Classification of Salivary based NS1 from Raman Spectroscopy with Support Vector Machine

A. R. M. Radzol.-*IEEE Member,* Khuan Y. Lee-*Sr. IEEE Member,* W. Mansor-*IEEE Member,*

Abstract— **Non-Structural Protein 1 (NS1) antigen has been recognized as a biomarker for diagnosis of flavivirus viral infections at early stage. Surface Enhanced Raman Spectroscopy (SERS) is an optical technique capable of detecting up to a single molecule. Our previous work has established the Raman fingerprint of NS1 with gold as substrate. Our current study aims to classify NS1 infected saliva samples from healthy samples, a first ever attempt. Saliva samples from healthy subjects, NS1 protein and NS1-saliva mixture samples were analyzed using SERS. The SERS spectra were then pre-processed prior to classification with support vector machine (SVM). NS1-saliva mixture at concentration of 10ppm, 50ppm and 100ppm were examined. Performance of SVM classifier with linear, polynomial and radial basis function (RBF) kernels were compared, in term of accuracy, sensitivity, and specificity. From the results, it can be concluded that SVM classifier is able to classify the samples into NS1 infected samples and normal saliva samples. Of the three kernels, performance in using polynomial and RBF kernel is found surpassing the linear kernel. The best performance is attained with RBF kernel with accuracy of [97.1% 93.4% 81.5%] for 100ppm, 50ppm and 10ppm respectively.**

I. INTRODUCTION

Non-structural protein 1 (NS1) is one of the eight non-structural proteins encoded by the genome of Flavivirus genus. Flavivirus belongs to the Flaviviridae family that includes dengue virus (DENV), West Nile Encephalitis virus (WNEV), Yellow fever virus (YFV), Japanese Encephalitis virus (JEV) etc that can brings about diseases such as encephalitis and hemorrhagic fevers with serious consequences, or even mortality [1-2] when medical attention is lacking or is not in time. NS1 was first reported in 1970 as a viral antigen in the sera of dengue-infected patients [1]. Even so, its structure and mechanistic function remain elusive thus far. It is found retained within the infected cells normally. However, it also localizes to the cell surface, secreted slowly from mammalian cells. The secreted and cell-surface-associated NS1 are highly immunogenic and both the proteins themselves and the antibodies they elicit have been implicated in the seemingly contradictory roles of protection and pathogenesis in the infected host [2]. NS1 is discovered to have an important yet unclear role in RNA replication [2]. Hence, it is accepted as biomarker for early detection of the related diseases in recent years[3, 4].

Raman spectroscopy is a technique used to study the vibration and rotational modes of molecule. It produces unique spectrum for each and every molecule from inelastic scattering of light[5]. Inelastic also known as Raman scattering are weak signals that impede practical application of Raman spectroscopy. Integration of Raman spectroscopy and nano-technology result in an enhanced technique of Raman spectroscopy, SERS. SERS enhances significantly Raman signal from molecule that have been absorbed with nano-size noble metal known substrate. Substrates are selective and sensitive. Their presence effects magnetic field of higher strength, which in turn increases Raman scattering and attenuates its intensity. The increase in intensity of Raman signal can be from 10^4 to 10^6 for regular cases, or as high as 10^8 and 10^{14} in special cases, with the right choice of substrate[6, 7]. Gold, silver and copper are substrates in common use. SERS has been demonstrated to be capable of detecting biological molecules such as proteins[8], DNA[9, 10], RNA[11], even the entire pathogen[12, 13]. This explains why it has been earmarked as a potential technique for early disease detection. In protein studies, Raman spectroscopy is found capable to identify secondary structure of protein by associating peaks to vibration mode of molecules and bonding of protein. Our previous work has been the first to establish Raman fingerprint of NS1 molecule[14].

SVM are supervised learning, non-probabilistic models with associated learning algorithms that analyze data and recognize patterns, for purpose of classification and regression analysis. With each training data tagged to one of the two classes, the SVM training algorithm predicts a model to represent training data as points in space. The data are mapped so that the two different classes are divided by a gap as wide as possible. New data are then mapped into that same space and predicted to one of the classes, based on which side of the gap they fall on. SVM algorithm incorporated with Principle Component Analysis (PCA) has been used to classify the colonic tissues into normal, benign hyperplastic polyps and malignant adeno-carcinomatous, based on their Raman spectra[15]. The study implements conventional SVM (C-SVM) and modified SVM (v-SVM) with the three kernel functions mentioned. Using leave-one-

This work was supported by Ministry of Education, Malaysia (600-RMI/ERGS 5/3/ (20/2013) and Research Management Institute (RMI) of Universiti Teknologi MARA (UiTM),

A. R. M. Radzol is with Faculty of Electrical Engineering, Universiti Teknologi MARA.(phone: 6013-4802620; fax: 603-55435077; e-mail: nikrozan@yahoo.com).

Khuan Y. Lee., is with Faculty of Electrical Engineering, Brain & Neuroscience Community of Research, Universiti Teknologi MARA, Shah Alam, 40450 Selangor Malaysia(phone: 603-55436088, e-mail: leeyootkhuan@salam.uitm.edu.my).

W. Mansor is with Faculty of Electrical Engineering Brain & Neuroscience Community of Research, Universiti Teknologi MARA, Shah Alam, 40450 Selangor Malaysia, (e-mail: wahidah231@salam.uitm.edu.my)

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

out cross-validation method, C-SVM with Gaussian RBF kernel function is found to achieve the highest diagnostic accuracy of 99.9%[15]. Another application of SVM on SERS spectra has been reported for discrimination of saliva samples between AIDS patients and healthy volunteers[16]. SVM with Gaussian RBF kernel function developed with R-software is used with reported sensitivity and specificity of 95.6% and 100%. However, SVM has not been applied to classify NS1 infected saliva samples from control saliva samples.

Our work here is the first attempt to investigate the feasibility to identify NS1 features from Raman spectra of control and disease saliva samples using supervised learning classifier; SVM. Section II elucidates theory on the classification technique. Section III looks into sample preparation and sample analysis with SERS. Section IV discusses on pre-processing procedures on raw Raman spectra, which was not found with previous studies and classification performance of SVM with linear, polynomial and RBF.

II. Support Vector Machine

SVM is a classification algorithm based on supervised learning method[17]. For a given sample of data, SVM is trained to maximize a particular mathematical function known as kernel function which separates the *d*-dimensional data perfectly into its two classes. The separating plane is known as optimal hyperplane which is established when distance to the closest negative example is equal to distance to the closest positive example. A kernel function is used to map the original data from input space to feature space to allow the hyperplane classifier to separate the input data into different classes with minimal error. Figure 1 briefly depicts the concept of SVM.

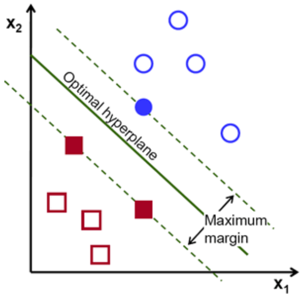

Figure 1: SVM algorithm create hyperplane which separate two classes of data

Set of vectors is said to be optimally separated by the hyperplane if it is separated without error and the distance between the closest vectors to the hyperplane is maximal.

$$w^T.x_i + b = 0 \qquad (1)$$

As shown in (1), x is test data in a separating hyperplane. The value of w and b are selected with maximum margin. The x-data can then be separated by minimizing the distance, $\|w\|^2$. Thus, the maximal distance between the closest vectors to the hyperplane can be obtained using (2) and (3) in the following.

$$(w^T.x_i) + b \gg 1 \text{ if } y_i = 1 \qquad (2)$$

$$(w^T.x_i) + b \ll -1 \text{ if } y_i = -1 \qquad (3)$$

Here, training vectors are $x_{i,}= 1, \ldots,n$ and this is a case of separating between two different classes of data. This can be represented as a soft classifier which linearly interpolates the margin as in (4). Say, for the vector *f(x)* in (4), the data is classified as '1' if x_i falls into class 1, while it is classified as '-1' if x_i falls into class 2.

$$f(x) = \begin{cases} 1, & (w^T.x_i) + b \gg 1 \\ -1, & (w^T.x_i) + b \ll -1 \end{cases} \qquad (4)$$

All the support vectors lie on the margin when the data is linearly separable. The number of support vectors preserved can thus be very small. Hence, it can be said that the hyperplane is determined by a small subset of the training set.

The kernel, represents a valid inner product in feature space. The training set is not linearly separable in an input space but it is linearly separable in the feature space. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Kernel Functions such as Linear Kernel, Polynomial Kernel and Radial Basis Function Kernel have been adopted in our work to transform non-linear classification to linear classification. The following equations show the mathematical representation of these kernel functions.

$$\text{Linear:} \quad K(x_i, x_j) = x_i.x_j + 1 \qquad (5)$$

$$\text{Polynomial:} \quad K(x_i, x_j) = (x_i.x_j + 1)^{\mathrm{p}} \qquad (6)$$

$$\text{Gaussian RBF:} \; K(x_i, x_j) = e^{\frac{\|x_j - x_i\|^2}{2\sigma^2}} \qquad (7)$$

III. Methodology

A. Samples Preparation

10 saliva samples were collected from 10 healthy volunteers aged 23-34 years old. All samples were collected in the morning (9-11am), abiding a published protocol [18]. The subject was also advised to refrain from taking foods, drink (except plain water) and brushing teeth at least 1 hour prior to saliva collection. About 10 minutes prior to the collection, subject was asked to gargle thoroughly for 1 minute. Thereafter, whole saliva of approximately 3ml was collected using unstimulated collection procedure. During the collection, ice was used to maintain the low temperature. The sample was then transferred into Eppendorf tubes and

was centrifuged at 14000rpm for 10 minutes to extract the supernatant.

NS1 glycoprotein protein (ab64456) purchased from Abcam was used in this study. It is a recombinant full length protein expressed in E. coli. Initial concentration of the protein is at 1mg/ml, equivalent to 1000ppm. To prepare NS1-saliva mixture at different concentration, the protein was diluted into 100ppm, 50ppm and 10ppm using de-ionized (DI) water as solvent.

30μL of diluted NS1 at all concentrations were mixed with 30μL saliva supernatant. To ensure homogeneity, the mixtures were centrifuged at 5000rpm for 1 minute. Then 10μL of every mixture was deposited onto the gold coated slide substrate and left dried before Raman analysis. A total of 15 mixtures at each selected concentration were prepared and analyzed.

B. *Raman Spectra of NS1-saliva Mixture*

The dried samples were analyzed by PeakSeeker Pro from Agiltron with near infrared excitation source of 785nm laser For all the spectra, the equipment is set to 300mW of power and exposure time of 10 seconds. The spectral region was recorded from 200cm^{-1} to 2000cm^{-1} with interval of 1cm^{-1}. The microscope objective lens was set to 50X with working distance of approximately 1.9mm. A total of useful150 spectra were obtained from 15 samples. Raman analysis was repeated to ensure reproducibility of spectra. For classification purpose, 25 spectra at each concentration were used.

IV. RESULT AND DISCUSSION

A. *Raman Spectra of Saliva and NS1-Saliva Mixture*

Figure 2 shows the average of 10 raw spectra of blank gold coated slide substrate with and without saliva. Only a slight difference between the two spectra can be observed.

From repeated measurement across a gold coated slide, the substrate exhibit similar features with variation in intensity. The standard deviation is ranging from 25.1 to 94.1 throughout the spectral region. The variation is due to the uneven size of gold nano-particles coating the slide as reported in our previous work[19].

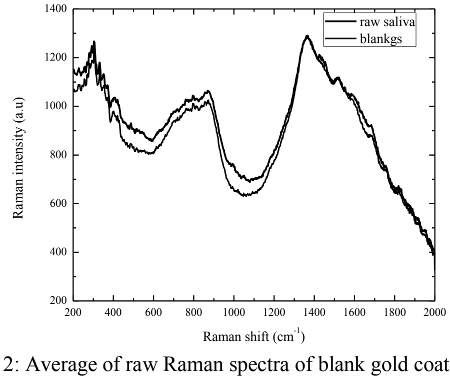

Figure 2: Average of raw Raman spectra of blank gold coated slide and saliva

As for saliva, no additional peak is observed in the spectral region. This is expected since 90% of saliva component is water. Water is well known as a weak Raman scatterer and do not produce any significant peak in the *fingerprint region* [20]. However, the standard deviation value of saliva is ranging from 36.3 to 255.3 exhibiting larger variation than the substrate.

Figure 2 shows the average spectra of samples; (a) saliva, (b) NS1-saliva mixture at 10ppm, (c) NS1-saliva mixture at 50ppm, (d) NS1-saliva mixture at 100ppm. At 100ppm and 50ppm, characteristic peak of NS1 at around 1000cm^{-1} is exhibiting appreciably high intensity. The peak is attributed from the breathing vibration of phenylalanine benzene ring as reported in our previous work[14]. As the concentration of mixture is decreased to 10ppm, the intensity of the peak is reduced as shown in Figure 3(b).

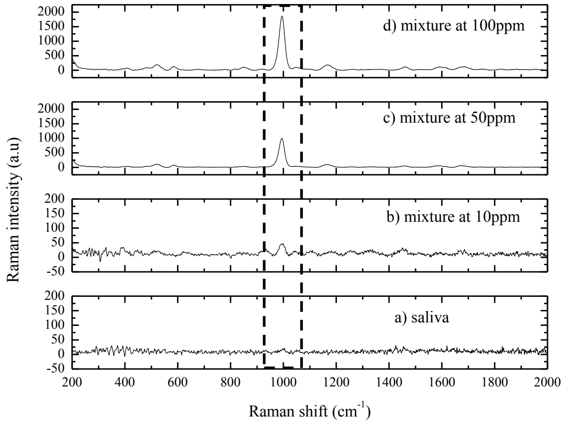

Figure 3: Raman spectra of NS1-saliva mixture at (a)100ppm, (b) 50ppm, (c)10ppm and (d) saliva

B. *SVM Classification of NS1*

For classification, region of interest (ROI), where spectral features of NS1 fingerprint reside, i.e. from 980cm^{-1} to 1020cm^{-1} forming a feature vector of [41x25] is used as the classifier inputs. Data are divided into training and test set in the ratio of 60:40.

Table 1 shows the accuracy attained by the SVM classifier. At a concentration of 100ppm, the classification accuracy achieved is higher than 97% for all kernels, except for linear kernel. Similar finding can be said of the sensitivity as shown in Table 2, where 94% or more of the samples detected with NS1 by SVM, except for linear kernel, are actual positives. In the case of specificity, as shown in Table 3, all the kernels of SVM are found able to not classifying samples without NS1 as samples with NS1. This encouraging performance is attributed to the characteristic peaks at 100ppm of concentration having an intensity level that makes it distinct from samples with only saliva, owing to their high concentration.

Performance of the SVM classifier at 50ppm follows similar trend, only with lower classification accuracy, sensitivity and specificity. The sensitivity performance is almost similar to 100ppm but the specificity performance for this concentration is slightly reduced indicating some misclassification of saliva data as mixture data.

At 10ppm, performance of the SVM classifier is further reduced, with accuracy, sensitivity and specificity less than 90% for all the kernel functions. As concentration of NS1 in the samples lessens, intensity of the characteristic peak from Raman scattering is also reduced, due to scarcity of the molecules. This makes detection of characteristic peak challenging. This is investigated further by displaying the pre-processed spectral data of saliva alone (red '*') and mixture of saliva with NS1 (green 'o') within the region of interest, as in Figure 3. As observed, the intensity of the characteristic peaks mingled with the spectral envelope of saliva.

Table 1: Accuracy of SVM classifier for every concentration

Kernel	10ppm	50ppm	100ppm
Linear	76.0	85.9	93.7
Polynomial	80.8	92.3	97.9
RBF	81.5	93.4	97.1

Table 2: Sensitivity of SVM classifier for every concentration

Kernel	10ppm	50ppm	100ppm
Linear	81.2	86.8	87.5
Polynomial	78.1	95.1	95.8
RBF	79.1	95.1	94.1

Table 3: Specificity of SVM classifier for every concentration

Kernel	10ppm	50ppm	100ppm
Linear	70.7	85.0	100
Polynomial	83.6	89.6	100
RBF	84.0	91.9	100

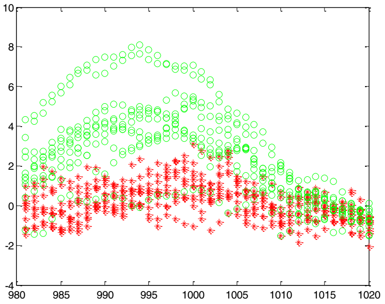

Figure 4: ROI of NS1-saliva mixture at 10ppm (green), saliva (red)

V. CONCLUSION

In this work, gold coated slide is used as substrate to analyze saliva samples of healthy volunteers and NS1-saliva mixture samples. For efficiency in classification, the region of interest for NS1-saliva Raman spectra is optimised to an enclosure about the NS1 fingerprint, i.e. from 980 to 1020cm^{-1}. It is found that the SVM classifier is able to detect the presence of NS1 with encouraging accuracy, sensitivity and specificity. Of the three kernels that are used, performance in using polynomial and RBF hyperplane to differentiate between the classes is illustrated to supersede the linear hyperplane. In addition, it is also found that the performance of the SVM classifier dependent on the intensity of the characteristic peak, attributed to the concentration of NS1. Where the best performance is attained with RBF kernel with accuracy of [97.1% 93.4% 81.5%] for 100ppm, 50ppm and 10ppm respectively.

REFERENCES

[1] P. R. Y. David A. Muller, "The flavivirus NS1 protein:Molecular and structural biology,immunology, role inpathogenesis and application asadiagnostic biomarker," 2013.

[2] H.-J. u. T. Brett D. Lindenbach, Charles M. Rice, "Flaviviridae: The Viruses and Their Replication," 2007.

[3] S. Alcon, *et al.*, "Enzyme-linked immunosorbent assay specific to dengue virus type 1 nonstructural protein NS1 reveals circulation of the antigen in the blood during the acute phase of disease in patients experiencing primary or secondary infections," *Journal of Clinical Microbiology,* vol. 40, pp. 376-381, 2002.

[4] S. Datta and C. Wattal, "Dengue NS1 antigen detection: A useful tool in early diagnosis of dengue virus infection," *Indian Journal of Medical Microbiology,* vol. 28, pp. 107-110, 2010.

[5] C. V. Raman, "A change of wave-length in light scattering [8]," *Nature,* vol. 121, p. 619, 1928.

[6] L. A. Dick, *et al.*, "Metal Film over Nanosphere (MFON) Electrodes for Surface-Enhanced Raman Spectroscopy (SERS): Improvements in Surface Nanostructure Stability and Suppression of Irreversible Loss," *The Journal of Physical Chemistry B,* vol. 106, pp. 853-860, 2001.

[7] M. Fan, *et al.*, "A review on the fabrication of substrates for surface enhanced Raman spectroscopy and their applications in analytical chemistry," *Analytica Chimica Acta,* vol. 693, pp. 7-25, 2011.

[8] Z. A. Combs, *et al.*, "Label-free raman mapping of surface distribution of protein A and IgG biomolecules," *Langmuir,* vol. 27, pp. 3198-3205, 2011.

[9] L. R. Allain and T. Vo-Dinh, "Surface-enhanced Raman scattering detection of the breast cancer susceptibility gene BRCA1 using a silver-coated microarray platform," *Analytica Chimica Acta,* vol. 469, pp. 149-154, 2002.

[10] C. Fang, *et al.*, "DNA detection using nanostructured SERS substrates with Rhodamine B as Raman label," *Biosensors and Bioelectronics,* vol. 24, pp. 216-221, 2008.

[11] J. D. Driskell, *et al.*, "Rapid microRNA (miRNA) detection and classification via surface-enhanced Raman spectroscopy (SERS)," *Biosensors and Bioelectronics,* vol. 24, pp. 917-922, 2008.

[12] R. M. Jarvis and R. Goodacre, "Discrimination of Bacteria Using Surface-Enhanced Raman Spectroscopy," *Analytical Chemistry,* vol. 76, pp. 40-47, 2003.

[13] S. Shanmukh, *et al.*, "Rapid and Sensitive Detection of Respiratory Virus Molecular Signatures Using a Silver Nanorod Array SERS Substrate," *Nano Letters,* vol. 6, pp. 2630-2636, 2006.

[14] A. R. M. Radzol, *et al.*, "Raman molecular fingerprint of non-structural protein 1 in phosphate buffer saline with gold substrate," 2013, pp. 1438-1441.

[15] E. Widjaja, *et al.*, "Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines," *International Journal of Oncology,* vol. 32, pp. 653-662, 2008.

[16] Y. Wang, *et al.*, "Preliminary study on the quick detection of acquired immure deficiency syndrome by saliva analysis using surface enhanced Raman spectroscopic technique," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, 2009, pp. 885-887.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[18] M. Navazesh, "Methods for collecting saliva," *Annals of the New York Academy of Sciences,* vol. 694, pp. 72-77, 1993.

[19] A. R. M. Radzol, *et al.*, "Nano-Scale Characterization of Surface Enhanced Raman Spectroscopic Substrates," *Procedia Engineering,* vol. 41, pp. 867-873, 2012.

[20] D. Catalina, "Raman Spectroscopy Applied to Biomolecule Characterization," in *Nanoantenna*, ed: Pan Stanford Publishing, 2013, pp. 1-34.