracy. To comparatively assess the effective diagnostic performances of GA-LDA algorithms, the ROC curve of PCA-LDA algorithms are also generated. It is illustrated by the improvement in the diagnostic sensitivities and specificities that GA-LDA algorithms give more effective diagnostic performances for differentiation of NPC from normal tissues. The integration areas under the ROC curves of GA-LDA algorithms and PCA-LDA algorithms are 0.956 and 0.924, respectively. These results further confirm that GA-LDA algorithms yield a better diagnostic accuracy than the PCA-LDA algorithms.

4 Discussions

Raman spectroscopy is a unique noninvasive detection technique that can produce abundant information about molecular composition and structure of biological tissue. It may become a promising clinical diagnostic tool by probing subtle molecular changes relevant to tissue pathology. The Raman signal is very weak, but with the development of confocal Raman microscopy spectroscopy techniques, the development of Raman spectroscopy in biomedical field is greatly enhanced.[44–47] NPC is a malignancy of the head and neck with a marked racial and geographic distribution. NPC detection in the early stage is often difficult due to the nonspecific symptoms. Several groups have been studying the diagnosis of NPC with Raman spectroscopy technique. For example, Lau et al. implemented a preliminary study of normal and NPC tissue from six patients with Raman spectroscopy, they found consistent differences in three bands 1290–1320, 1420–1470, 1530–1580 cm^{-1}.[40] Feng et al., researched surface-enhanced Raman scattering of human nasopharyngeal tissue with gold nanoparticle. They found that Raman spectral imaging based on three characteristic peaks at 962, 725, and 1366 cm^{-1}, confirmed there were vigorous metabolism and strong enzymatic activities in tumor tissues.[41] This study demonstrates the great potential of Raman spectra for probing the biochemical heterogeneities of nasopharyngeal cancerous tissues using gold nanoparticles. But there are

Table 1 Tentative assignment of significant Raman bands identified by GA-LDA algorithm.

Peak position (cm^{-1})	Vibrational assignments	Intensity change (cancerous – normal)	*p*-value
792–805	*vs*(O–P–O) ring breathing	+	1.87*E* – 14
867–880	*v*(C–C) collagen	–	4.12*E* – 8
996–1009	*vs*(C–C) Phenylalanine	+	2.31*E* – 9
1086–1099	*vs* (PO_2^-) nucleic acids	+	1.09*E* – 6
1288–1304	δ(CH2), δ(CH3CH2) lipid	–	2.71*E* – 4
1663–1670	*v*(C=O) Amide I	+	3.46*E* – 3
1742–1752	*v*(C=O) phospholipids	–	1.42*E* – 3

Note: *v*-stretching mode; *vs*-Symmetric breathing; δ-bending mode. The mean intensity changes (increase: +; decrease: –) and the *P*-values of unpaired two-sided Student's t-test on the Raman band intensities of normal and cancerous nasopharyngeal tissues.

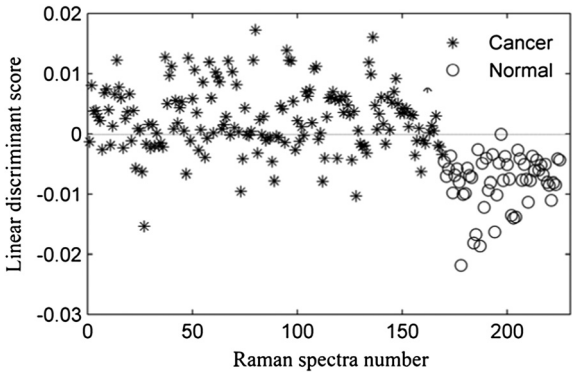


Fig. 5 Scatter plot of the linear discriminant scores of normal and NPC tissue using GA-LDA. The separate line produces diagnostic sensitivity of 69.2% (117/169) and specificity of 100.0% (56/56) for discriminating NPC from normal nasopharyngeal tissue.

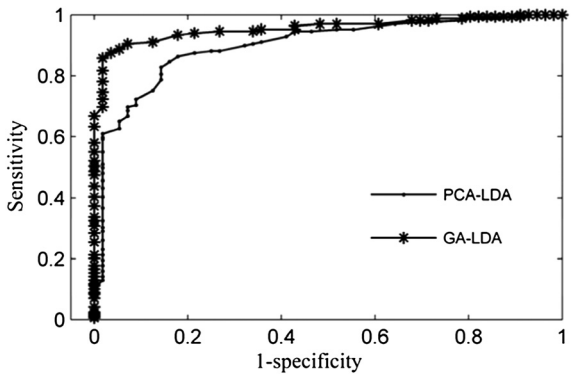


Fig. 6 Receiver operating characteristic (ROC) curves of Raman spectra discrimination results from NPC and normal tissue with PCA-LDA and GA-LDA algorithms together with the leave-one-out cross-validation method. The integration area under the ROC curves for PCA-LDA and GA-LDA are 0.924 and 0.956, respectively, illustrating the efficacy of GA-LDA in NPC diagnosis with Raman spectroscopy.

fluctuations depended on the excitation light intensity in the Raman scattering signals at the tissue level. Later Feng et al., measured blood plasma surface-enhanced Raman spectroscopy of NPC patients and normal volunteers. They successfully separated the two groups' spectra with sensitivity of 90.7% and specificity of 100% by PCA-LDA technique.[37]

In this work, we have studied Raman spectral properties of NPC and normal tissue with confocal Raman microscopy and explored the potential to improve the diagnostic performance with feature selection technique based on GA-LDA algorithms. The distinct spectra differences of normal and NPC tissue are observed from difference spectra in Fig. 2(c). It indicates that there is a great potential of differentiating dysplasia from normal nasopharyngeal tissue with Raman spectroscopy. The main intensity differences are located at around 788 cm^{-1} [νs(O—P—O), cytosine, uracil ring breathing], 1004 cm^{-1} [νs(C—C) phenylalanine], 1339 cm^{-1} [CH_3CH_2 wagging of collagen and polynucleotide chain], 1449 cm^{-1} [$\delta(CH_2)$ of proteins and lipids], 1660 cm^{-1} [amide I ν(C═O) collagen, α-Helix/ν(C═C) of lipid]. These results are basically in agreement with the other reports about NPC.[5,40,41] In the same range of 950 to 1650 cm^{-1}, our Raman spectral shape is the same to the Lau et al.'s, results, but different with the Feng et al.'s. The changes of Raman spectral shape are due to the use of surface enhanced Raman spectroscopy technology which only enhances some peaks.[40,41]

However, the Raman spectral pattern of normal and cancer tissue is very similar. There is an urgent need to develop the efficient diagnostic algorithms to interpret these tiny spectral changes. To solve the above problems, we employ feature selection technique based on GA-LDA algorithm to correlate significant spectral bands with tissue pathology. The GA-LDA results display that diagnostically important variables are limited to several Raman spectral bands that are related to proteins, nucleic acids and lipids. The important seven spectral bands are distributed in the regions around 792–805, 867–880, 996–1009, 1086–1099, 1288–1304, 1663–1670, and 1742–1752 cm^{-1}. These Raman bands mainly correlate with dysplasia progression. For instance, the Raman peak 867–880 cm^{-1} (C—C stretching mode of collagen) is found to decrease significantly, indicating that there is a relative reduction of collagen content in cancer tissue. The main cause is that cancer cells spread to underlying stromal layer and express a class of metalloprotease, leading to an overall decrease of collagen content in cancer tissue.[48] The decrease of Raman bands 1288–1304 cm^{-1} (CH3CH2 bending of lipid) and 1742–1752 cm^{-1} (C═O stretching of phospholipids) associated with lipids illustrates that the proportion of fat content in the cancerous tissue is greatly reduced. The likely reason is that amplified cancer cells consume a lot of fat, resulting in decrease of lipid molecules.[49,50] Nucleic acid-related Raman bands of 1086–1099 cm^{-1} (PO_2^- Symmetric breathing of nucleic acids) are significantly enhanced, as cancer cells proliferate indefinitely, and DNA replicates greatly, leading to increased content of cells DNA.[41,42] In the Raman bands of 1086–1099 cm^{-1} and 1530–1560 cm^{-1}, our difference spectra are consistent with Lau et al.'s, results. While in the Raman bands of 1420 and 1470 cm^{-1}, our difference spectra are different from their results. This may be attributed to the measurement conditions and amount of sample. We measured 63 patients with the beam spot diameter about 5 μm, and Lau measured 6 patients with the beam spot diameter of 3.5 mm.

There are several indicators to evaluate the diagnostic capabilities of a particular algorithm such as sensitivity, specificity, accuracy and AUC, in which AUC can be a comprehensive reflection of the diagnostic algorithm to identify the disease. Other teams have also chosen accuracy as evaluation standard.[51] In this study, AUC is adopted as the fitness function of GA and has obtained better search results in Fig. 3. The best AUC is 0.956, corresponding diagnostic accuracy 0.769. We also attempt to use the diagnostic accuracy as fitness function to search the characteristic Raman bands, and get the best AUC 0.944, accuracy 0.780. Although the results of the evaluation functions are close, we believe that the AUC is more suitable to assess the diagnostic model.

The diagnostic model built by GA-LDA technique searching important Raman feature spectra of nasopharyngeal tissue produces a sensitivity of 69.2%, specificity of 100.0% and overall accuracy of 76.9%. The area of 0.956 under the ROC curve further affirms the efficient diagnostic performances of GA-LDA-based algorithms for classification NPC from normal tissue. To compare the performances of GA-LDA algorithms with conventional multivariate statistical technique, PCA-LDA algorithms are applied to classify the same Raman dataset. The cumulative contribution rate in the first 20 PCs is 95.3% of

total variance, corresponding diagnostic accuracy of 73.3%, sensitivity of 63.3% and specificity of 94.6% with unpaired two-side t-test ($p < 0.05$) together with the leave-one-out cross-validation. The integration area of ROC curve is 0.924 for PCA-LDA algorithms as display in Fig. 6. Although PCA-LDA algorithms yield good classification effects, it can't explain the contributions to discriminant results come from what biochemical component associated with tissue malignancy because PCA variables are extracted from entire Raman spectra of nasopharyngeal tissue. The GA-LDA algorithms provide a simplified diagnostic model with improved diagnostic accuracy by the selection of significant Raman feature bands related to biochemical components such as proteins, nucleic acids and lipids. The possible reason of GA-LDA algorithms improving diagnostic performances is that GA is a heuristic optimization technique that can search characteristic Raman spectra with distinct classification attributes.[30] Other groups have also applied feature selection technique based on GA with Raman spectroscopy for the classification. For example, Duraipandian et al., use GA partial least squares discriminant analysis and Raman spectroscopy to identify biomolecular changes of cervical tissue associated with dysplastic transformation. They achieved a diagnostic accuracy of 82.9%, sensitivity of 72.5% and specificity of 89.2% for precancer detection.[27] Lavine et al., successfully classified several wood types by the application of GA to extract features of woods' Raman spectroscopy.[52] All of above results demonstrate that the GA-LDA algorithms are very attractive techniques that are expected to provide a more accurate diagnostic model for the clinical application of Raman spectroscopy.

5 Conclusion

In conclusion, an efficient diagnostic model based on GA-LDA algorithms picking out the characteristic Raman bands associated with biochemical components is developed and applied to classify the NPC from normal tissue. An improved diagnostic accuracy of 76.9%, sensitivity of 69.2% and specificity of 100.0% is obtained. Compared with the PCA-LDA algorithms, the GA-LDA technique can build a simpler diagnostic model with clearer physical meaning and higher diagnostic efficiency. This work demonstrates that the Raman spectroscopy associated with GA-LDA diagnostic algorithms has enormous potential to detect and diagnose NPC.

Acknowledgments

The authors would like to acknowledge the financial support of the Dongguan Science and Technology Project (201010815207), the National Natural Science Foundation of China (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).

References

1. W. I. Wei and J. S. T. Sham, "Nasopharyngeal carcinoma," *The Lancet* **365**(9476), 2041–2054 (2005).
2. J. H. C. Ho, "An epidemiologic and clinical study of nasopharyngeal carcinoma," *Int. J. Radiat. Oncol. Biol. Phys.* **4**(3–4), 183–198 (1978).
3. K. Tabuchi et al., "Early detection of nasopharyngeal carcinoma," *Int. J. Otolaryngol.* **2011**, 638058 (2011).
4. Y. Hu et al., "Classification of normal and malignant human gastric mucosa tissue with confocal Raman microspectroscopy and wavelet analysis," *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **69**(2), 378–382 (2008).
5. A. T. Harris et al., "Raman spectroscopy in head and neck cancer," *Head Neck Oncol.* **2**, 26 (2010).
6. E. B. Hanlon et al., "Prospects for in vivo Raman spectroscopy," *Phys. Med. Biol.* **45**(2), R1–R59 (2000).
7. A. Shapiro et al., "Raman molecular imaging: a novel spectroscopic technique for diagnosis of bladder cancer in urine specimens," *Eur. Urol.* **59**(1), 106–112 (2011).
8. N. Li et al., "Micro–Raman spectroscopy study of the effect of mid-ultraviolet radiation on erythrocyte membrane," *J. Photochem. Photobiol. B Biol.* **112**, 37–42 (2012).
9. T. Kawabata et al., "Optical diagnosis of gastric cancer using near-infrared multichannel Raman spectroscopy with a 1064-nm excitation wavelength," *J. Gastroenterol.* **43**(4), 283–290 (2008).
10. H. Abramczyk et al., "The hallmarks of breast cancer by Raman spectroscopy," *J. Molec. Struct.* **924–926**, 175–182 (2009).
11. C. A. Lieber et al., "In vivo nonmelanoma skin cancer diagnosis using Raman microspectroscopy," *Laser. Surg. Med.* **40**(7), 461–467 (2008).
12. C. M. Krishna et al., "Raman spectroscopy studies for diagnosis of cancers in human uterine cervix," *Vibrational Spectrosc.* **41**(1), 136–141 (2006).
13. P. Crow et al., "Assessment of fiberoptic near-infrared Raman spectroscopy for diagnosis of bladder and prostate cancer," *Urology* **65**(6) 1126–1130 (2005).
14. B. Brożek-Płuska et al., "Breast cancer diagnostics by Raman spectroscopy," *J. Mol. Liq.* **141**(3), 145–148 (2008).
15. K. E. Shafer-Peltier et al., "Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosisin vivo," *J. Raman Spectrosc.* **33**(7), 552–563 (2002).
16. M. Sattlecker et al., "Assessment of robustness and transferability of classification models built for cancer diagnostics using Raman spectroscopy," *J. Raman Spectrosc.* **42**(5), 897–903 (2011).
17. S. K. Teh et al., "Diagnostic potential of near-infrared Raman spectroscopy in the stomach: differentiating dysplasia from normal tissue," *Brit. J. Cancer* **98**(2), 457–465 (2008).
18. B. Yan et al., "Discrimination of parotid neoplasms from the normal parotid gland by use of Raman spectroscopy and support vector machine," *Oral Oncol.* **47**(5), 430–435 (2011).
19. M. Gniadecka et al., "Diagnosis of basal cell carcinoma by Raman spectroscopy," *J. Raman Spectrosc.* **28**(2–3), 125–129 (1997).
20. H. Shinzawa et al., "Multivariate data analysis for Raman spectroscopic imaging," *J. Raman Spectrosc.* **40**(12), 1720–1725 (2009).
21. H. W. Ressom et al., "Peak selection from MALDI-TOF mass spectra using ant colony optimization," *Bioinformatics* **23**(5), 619–626 (2007).
22. R. M. Balabin and S. V. Smirnov, "Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data," *Anal. Chim. Acta* **692**(1–2), 63–72 (2011).
23. A. Durand et al., "Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton–viscose textiles," *Anal. Chim. Acta* **595**(1–2), 72–79 (2007).
24. R. M. Jarvis and R. Goodacre, "Genetic algorithm optimization for preprocessing and variable selection of spectroscopic data," *Bioinformatics* **21**(7), 860–868 (2005).
25. C. R. W. Leping Li, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics* **17**(12), 1131–1142 (2001).
26. M. A. J. Asha Gowda Karegowda and A. S. Manjunath, "Feature subset selection using cascaded GA &CFS: a filter approach in supervised learning," *Int. J. Comput. Appl.* **23** (2), 1–10 (2011).
27. S. Duraipandian et al., "*In vivo* diagnosis of cervical precancer using Raman spectroscopy and genetic algorithm techniques," *Analyst* **136**(20), 4328–4336 (2011).
28. Z. Zhao et al., "A spectrum signals detection method for surface enhanced Raman scattering under high fluorescence and background noise," *Spectrosc. Spectral Anal.* **30**(8), 2146–2150 (2010).

29. H. L. Jianhua Zhao, D. I. Mclean, and H. Zeng, "Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy," *Appl. Spectrosc.* **61**(11), 1225–1232 (2007).
30. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, New York (1989).
31. J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*, Michigan Press, Ann Arbor, Michigan (1975).
32. P. Rocca et al., "Evolutionary optimization as applied to inverse scattering problems," *Inverse Probl.* **25**(12), 123003 (2009).
33. C. Pinkey, D. Kusum, and P. Millie, "Optimizing cnc turning process using real coded genetic algorithm and differential evolution," *Global J. Technol. Optimization* **2**, 57–165 (2011).
34. M. Cui, "A float code genetic algorithm based on orthonormal multiwavelet denoising mutation," in *Int. Joint Conf. on Computational Sciences and Optimization, CSO 2009*, Vol. 1, pp. 172–175, IEEE, Sanya, Hainan (2009).
35. Y. T. Kao and E. Zahara, "A hybrid genetic algorithm and particle swarm optimization for multimodal functions," *Appl. Soft. Comput.* **8**(2), 849–857 (2008).
36. I. Notingher et al., "Spectroscopic study of human lung epithelial Cells (A549) in culter: living cells versus dead cells," *Biopolymers* **72**(4), 230–240 (2003).
37. S. Feng et al., "Nasopharyngeal cancer detection based on blood plasma surface-enhanced Raman spectroscopy and multivariate analysis," *Biosens. Bioelectron.* **25**(11), 2414–2419 (2010).
38. Z. Huang et al., "Near-infrared Raman spectroscopy for optical diagnosis of lung cancer," *Int. J. Cancer* **107**(6), 1047–1052 (2003).
39. H. Wang et al., "Depth-resolved in vivo micro–Raman spectroscopy of a murine skin tumor model reveals cancer-specific spectral biomarkers," *J. Raman Spectrosc.* **42**(2), 160–166 (2011).
40. D. P. Lau et al., "Raman spectroscopy for optical diagnosis in normal and cancerous tissue of the nasopharynx-preliminary findings," *Laser. Surg. Med.* **32**(3), 210–214 (2003).
41. S. Feng et al., "Gold nanoparticle based surface-enhanced Raman scattering spectroscopy of cancerous and normal nasopharyngeal tissues under near-infrared laser excitation," *Appl. Spectrosc.* **63**(10), 1089–1094 (2009).
42. N. Stone et al., "Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers," *J. Raman Spectrosc.* **33**(7), 564–573 (2002).
43. J. W. Chan et al., "Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells," *Biophys. J.* **90**(2), 648–656 (2006).
44. T. Bocklitz et al., "A comprehensive study of classification methods for medical diagnosis," *J. Raman Spectrosc.* **40**(12), 1759–1765 (2009).
45. E. Zuser, T. Chernenko, J. Newmark, M. Miljkovic, and M. Diem, "Confocal Raman microspectral imaging (CRMI) of murine stem cell colonies," *The Analyst.* **135**(12), 3030–3033 (2010).
46. W. G. L. P. J. Caspers and G. J. Puppels, "Combined *in vivo* confocal Raman spectroscopy and confocal microscopy of human skin," *Biophys. J.* **85**(1), 572–580 (2003).
47. M. Larraona-Puy et al., "Development of Raman microspectroscopy for automated detection and imaging of basal cell carcinoma," *J. Biomed. Opt.* **14**(5), 054031 (2009).
48. I. Stamenkovic, " Matrix metalloproteinases in tumor invasion and metastasis," *Sem. Cancer Biol.* **10**(6), 415–453 (2000).
49. A. S. H. Karen et al., "Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosis *in vivo*," *J. Raman Spectrosc.* **3**(7), 552–563 (2002).
50. A. Beljebbar et al., "Identification of Raman spectroscopic markers for the characterization of normal and adenocarcinomatous colonic tissues," *Crit. Rev. Oncol. Hematol.* **72**(3), 255–264 (2009).
51. M. S. Bergholt et al., "*In vivo* diagnosis of gastric cancer using Raman endoscopy and ant colony optimization techniques," *Int. J. Cancer.* **128**(11), 2673–2680 (2011).
52. B. K. Lavine et al., "Raman spectroscopy and genetic algorithms for the classification of wood types," *Appl. Spectrosc.* **55**(8), 960–966 (2001).