International Journal of Pharmacy and Pharmaceutical Sciences

ISSN- 0975-1491 Vol 8, Issue 6, 2016

Review Article

SIGNAL PROCESSING FOR RAMAN SPECTRA FOR DISEASE DETECTION

AFAF ROZAN MOHD RADZOL[a], LEE YOOT KHUAN[a,b,c], WAHIDAH MANSOR[a,b,c], FAIZAL MOHD TWON TAWI[d]

[a]Faculty of Electrical Engineering, [b]Computational Intelligence Detection RIG, [c]Pharmaceutical and Lifesciences Communities of Research, Universiti Teknologi MARA, Shah Alam, Selangor DE, Malaysia, [d]Jabatan Kejuruteraan Elektrik, Politeknik Seberang Perai Permatang Pauh, P Pinang, Malaysia
Email: afafrozan944@ppinang.uitm.edu.my

Received: 26 Jan 2016 Revised and Accepted: 20 Apr 2016

ABSTRACT

Raman Spectroscopy enables in-depth study into the molecular structure of solid, liquid and gasses from its scattering spectrum. As such, the spectrum could offer a biochemical fingerprint to identify unknown molecules. Surface Enhanced Raman Spectroscopy (SERS) amplifies the weak Raman signal by 10^{+3} to 10^{+7} times, revolutionary making the method appealing to the research community. SERS has been proven useful for disease detection from a medium such as a cell, serum, urine, plasma, saliva, tears. The spectra displayed are noisy and complicated by the presence of other molecules, besides the targeted one. Moreover, the difference between the infected and controlled samples is far too minute for detection by the naked human eyes. Hence, signal processing techniques are found crucial to single out fingerprint of the target molecule from biological spectra. Our work here examines signal processing techniques attempted on SERS spectra for disease detection, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Logistic Regression Analysis (LRA). It is found that PCA-LDA is the most popular (45%), ensued by PCA-ANN (33%) and SVM (22%). PCA-SVM yields the highest in accuracy (99.9%), followed by PCA-ANN (98%) and LRA (97%). PCA-LDA and SVM score the highest in both sensitivity-specificity.

Keywords: Raman Spectra, Surface Enhanced Raman Spectroscopy (SERS), Neural Network (NN), Support Vector Machine (SVM), Logistic Regression Analysis (LRA), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA).

INTRODUCTION

Early diagnosis could offer life a second chance for those suffering from terminal diseases. This is because it allows early intervention and treatment. There have been interests shared amongst fundamental scientists, biomedical engineers and medical personnel to develop diagnostic tools, which are able to provide early detection of diseases. This has spun new research areas such as nanomaterial-based sensors, photonic sensors, biosensors, compounded signal processing techniques and nanotechnology medical systems.

Raman Spectroscopy provides a mean to study the structural property of solid, liquid and gasses to a molecular scale, from its scattering spectrum. This offers a detailed biochemical fingerprint, useful for identification of unknown molecule [1]. In biomedical application, it has been used for disease detection. Nevertheless, the Raman signal is so weak that it is useless to the users. Researchers have tried different ways in sample preparation, sample illumination and scattered light detection to enhance the intensity of Raman signal. One of the successful techniques is Surface Enhanced Raman Spectroscopy (SERS) [2].

SERS is a form of Raman spectroscopy which amplifies the intensity of signal through adsorption to or interaction with metal surfaces, usually nanoscale featured gold or silver surfaces, or, gold or silver colloids [3]. Raman spectra obtained from SERS is capable of providing information about the molecular structure of the sample [4], and hence serve as fingerprints for chemical and biological systems [5]. SERS shares the advantages of Raman spectroscopy: (i) amount of sample required is minimal; (ii) preparation for spectroscopy is minimal; (iii) analysis is simple and fast; (iv) test is non-destructive and easily reproducible [6-8]. Owing to this, it is attracting more and more biomedical applications, in particular for disease detection, such as breast cancer [9, 10], lung cancer [11, 12], head and neck cancer [13, 14], skin diseases [4], colon and rectum cancer [15], nasopharyngeal cancer [16], gastric cancer [17], cervical cancer [18], prostate cancer [19], diabetes [20] and Acquired Immune Deficiency Syndrome (AIDs) [21]. Existing works have shown application of SERS to produce Raman spectra of infected samples from entities, such as cell [22, 14], tissue [11, 23], serum [15-17, 19, 24], plasma [16, 17], urine [20], saliva [25, 26] and tears [27, 28], to detect for biochemical anomalies, in comparison with the controlled samples. Raman spectra obtained from the biological samples usually display complex patterns, consisting of peaks representing a mixture of molecules, namely proteins, nucleic acids, lipids, sugars, etc. Furthermore, the difference between the infected and controlled samples is too minute to be identified by the naked eyes. Hence, extraction of signature features from Raman spectra for detection of disease with signal processing techniques is essential.

This paper intends to illustrate a confluence of Raman spectrometry, a fundamental analysis tool for the pharmaceutical research community, with biomedical signal processing, fundamental tools for engineers that enable automation. Integration of these two interdisciplinary tools holds the potential that enables rapid, non-destructive Raman for on-line process monitoring and analysis in the pharmaceutical industry. Signal processing technique for the purpose of automated detection, in general, employs the following stages: signal pre-processing, signal representation, feature extraction, feature reduction, feature selection and classification. For analysis of complex Raman spectra of biological samples, previous works with signal processing techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression Analysis (LRA) or a combination of these, were reported.

This paper first provides background, theories and algorithmic procedures underlying these signal processing techniques. It then delves further into implementation details of these techniques and their compounded forms on Raman spectra, with applications on patient data for disease detection from samples of blood, saliva, tissue, cell and so on, to draw similar application in pharmaceutical research. Works elaborated in this review are from high impact journals and proceedings indexed in established databases such as SCOPUS, Web of Science, ELSEVIER, PubMED and MEDLINE, with the inclusion criteria of Raman or SERS for analysis, signal processing techniques for feature extraction and/or classification, performance evaluation as well as biomedical and pharmaceutical application oriented.

Signal processing techniques for the analysis of Raman spectra

This Section explains theories and methodologies underlying the signal processing techniques. Table 1 summarizes the signal processing techniques applied on Raman spectra for disease detection from our literature survey. In terms of usage, it is found that PCA-LDA is the most popular (45%), ensued by PCA-ANN (33%) and SVM (22%).

Fig. 1 displays detection performance attained by the different signal processing techniques tabulated in table 1. It can be observed that PCA-SVM yields the highest accuracy (99.9%), followed by PCA-ANN (98%) and LRA (97%). On the other hand, the performance of PCA-LDA and PCA-SVM place them at the leftmost of the ROC graph, being optimal in both sensitivity and specificity.

Principle component analysis

Principal Component Analysis (PCA) is a multivariate analysis technique for unsupervised reduction in dimension and/or classification. It transforms a large chunk of data into fewer new variants by reducing redundancy and minimizing noise. After transformation, the shape and location of the original spectra change as it migrates to a different space. Classification is based on these new variant features. This makes it useful for the analysis of Raman spectra obtained from biological samples, which contain a high volume of data with complex characteristic.

PCA reduces the usually high volume of spectral data to a few principal components, a combination of new datasets by the following equation,

$$O(\chi) = P_1C_1.P_1(\chi) + P_2C_2.P_2(\chi) + \cdots + P_nC_n.P_n(\chi) (1)$$

$O(\chi)$ are the original spectra; $P_n(\chi)$ is the principal component spectroscopy and P_nC_n are the principal components of the spectra [29].

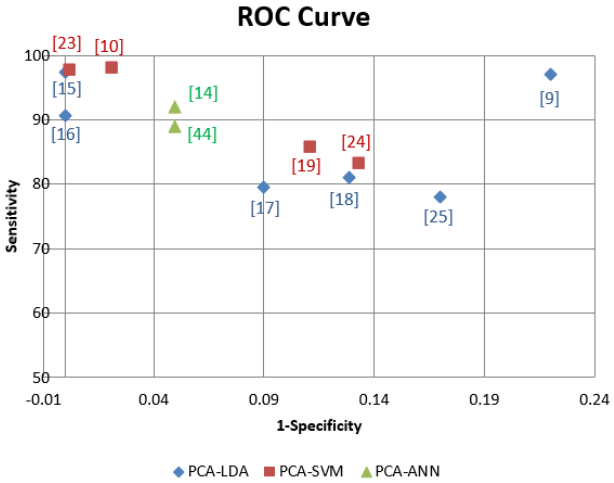

Fig. 1: Receiving operating curve of signal processing techniques for Raman spectra analysis in disease detection

Table 1: Signal processing techniques for Raman spectra analysis in disease detection.

Disease	Sample	Substrate	Technique	Percentage		
				Acc	Sen	Spe
Nasopharangeal cancer [16]	Blood plasma	Ag NP	PCA-LDA	NA	90.7	100
Gastric cancer [17]	Blood plasma	Ag NP	PCA-LDA	NA	79.5	91
Colorectal cancer [15]	Blood serum	Au NP	PCA-LDA	NA	97.4	100
Lung cancer [25]	Saliva	Oven heated, Ag NP	PCA-LDA	80	78	83
Cervical cancer [22]	Cell	In vivo using, Fibre optic probe	PCA-LDA	84.1	81	87.1
Breast cancer [9]	Blood serum	Aluminium	PCA-LDA	NA	97	78
Diabetes [20]	Urine	Aluminium holder	PCA-QDA	70	NA	NA
Flavivirus infection [35]	Saliva	Gold coated slide	LDA	93.75	87	100
Colonic cancer [23]	Tissue	-	PCA-SVM	>98	>97.7	>99.8
Breast cancer [10]	Tissue	Quartz cuvette	PCA-SVM	99.7	98	97.9
Esophageal cancer [24]	Blood serum	Ag NP	PCA-SVM	85.2	83.3	86.7
AIDS [21]	Saliva	Nanochip	SVM	90.9	95.6	100
Flavivirus infection [39]	Saliva	Gold coated slide	PCA-SVM	98.71	98.97	98.44
Prostate cancer [19]	Tissue &Cell	Glass slide	PCA-SVM	NA	85.7	88.9
Thalassemia [45]	Cell	-	PCA-ANN	97.6	NA	NA
Skin Lesion [4]	Skin	-	PCA-ANN	94.8±3	NA	NA
Thyroid [14]	Cell line	-	PCA-ANN	NA	92	95
Liver cancer [44]	Blood serum	Sample tube	PCA-ANN	80	89	95
Lung cancer [11]	Saliva	Nanochip	LRA	96.9	-	-

Silver nanoparticle (Ag-NP), Principal Component Analysis (PCA), Artificial Neural Network (ANN), Logistic Regression Analysis (LRA), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Accuracy (Acc), Sensitivity (Sen), Specificity (Spe).

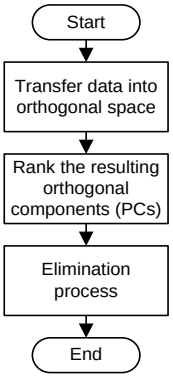

Fig. 2: Algorithm for principal component analysis

Fig. 2 describes steps in the development of PCA algorithm. Firstly, PCA transforms the input feature data into an orthogonal space using orthogonal linear transformation. The outcome is orthogonal components known as PCs. Secondly; the PCs are arranged according to their variance. Variance is a measure of variability in a sample distribution and is expressed as the average squared deviation of each sample from its mean as follows,

$$\text{Variance} = \frac{\sum(\text{sample}-\text{mean})^2}{\text{total sample}} \ldots\ldots (2)$$

While the percentage of variance is given as follows,

$$\%\ \text{of variance} = \frac{100\times\text{variance of the n}^{\text{th}}\ \text{PCs}}{\text{total variance}} \ldots\ldots (3)$$

The final step eliminates PCs with the least contribution to variance in the dataset. In principle, a selection of PCs is enough to account for the total variance in the observed variables. PCs with the largest

variance are ranked first while those with the least variance are ranked last. However, in practice, the cost function is used in addition to ranking the significance of PCs components. The three cost functions in common use are:

i. Eigenvalue One Criterion (EOC)–This criterion keeps PCs with eigenvalue equal or greater than one, as shown in fig. 3, for their variance, is higher [30].

ii. Cumulative percent of variance (CPV)–This criterion retains components of which their CPV accounts for a designated threshold or higher, usually 80% [29].

iii. Screen test–This is usually used in conjunction with (i). It plots the graphic representation of the relationship between eigen values and PCs as illustrated in fig. 3. PCs with a larger gap between other components are kept, such as the 1st, 2nd, and 3rd PCs, while the 4th and higher PCs with small gaps are eliminated [31, 32].

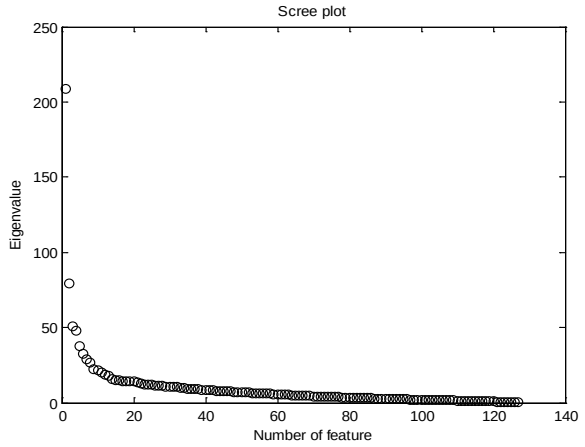

Fig. 3: Scree plot of control and NS1 adulterated saliva dataset (adapted from [32])

PCA is a technique for extracting significant components. For Raman spectra analysis, PCA is found integrated with classification techniques such as LDA, ANN, and SVM which will be discussed in the following sections.

Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a supervised multivariate analysis method to reduce the dimension of data and/or classification, with the assumption that the covariance matrix for each class is identical. It can be used to discriminate between two or more groups of data once a suitable linear transformation is determined. After transformation, the data location does not change. The transformation marks a decision region between the given classes, according to a criterion that aims to increase the separability between classes. Classification by LDA is based on data, unlike PCA which is based on new variant features. It is widely used in statistics, pattern recognition and machine learning [33, 34].

Fig. 3 describes steps in the computation of the LDA algorithm. The algorithm starts by projecting the input data onto the LDA space with the following equation,

$$Z_i = A^T Y_i \quad \ldots (4)$$

Where, i=1,2,⋯,n, Y_i is the input; Z_i is the corresponding data in the LDA space and A is the linear transformation matrix.

Next, scatter matrix data analysis is applied to select the major difference between classes. Here, the within-class and between-class scatter are made the criterion for class separability. First, the sample meanμ of all the LDA data, the sample mean of classes μ_i and covariance matrix S_i for each group are computed using (5) and (6). Ni is the number of data in the i^{th}-group.

$$\mu_i = \frac{1}{N_i}\sum Z_i \quad \ldots (5)$$

$$S_i = \sum_{i=1}^{N_k}(Z_i^k - \mu_k)(Z_i^k - \mu_k)^T \quad \ldots (6)$$

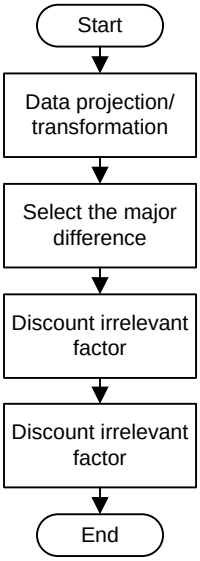

Fig. 4: Algorithm for linear discriminant analysis

Second, the between-class scatter matrix S_b and the within-class scatter matrix S_w are obtained using (7) and (8), where N is the total number of data; NK is the number of data in k^{th}-group.

$$S_b = \sum_{k=1}^{L}\frac{N_k}{N}(u_k - u)(u_k - u)^T \quad \ldots (7)$$

$$S_w = \frac{1}{N}\sum_{k=1}^{L}\frac{N_k}{N}(Z_i^k - u_k)(Z_i^k - u_k)^T \quad \ldots (8)$$

Finally, the major difference between classes comes from solution to the following generalized eigenvalue problem, where λ is the eigenvalue while w_i's are the classes.

$$S_b w = \lambda S_w \rightarrow [w_1, w_2, \ldots, w_3] \quad \ldots (9)$$

In order to discount the irrelevant factor, a non-redundant set of features consisting of only eigenvectors corresponding to non-zero eigenvalues are kept while those corresponding to zero eigenvalues are eliminated. For classification, the Euclidean distance, i.e. distance between two points measured by a ruler, is adopted.

For application to Raman spectra, this technique is found used in tandem with PCA, known as PCA-LDA algorithm [9, 12, 15-17], for data reduction. The principal components extracted from PCA are discriminated by LDA, as illustrated by the following examples.

Detection of nasopharyngeal cancer using SERS technique was investigated [16]. SERS spectra acquired from 15 µl blood plasma samples with 15uL silver nanoparticles colloid (Ag NP) substrate were first processed with PCA to extract the principal components, which were then passed to LDA to discriminate for samples with nasopharyngeal cancer. Prior to PCA-LDA algorithm, the spectra were pre-processed to remove the fluorescence background using the multi-polynomial fitting algorithm. Then the spectra were normalized using integration of the area under the curve. The result from this reported sensitivity and specificity of 95.4% and 100% respectively. When the same technique was applied to detection of gastric cancer, a sensitivity of 79.5% and specificity of 91% were attained [17]. A similar technique with a different substrate type, gold nanoparticle colloids (Au NP), detects colorectal cancer with a sensitivity of 97.4% and specificity of 100% [15].

PCA-LDA analysis of Raman spectra of blood serum from breast cancer patients was also reported [9]. Blood serum samples of 11 patients and 12 healthy volunteers were first analyzed using Raman system with aluminum substrate. The raw spectra were pre-

processed by Savitsky-Golay filter for smoothing and cubic spline interpolation for regression to remove baseline drift. By using PCA, 10 PCs from spectra of patients and 7 PCs from spectra of healthy volunteers were identified for discrimination between the two groups. A sensitivity of 92.2% and specificity of 86.0% were achieved from cross-validation technique of PCA-LDA. Of recent, another group working on the same research problem furthered the investigation to discriminate between the different stages of the disease, with a similar analysis. Its objective was to detect luminal A tumor, an indicator for early stage breast cancer. The group reported that early stage breast cancer indicator gave a better diagnostic performance, 90% of sensitivity and 95% of specificity, than the advance stage indicator of 80% and 85% only respectively [35].

Li et al. reported on discrimination of lung cancer patients by analyzing the Raman spectra of their saliva samples [12, 25]. In the study, saliva samples from 21 lung cancer patients and 22 normal subjects were collected. Microwave oven heated silver colloid was used as the substrate. Prior to discrimination by PCA-LDA, the raw spectra were pre-processed by normalization, smoothing, and baseline correction. An accuracy of 80%, the sensitivity of 78% and specificity of 83% were attained.

A recent study reported an application of PCA-LDA on Raman spectra for in vivo diagnosis of cervical cancer [22]. In vivo measurement of normal [n=993] and dysplasia [n=247] cervixes were measured using an NIR Raman system coupled with a ball lens fiber optics confocal Raman from 84 non-pregnant patients. Before the measurement, a 5% acetic acid was applied to the cervix to distinguish between normal and abnormal epithelium. Principal components from PCA were fed into LDA classifier and validated using leave-one-out cross-validation method. The method yielded a diagnosis accuracy of 84%, the sensitivity of 81% and specificity of 87.1%.

In another study recently, the Raman spectra of urine collected from patients with diabetes mellitus and hypertension was measured to evaluate the risk of developing renal lesion [20]. 100 µl of urine was placed in an aluminum holder with the vessel and analyzed using Raman system. From PCA, urea, creatinine, and glucose were identified as significant features for classification between the groups. QDA classifier, a higher order of LDA, was found to achieve 70% of overall classification rate.

Our preliminary attempt to classify salivary non-structural protein 1 (NS1), a biomarker for early detection of Flavivirus infection, from their Raman spectra using LDA has reported an encouraging performance, of 93.75% in accuracy, 87% in sensitivity and 100% in specificity [36]. Dengue fever, Yellow fever, Japanese encephalitis, Tick-borne encephalitis are amongst the diseases caused by Flavivirus infection [37]. In our study, 40 spectra of NS1 adulterated saliva at different concentrations are classified. The input to the LDA classifier consists of 16 features and the ratio of training to test sets is 80:20.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is an algorithm based on supervised learning, non-probabilistic models with associated learning algorithms, for the purpose of regression analysis and classification, to analyze and differentiate between data or patterns [38]. By tagging each training data to one of the two classes, the SVM training algorithm predicts a model to represent the training data. The data in the input space are first transformed into a feature space. Then a separation hyperplane is introduced to divide the data into the different classes. For linearly separable data, only a simple straight hyperplane is sufficient to classify the data. However, for non-linearly separable data, the transformation requires a high-dimensional feature space together with soft margin and a kernel function. The soft margin chooses a hyperplane classifier that allows misclassification of data while maximizing the margin so that the hyperplane classifier can separate the input data into different classes in the feature space with minimal error. New data are then transformed into that same feature space to be relocated into one of the classes, based on which side of the hyperplane do they fall on. The transformation algorithm is known as the kernel function. The kernel function is the determinant component to this algorithm. The hyperplane is optimal when the large distance between hyperplane and data point is obtained. Theoretically, the best model of SVM depends on the regularization parameter C and kernel function parameters, which includes optional constant(c), slope (α), polynomial degree (d) and RBF sigma (σ). C is a soft margin parameter of the error term. A Higher value of C indicates a good proportion of the training data are classified correctly; lower C results in a more flexible hyperplane that try to minimize the margin error. The kernel function is expressed as K, where, x_i and x_j are the input vectors, φ is a function that maps x into higher dimensional space,

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \quad \ldots\ldots (10)$$

Linear, polynomial and Gaussian Radial Basis Functions as in (11), (12) and (13) are kernel functions commonly used in SVM.

$$Linear: K(x_i, x_j) = x^T . x_j + c \quad \ldots\ldots (11)$$

$$Polynomial: K(x_i, x_j) = (\alpha x_i . x_j + c)^d \quad \ldots\ldots (12)$$

$$Gaussian\ RBF: K(x_i, x_j) = e^{\frac{\|x_j - x_i\|^2}{2\sigma^2}} \quad \ldots\ldots (13)$$

Fig. 4 shows steps in developing the SVM algorithm. Firstly, sets of input data vectors are arranged in an n-dimensional space and transformed into a feature space by the chosen kernel function. After that, the data are trained to search for the Optimum Separating Hyperplane (OSH) so that the distance from the hyperplane to the nearest positive and negative data point is maximized. Then, the hyperplane classifier is obtained by calculating the relative positions of the projection points of the two vectors on the hyperplane. Finally, the samples are classified accordingly.

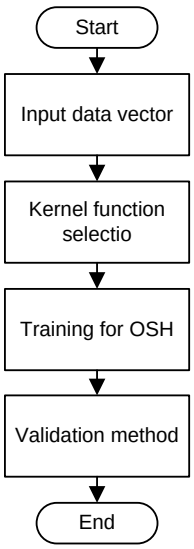

Fig. 5: Algorithm for support vector machine

From our literature study as follows, it is found that majority applied PCA prior to SVM algorithm [11, 18, 23-25] to reduce the dimension of the input vector, similar to LDA in as discussed in the previous section.

A total of 817 spectra were captured from three groups of colonic tissues specimen, i.e. 41 normal, 18 hyperplastic polyps and 46 adenocarcinomatous, using a 785 nm NIR Raman system. It was intended to classify them into normal, benign hyperplastic polyps and malignant adenocarcinomatous [23]. PCA retained 18 significant PCs cumulating to 91.4% of the original spectra for input to SVM classifier. This study implemented conventional SVM (c-SVM) and modified SVM (v-SVM) with the three kernel functions (11), (12) and (13) mentioned above. Using leave-one-out cross-validation method,

c-SVM with Gaussian RBF kernel function was found to achieve the highest diagnostic accuracy of 99.9%. The fluorescence background was removed, and the embedded noise was smoothed prior to normalization of the spectra for SVM classification.

Another application of PCA-SVM was found in the prognosis and diagnosis of castration-resistant prostate cancer (CRPC) [19]. PC cell lines and tissues of 50 patients diagnosed with androgen-dependent prostate cancer (ADPC) and CRPC were measured using Raman system with a laser wavelength of 632.8 nm. The raw spectra were pre-processed by a first order Savitsky-Golay smoothing filter and auto-fluorescence background subtraction algorithm before PCA. Then SVM classifier with RBF kernel optimized using leave-one-out cross-validation was used, which reported a classification performance of 85.7% for sensitivity and 88.9% for specificity.

PCA-SVM algorithm has also been used to analyze the Raman spectra of blood serum of esophageal cancer patients [24]. The spectra were obtained from 30 pathological confirmed and 31 healthy volunteers. Silver colloid substrates were mixed with the serum to enhance the intensity of Raman scattering. The spectra were pre-processed for smoothing and baseline removal. The performance of the PCA-SVM was compared with the conventional SVM (c-SVM). Highest accuracy attained by c-SVM with linear and RBF kernel was 72.1% and 83.6%, relative to 77% and 85.2% for PCA-SVM with the same kernel. PCA was observed to have improved the classification accuracy of SVM, besides reducing data for post-processing.

Of recent, PCA-SVM was used to discriminate between Raman spectra of normal, benign and cancerous breast tissues [10]. Spectra totaling at 491 were acquired from breast tissues taken from 15 patients. Two Raman systems with a different laser source, 532 nm, and 785 nm, were used in the study. From PCA, 16 significant components were chosen for classification by SVM. A performance of 98% of sensitivity and 97.9% of specificity was reported in discriminating cancerous tissues from normal and benign tissues. However, the paper did not mention the type of kernel used in their study.

In discriminating saliva samples between AIDS patients and healthy volunteers, SVM alone with Gaussian RBF kernel function was applied on the SERS spectra of these saliva samples [21]. Sensitivity and specificity of 95.6% and 100% respectively were reported.

A recent research from our team found that saliva samples adulterated with NS1 can be detected using SVM. Saliva samples adulterated with NS1 at different concentrations deposited onto substrate adsorbed with gold nanoparticles were analyzed using Raman spectroscopy. Even with NS1 at a low concentration of 10 ppm, the SVM classifier with RBF kernel attained accuracy, sensitivity and specificity of 81.5%, 79.1% and 84% respectively [39]. An improvement to the algorithm was introduced with PCA as the feature extraction technique and Linear kernel SVM as the classification technique for the spectra. The performance of classification was found increased, with 98.71% of accuracy, 98.97% of sensitivity and 98.44% of specificity using Cattel's Scree test as the criterion to select the significant principal components [40].

Diagnosis of parotid gland tumor was also attempted from Raman spectra. It was an ex-vivo study using the parotid tumor and normal tissues as samples, the accuracy achieved for classification of malignant and normal samples, using SVM with Gaussian radial basis (RBF) kernel, was 100%. The accuracy was lower at 98.3% for classification between benign and normal samples [40].

Considering a less invasive option than the above, using blood serum in place of tissues as samples, the same group repeated the study with the same classification algorithm [41]. Blood serum of 0.4 ml was mixed with 4 ml gold nanoparticles and incubated for 2 h at-4^{o}C prior to Raman analysis. The accuracy for classification between malignant and normal samples was 88.3%, while sensitivity and specificity were respectively 97.4% and 73.7%. The classification performance between benign and normal samples was slightly lower at the accuracy of 84.1%, the sensitivity of 90.8% and specificity of 74.3% [42].

Artificial neural network

Neurons of Artificial Neural Network (ANN) function as switches to receive inputs from other neurons. The status of neuron output is either 'activated' or 'inactive', depending on the sum of the multiplication of the inputs and weights feeding the neuron as illustrated in fig. 6. The weight by which the input is multiplied corresponds to the strength of the synapse [43]. ANN has been found successful in solving problems ranging from speech recognition, clustering, prediction system, pattern recognition and classification of diseases.

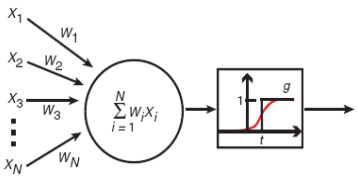

Fig. 6: Graphical representation of the McCulloch-Pitts model neuron (adapted from [43])

Fig. 7 shows a general procedural flow for ANN. First, the original spectral data undergo the pre-processing stage. This stage is used to suppress the background. For example, there may be superfluous features in the scattered data which need to be trimmed with feature selection technique. The features selected are then used as input to the ANN classifier, such as Back Propagation classifier, MLP classifier, which operates the following procedures, (i) Defining network architecture; (ii) Inferring the weight; (iii) Adapting hyper-parameters. Finally, the classifier interprets the result of classification of training patterns.

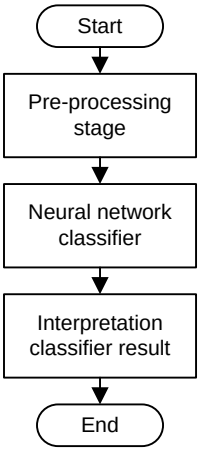

Fig. 7: Algorithm for artificial neural network

Harris et al. investigated the possibility to use ANN to discriminate between cancerous and normal cells from their Raman spectra [14]. Natural biological cells are a complex mixture of molecules (proteins, nucleic acids, lipids, and sugars) which produce Raman spectra with complex background. Hence, preliminary work was carried out with well-characterized cultured cells at standardized laboratory conditions and Raman spectrometer, first to understand the spectra produced. Then ANN was used to classify the cancerous cells from the normal cells, reporting specificity of 95% and sensitivity of 92%. Encouraged by the promising results, another study was conducted to discriminate between five different types of thyroid cell line [15], replacing ANN with Genetic Algorithm (GA). The discrimination sensitivity is found to decrease, between ranges of 61% to 91%, depending on the cell line type.

PCA-ANN and PCA-LDA algorithms were compared in their ability to discriminate Raman spectra of blood serum between

normal (n=31), liver cancer (n=27) and liver cirrhosis volunteers (n=23). Prior to the discrimination algorithm, the spectra were smoothed with the least square method. Using PCA, the spectra was reduced to only two principal components, which carry 97% of the total variance. The selected PCs were then used as inputs to the ANN and LDA algorithm. PCA-ANN was found to yield a higher performance than that of PCA-LDA, with sensitivity and specificity of 89% and 95%, as to the sensitivity of 88% and specificity of 79% [44].

ANN (BP algorithm) with inputs optimized by PCA was used to detect abnormal erythrocyte cell from Raman spectra of a single erythrocyte cell. ANN (BP algorithm) is the simplest form of ANN. It learns by minimizing the feedback error with an objective function. Based on a population of 11 patients with non-deletional HbH disease (HbH-CS), 11 thalassemias patients and 11 normal donors, the predictive accuracy was surprisingly as high as 97.9% [45].

PCA-ANN (MLP) was also applied to classification of skin lesion [4]. Raman spectra of five different types of skin lesions, i.e. basal cell carcinoma, malignant melanoma, normal skin, benign pigmented skin tumor and benign skin lesion, were obtained from skin samples in vitro on the skin surface of the punch biopsies or curetted lesions. Using multilayer perceptron (MLP) network, where the posterior probabilities of two-layer feedforward neural network are given in (8),

$$h_j(x) = \tanh\left(\sum_{i=1}^{I} \omega_{ji} x_i + \omega_{jo}\right). \quad (8)$$

where x_i are the inputs to the hidden layer weights, ω_{ji} is the input to hidden layer bias, and ω_{jo} is the output of the jth sigmoid activation function of the hidden layer. The network output of the output layer is given by (9),

$$y_k(x) = \sum_{j=1}^{H} \omega_{kj} h_i(x) + \omega_{ko} \ldots. \quad (9)$$

Where ω_{kj} are the hidden to output weights, ω_{ko} are the input to hidden biases and H is the number of units in the hidden layer. The classification rate reached 94.8%±2.7%.

Logistic regression analysis

Logistic regression analysis (LRA) is a multiple regression analysis for problems in which the outcome variable is categorical. LRA predicts the dichotomous (Yes/No) or binomial (1/0) outcome of response (dependent variable) using one or several predictors (independent variable). Hence, it is suitable for detection of anomalies in biomedical application.

Fig. 7 depicts steps in the LRA algorithm. It starts by deciding the choice of dependent and independent variables to insert into the general equation of LRA. Then, a logic equation is derived from testing the data, as expressed in (10-11),

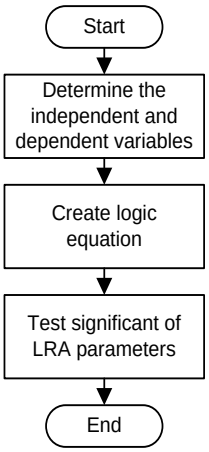

Fig. 8: Algorithm for logistic regression analysis

$$g(x) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \ldots\ldots. \quad (10)$$

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \ldots\ldots \quad (11)$$

Where, $\pi(x)$ is a predictor while $g(x)$ is the logic estimate value. $\beta_o, \beta_1 \ldots \beta_n$ are coefficient values for constant variables from Maximum Likelihood Estimation. The final step is to test the significance of LRA coefficients with Wald statistic as shown in (12).

$$W = \frac{\beta_i}{SE(\beta_i)} \ldots\ldots \quad (12)$$

Where β_i is the i^{th} coefficient value of the logistic regression while SE is the standard error for the i^{th}-coefficient value.

In a preliminary study, the SERS spectra of saliva samples acquired from 45 healthy and 19 lung cancer patients were compared [11]. Independent sample T-test was conducted on the spectra. Distinctive peaks have been identified as biomarkers for lung cancer. Discrimination with LRA achieved an accuracy of 96.9%.

CONCLUSION

SERS spectra of biological samples such as tissues, blood serum, blood plasma and saliva can be used to distinguish infected samples from normal samples. However, due to the complex characteristic of the spectra, it is essential first to process the signal to extract the significant features, at a molecular level, to represent the biomarker for disease detection. Theory, algorithmic procedure, and performance of signal processing techniques, PCA-LDA, PCA-ANN, PCA-SVM, SVM and LRA are examined in this paper. It is found that PCA-LDA is the most popular (45%), ensued by PCA-ANN (33%) and SVM (22%). PCA-SVM yields the highest in accuracy (99.9%), followed by PCA-ANN (98%) and LRA (97%). PCA-LDA and SVM score the highest, in terms of sensitivity-specificity. Application of these techniques on SERS spectra has shown encouraging performance, which could lead to novel promising screening or diagnostic procedure.

ACKNOWLEDGMENT

The authors wish to thank the Ministry of Higher Education (MOSTI), Malaysia for providing the research funding 100-RMI/SF 16/6/2 (14/2015); the Director of Non-Destructive Biomedical and Pharmaceutical Research Centre, Faculty of Pharmacy, Assoc Prof Dr. Wong Tin Mun and Head of Centre of NANO-Sci-Tech, Institute of Science, Prof Dr. Mohd Rusop, UiTM for kindly lending their equipment for our research project; the Research Management Institute and the Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia, for the support and assistance given to the authors in carrying out this research.

CONFLICT OF INTERESTS

Declared none

REFERENCES

1. Raman CV. A change of wave-length in light scattering. Nature 1928;121:619.
2. Fleischmann M, Hendra PJ, McQuillan AJ. Raman spectra of pyridine adsorbed at a silver electrode. Chem Phys Lett 1974;26:163-6.
3. Otto A. Surface enhanced raman spectroscopy (SERS). Surf Sci 1982;117:330.
4. Sigurdsson S, Philipsen PA, Hansen LK, Larsen J, Gniadecka M, Wulf HC. Detection of skin cancer by classification of Raman spectra. Biomed Eng IEEE Transactions 2004;51:1784-93.
5. Tsung-Heng T, Ting-Ting L, Yung-Ching H, Yu C, Tian-Jiun L, You-Hsuan L, *et al.* A multiscale approach for surface-enhanced Raman spectroscopy (SERS) spectrum representation and its application to bacterial discrimination. Proc Biomed Eng Inf China 2008;2:1247-52.
6. Vandenabeele P, Moens L. Introducing students to Raman spectroscopy. Anal Bioanal Chem 2006;385:209-11.
7. Movasaghi Z, Rehman S, Rehman IU. Raman spectroscopy of biological tissues. Appl Spectrosc Rev 2007;42:493-541.

8. Campion A, Kambhampati P. Surface-enhanced Raman scattering. Chem Soc Rev 1998;27:241-50.
9. Pichardo-Molina JL, Frausto-Reyes C, Barbosa-García O, Huerta-Franco R, González-Trujillo JL, Ramírez-Alvarado CA, *et al.* Raman spectroscopy and multivariate analysis of serum samples from breast cancer patients. Lasers Med Sci 2007;22:229-36.
10. Zhou L, Sun Y, Li J, Boydston-White S, Masilamani V, Zhu K, *et al.* Resonance Raman and Raman spectroscopy for breast cancer detection. Technol Cancer Res Treat 2013;12:371-82.
11. Yan W, Shuang S, Dian Q, Anyu C, Zijian C, Yulu Y, *et al.* Preliminary study on early detection technology of lung cancer based on surface-enhanced Raman spectroscopy. Proc Biomed Eng Inf China 2010;1:2081-4.
12. Li X, Yang T, Li S, Yu T. Surface-enhanced Raman spectroscopy differences of saliva between lung cancer patients and normal people. Proc SPIE-OSA Biomedical Optics SPIE; 2011. p. 808722–5.
13. Harris A, Rennie A, Waqar-Uddin H, Wheatley S, Ghosh S, Martin-Hirsch D, Fisher S, *et al.* Raman spectroscopy in head and neck Cancer. Head Neck Oncol 2010;2:1-6.
14. Harris A, Garg M, Yang X, Fisher S, Kirkham J, Smith D, *et al.* Raman spectroscopy and advanced mathematical modeling in the discrimination of human thyroid cell lines. Head Neck Oncol 2009;1:1-6.
15. Lin D, Feng S, Pan J, Chen Y, Lin J, Chen G, *et al.* Colorectal cancer detection by gold nanoparticle based surface-enhanced Raman spectroscopy of blood serum and statistical analysis. Opt Express 2011;19:13565-77.
16. Feng SY, Chen R, Lin J, Pan J, Chen G, Li Y, *et al.* Nasopharyngeal cancer detection based on blood plasma surface-enhanced Raman spectroscopy and multivariate analysis. Biosens Bioelectron 2010;25:2414-9.
17. Feng SY, Pan JJ, Wu YA, Lin D, Chen YP, Xi GQ, *et al.* Study on gastric cancer blood plasma-based on surface-enhanced Raman spectroscopy combined with multivariate analysis. Sci China: Life Sci 2011;54:828-34.
18. Duraipandian S, Zheng W, Ng J, Low JJH, Ilancheran A, Huang Z. *In vivo* diagnosis of cervical precancer using Raman spectroscopy and genetic algorithm techniques. Analyst 2011;136:4328-36.
19. Wang L, He D, Zeng J, Guan Z, Dang Q, Wang X, *et al.* Raman spectroscopy, a potential tool in diagnosis and prognosis of castration-resistant prostate cancer. J Biomed Opt 2013;18:87001-7.
20. Bispo JAM, de Sousa Vieira EE, Silveira JL, Fernandes AB. Correlating the amount of urea, creatinine, and glucose in urine from patients with diabetes mellitus and hypertension with the risk of developing renal lesions by means of Raman spectroscopy and principal component analysis. J Biomed Opt 2013;18:087004. Doi:10.1117/1.JBO.18.8.087004. [Article in Press]
21. Wang Y, Hua L, Liu J, Qu D, Chen A, Jiao Y, *et al.* Preliminary study on the quick detection of acquired immune deficiency syndrome by saliva analysis using surface enhanced Raman spectroscopic technique. Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009. Minneapolis, US: IEEE; 2009. p. 885–7.
22. Duraipandian S, Zheng W, Ng J, Low JJH, Ilancheran A, Huang Z. Near-infrared-excited confocal Raman spectroscopy advances *in vivo* diagnosis of cervical precancer. J Biomed Opt 2013;18:067007. Doi:10.1117/1.JBO.18.6.067007. [Article in Press]
23. Widjaja E, Zheng W, Huang Z. Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines. Int J Oncol 2008;32:653-62.
24. Li SX, Zeng QY, Li LF, Zhang YJ, Wan MM, Liu ZM, *et al.* Study of support vector machine and serum surface-enhanced Raman spectroscopy for noninvasive esophageal cancer detection. J Biomed Opt 2013;18:27008. Doi:10.1117/1.JBO.18.2.027008. [Article in Press].
25. Li X, Yang T, Lin J. Spectral analysis of human saliva for detection of lung cancer using surface-enhanced Raman spectroscopy. J Biomed Opt 2012;17:0370031. Doi:10.1117/1.JBO.17.3.037003. [Article in Press]
26. Kho KW, Malini O, Shen ZX, Soo KC. Surface enhanced Raman spectroscopic (SERS) study of saliva in the early detection of oral cancer. Proceeding of SPIE; Bellingham, WA: SPIE; 2005. p. 84–91.
27. Filik J, Stone N. Analysis of human tear fluid by Raman spectroscopy. Anal Chim Acta 2008;616:177-84.
28. Reyes-Goddard JM, Barr H, Stone N. Surface enhanced Raman scattering of herpes simplex virus in the tear film. Photodiagn Photodyn Ther 2008;5:42-9.
29. Jolliffe IT. Principal component analysis. 2nd ed. Encyclopedia of statistics in behavioral science. New York, US: Springer-Verlag; 2002.
30. Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas 1960;20:141–51.
31. Cattell RB. The scree tests for the number of factors. Multivariate Behav 1966;1:245–76.
32. Radzol ARM, Lee KY, Mansor W, Othman NH. Principal component analysis for detection of NS1 molecules from Raman spectra of saliva. Proceeding of 11th International Colloquium on Signal Processing and Its Applications; Kuala Lumpur, Malaysia: IEEE; 2015. p. 168–73.
33. Fisher RA. The use of multiple measurements in taxonomic problems. Annu Eugen 1936;7:179–88.
34. McLachlan GJ. Discriminant analysis and statistical pattern recognition. New Jersey: John Wiley; 2004.
35. Cervo S, Mansutti E, Del Mistro G, Spizzo R, Colombatti A, Steffan A, *et al.* SERS analysis of serum for detection of early and locally advanced breast cancer. Anal Bioanal Chem 2015;407:7503-9.
36. Twon Tawi FM, Lee KY, Mansor W, Radzol ARM. Automated detection of non-structural protein 1 in saliva from Raman spectrum with linear discriminant analysis. Aust J Basic Appl Sci 2014;8:27-32.
37. Muller DA, Young PR. The flavivirus NS1 protein: molecular and structural biology, immunology, role in pathogenesis and application as a diagnostic biomarker. Antiviral Res 2013;98:192-208.
38. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273-97.
39. Radzol ARM, Lee KY, Mansor W. Classification of salivary-based NS1 from Raman spectroscopy with support vector machine. Proceeding of 36th Annual International Conference of the IEEE Engineering in Medicice and Biology Society, Chicago, US: IEEE; 2014. p. 1835–8.
40. Radzol ARM, Lee KY, Mansor W. Model selection for PCA-linear SVM for automated detection of NS1 molecule from Raman spectra of the salivary mixture. Proceeding of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy: IEEE; 2015. p. 2824–7.
41. Yan B, Wen Z, Li Y, Li L, Xue L. An intraoperative diagnosis of parotid gland tumors using Raman spectroscopy and support vector machine. Laser Phys 2014;24. Available from: http://dspace.xmu.edu.cn/handle/2288/93743. [Last accessed on 10 Dec 2016].
42. Yan B, Li B, Wen Z, Luo X, Xue L, Li L. Label-free blood serum detection by using surface-enhanced Raman spectroscopy and support vector machine for the preoperative diagnosis of parotid gland tumors. BMC Cancer 2015;15:1–9.
43. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115-33.
44. Yang T, Li X, Yu T, Sun R, Li S. Spectral discrimination of serum from liver cancer and liver cirrhosis using Raman spectroscopy. Proceeding of SPIE-Clinical and Biomedical Spectroscopy and Imaging II; SPIE; 2011. p. 808720.
45. Chen X, Wang G, Tao Z, Liu J, Yao H, Huang S, *et al.* Raman spectral discrimination of thalassemia erythrocytes based on PCA arithmetic and BP network model. Zhongguo Jiguang/Chin J Lasers 2009;36:2448-54.