PCA-Polynomial-ELM Model Optimal for Detection of NS1 Adulterated Salivary SERS Spectra

Nur Hainani Othman¹, Khuan Yoot Lee¹,² *, Afaf Rozan Mohd Radzol²,³, Wahidah Mansor¹,², Nur Amanina Yusoff ¹

¹Faculty of Electrical Engineering, Universiti Teknologi MARA, Shah Alam, Selangor DE, Malaysia
²Computational Intelligence Detection RIG, Pharmaceutical & Lifesciences Communities of Research, Universiti Teknologi MARA, Shah Alam, Selangor DE, Malaysia
³Faculty of Electrical Engineering, Universiti Teknologi MARA, Cawangan Pulau Pinang, Permatang Pauh, Pulau Pinang, Malaysia

*leeyootkhuan@uitm.edu.my

Abstract. Of recent, there has been much interest in the application of Surface Enhance Raman Spectroscopy (SERS) analysis in the detection of diseases such as dengue. Early diagnosis of dengue affords early intervention, greater chance of cure and prevention of mild dengue progressing into life threatening stage. SERS produces, on the interaction of photons from laser beam with saliva samples, a spectral image of its composition here. In the case of dengue fever, Non–Structural Protein 1 (NS1), being its biomarker, is the biochemical fingerprint to be revealed by SERS. NS1 presents in body fluid such as blood and saliva of patients since day one of infection, that makes NS1 a favourite alternative to antibody types of biomarker. However, the concentration of NS1 in saliva is low, yielding a low intensity SERS spectrum. In addition, the spectrum is usually interfered with undesirable noisy features. Extreme Learning Machine (ELM) is a fast algorithm with its strength in data pattern generalization. It has been applied in pattern recognition and machine learning for classification and regression, with encouraging performance. Our work here intends to determine an optimal polynomial-ELM model in classifying SERS spectra of saliva samples adulterated with NS1, amongst the different models subject to three different termination criteria of Principal Component Analysis (PCA). Performance of ‘100%’ is attained for accuracy, sensitivity, specificity and precision, while ‘1’ for kappa, by combining the cumulative percent of total variance (CPV) termination criterion and polynomial-ELM model of power 2 and constant 0.5.

1. Introduction
Dengue fever (DF) is a viral disease spread by female Aedes mosquitoes. When a mosquito bites a person carrying dengue virus in their blood, the mosquito becomes infected and transmits the virus to another person. Dengue virus is one of the flavivirus genome, a member of the Flaviviridae family [1]. It has three structural proteins and seven non-structural proteins, in which NS1 is one of the seven. NS1 plays an important role in viral replication and is essential to the existence of the virus [2]. The advantage of using antigen such as NS1 is that, detection at early stage of DF, without having to wait for the antibodies to form, is possible. DF is the most common mosquito-borne disease and a major health concern in tropic and subtropical regions. According to World Health Organization, more than 40% of
the world’s population, representing 2.5 billion people, are at risk of dengue infection [3]. There are several methods to detect NS1 such as Enzyme-Linked Immunosorbent Assay (ELISA), Reverse Transcription Polymerase Chain Reaction (RT-PCR), with blood as the diagnostic medium. As such, they are prone to blood-related infections [4]. Detection of NS1 is one of the many techniques to diagnose dengue virus, such as virus isolation, detection of Ribonucleic acid (RNA), immunohistochemistry, Immunoglobulin M (IgM), Immunoglobulin G (IgG) [5].

Raman spectroscopy (RS) yields information on atomic vibration, which allows test recognizable proof and quantization based on inelastic scattering of monochromatic light [6]. Signals produced by RS are so weak that limit its application till the discovery of SERS. SERS provides greatly enhanced RS signals from Raman active analyte that have been adsorbed onto nano-metal surfaces [7]. The intensity of RS signals has been observed to increase by $10^4$ at the least, to as high as $10^{14}$ [6][8]. The electromagnetic and chemical mechanism contribute to two-third of the enhancement in intensity [7]. The electromagnetic enhancement owes to the roughness of nano-metal surface, while the chemical enhancement is a cause effect of adsorbate electronic states as a result of chemisorption of the analyte [7]. The discovery of SERS encourages RS to be broadly acknowledged as potential diagnostic method for a wide-ranging applications such as, detection of viruses in sputum [9], diabetes in serum [10], melamine in milk [11] and gastric cancer in blood plasma [12]. Applications of RS for detection of DF from NS1 have been reported, yet using blood serum as the diagnostic medium [13][14].

In this paper, the application of SERS for biochemical analysis of saliva samples for detection and classification of NS1-Df is explored. PCA and polynomial-ELM are used to extract features and identify salivary SERS spectra adulterated with NS1. To the best of our knowledge, besides our research group, this is the first report on SERS analysis of saliva samples to reveal NS1-Df for purpose of detection and classification. Section 2 presents theory on PCA and polynomial-ELM classifier. Section 3 elaborates on the NS1 adulterated dataset and methodology employed in this study. Section 4 discusses results from classification of NS1 salivary SERS spectra with different PCA termination criteria and polynomial-ELM vector parameters.

2. Theoretical Background

2.1. Principal Component Analysis, PCA
PCA is probably the oldest multivariate analysis technique introduced by Pearson (1901) [15]. This technique is popularly in use after the introduction of electronic computer. Currently, its algorithm resides in almost every statistical computer package [15].

Being a data reduction technique, PCA reduces the dimensionality of the variables in a large dataset, yet maintains the important features as much as possible [15][16]. In other words, with the use of mathematical procedure, PCA reduces high dimensional data to low dimensional data by transforming the correlated variables into a smaller number of uncorrelated variables, known as principal components (PCs) [17]. Eigenvalue of dataset is a measure of variability in the sample distribution that reflects the variance. PCs are ranked according to variance of datasets. The first place is occupied by PC with the largest variance, while the last place by PC with the least variance. The formula to measure the variability for average squared deviation of each sample from its mean is as follows [16],

$$s^2 = \frac{\Sigma (\text{sample} - \text{mean})^2}{\text{Total sample} - 1}$$ (1)

To facilitate data analysis and interpretation of the original data, only significant PCs are selected after ranking using PCA termination criteria. The three termination criteria considered in this study are Eigenvalue-One Criterion (EOC), Cattell’s Scree test and Cumulative Percent of Variance (CPV) [18].

2.2. Extreme Learning Machine, ELM
ELM is first developed by Huang in 2004 and published in 2006. It consists of a three layer neural network with sigmoid activation function [19]. It is suitable for implementation on a single-hidden layer feedforward neural networks (SLFNs) [20]. ELM is a learning algorithm well-known for high learning
speed, good pattern generalization, least of human intervention and ease of implementation. It not only delivers smaller training error but also better performance.[19][20]. The advantages of ELM to conventional gradient based learning methods are, (i) Stopping criteria, learning rates, learning epochs and local minima are auto-optimized; (ii) Learning algorithm is fast, adopting one pass without reiteration. The learning phase typically completes in seconds or less. It is also suitable for all activation functions [19][20].

The polynomial kernel is a non-stationary kernel. It is well suited for problems where all the training data are normalized. Polynomial is commonly defined in mathematical form as follow [20]:

$$k(u,v) = (u.v + n)^{power}$$

where $n$ is constant and power is the power for the function, while $u$ and $v$ are vectors in the input space computed from training or test samples. The polynomial power starts from two, according to the polynomial function principle. The polynomial constant and polynomial power are varied to obtain optimized vector parameters.

3. Methodology

Figure 1 illustrates steps in optimizing the selection of PCA termination criteria [Scree, CPV, EOC] and parameters of polynomial-ELM classifier [constant, power], for use with SERS spectra of NS1 adulterated saliva.

The algorithm consists of two parts: feature extraction by PCA and feature classification by polynomial-ELM. SERS spectra of 64 samples each from the control group and the NS1 adulterated group, are acquired from the UiTM-NMRR-12868-NS1-DENV database. The control group data are contributed by healthy volunteers between 23-24 years old, while the NS1 adulterated data are acquired by mixing the control group saliva with pure NS1 protein at different concentrations [2]. Prior to feature extraction and feature classification, the spectra are pre-processed to eliminate spurious noise, background noise, fluorescent effect, signal drifts and spikes from cosmic ray [6][21].

PCA with different termination criteria, i.e. Scree test [128 x 5], CPV [128 x 70] and EOC [128 x 115] are served as inputs to the polynomial-ELM classifiers. Then, polynomial constant $n$ and polynomial power, as defined in Equation (2), are varied and their effects recorded. In the first case, the polynomial constant $n$ is varied, as the polynomial power is kept constant; vice versa in the second case. The range of variation for polynomial constant $n$ is from 0.000001 to 10000, while that for polynomial power is from 2 to 50. For both cases, performance of the classifier models is analysed using indicators such as accuracy, precision, sensitivity, specificity and kappa value. Based on this, the best classifier model is determined.

4. Result and Discussion

This section presents results from classification of NS1 adulterated salivary SERS spectra by using PCA termination criteria integrated with polynomial kernel of ELM classifier. Altogether 3381 polynomial-ELM classifier models, of 69 different polynomial $n$-values and 49 polynomial power-values, are assessed.

With reference to Figure 2, Figure 3 and Figure 4, the best performance in [Accuracy, Precision, Kappa] is [97.37%, 94.74%, 0.95], [100%, 100%, 1.0] and [100%, 100%, 1.0] for Scree-, CPV- and EOC-Polynomial-ELM classifier model with [polynomial constant, polynomial power] of [0.000001, 9], [0.5, 2] and [0.5, 2] respectively. Figure 5 shows the ROC graphs for these best performed classifiers. The nearest point with the straight line indicates the best performance of the ROC graph. CPV and EOC-polynomial-ELM classifier models are observed with the best sensitivity and specificity, which lies at the upper left corner of the ROC graph, with coordinate [0,1].
As observed from Figure 2, Figure 3 and Figure 4, trending in the accuracy performance is tracked by kappa and precision. When accuracy trends up, the kappa value and precision increases, and conversely. This is because the kappa algorithm is dependent on accuracy. The kappa algorithm compares observational probability of agreement, which is the classifier accuracy, to the hypothetical expected probability of agreement under an appropriate set of baseline constraints such as, total independence of observer classification, which is also known as chance agreement [22]. Theoretically, the precision algorithm does not depend on accuracy. The classifier can be precise but not accurate, and vice versa. However, in this study, the classifier is found precise when it is accurate; imprecise when it is inaccurate, which just shows that the classifier is reliable.
Figure 4. Optimal Polynomial-ELM classifier performance for EOC termination criteria

Figure 5. ROC of Polynomial-ELM classifier for different PCA termination criteria

5. Conclusion

Here, the effect of PCA termination criteria and polynomial-ELM parameters in identifying NS1 from SERS spectra of saliva are investigated. As far as our knowledge, research from our group has been the first to investigate the presence of salivary NS1 with SERS, which offers favorable circumstances such as, early, non-invasive and simple sample collection and sample preparation. This is intended as an alternative technique in distinguishing NS1 from saliva, which will profit diagnosis of NS1-DF related infections in the future. For our study here, of the PCA termination criteria, the CPV and EOC criteria which transform the original signal into 70 PCs and 115 PCs respectively, reported the best performance with ‘1’ for kappa; ‘100%’ for accuracy, precision, sensitivity and specificity and sensitivity-specificity pair at coordinate [0,1] of the ROC graph. With further consideration to the number of PCs, CPV uses less PCs in comparison to EOC, which could reduce computational time and load. Even though both CPV and EOC criteria yield similar performance, CPV integrated with polynomial-ELM classifier of polynomial constant 0.5 and power 2 is reckoned with the most optimized classification performance.

Acknowledgments

The author would like to thank the Research Management Institute, for providing the research funding 600-IRMI/REI 5/3 (02 1/2018); Faculty of Electrical Engineering, Universiti Teknologi MARA, Selangor, Malaysia, for the support and assistance given to the authors in carrying out this research.

References

[1] Chen H, Lai Y and Yeh T 2018 Dengue virus non-structural protein 1: a pathogenic factor, therapeutic target and vaccine candidate J. Biomed. Sci. 25 1–11

[2] Radzol A R M, Lee K Y and Mansor W 2014 Classification of salivary based NS1 from Raman Spectroscopy with support vector machine Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Chicago, IL, USA) p 1835–1838

[3] Bhatt S et al 2013 The global distribution and burden of dengue Nature 496, 504–507

[4] Radzol A R M, Lee K Y and Mansor W 2013 Nonstructural protein 1 characteristic peak from NS1-saliva mixture with Surface-Enhanced Raman spectroscopy Annu. Int. Conf. IEEE Eng. Med.
Hussain A, Rehman S U, Aslam S, Javed N and Abbas Z 2014 Nano-gold particles mediated detection of NS1; an early diagnostic marker of dengue virus infection J. Anim. Plant Sci. 24 1110–1115

Radzol A R M, Lee K Y, Mansor W and Azmin A 2015 Optimization of Savitzky-Golay smoothing filter for salivary surface enhanced Raman spectra of non structural protein 1 Proc. IEEE Reg. 10 Annu. Int. Conf. (Bangkok, Thailand) p 9–14

Radzol A R M, Lee K Y, Mansor W and Yahaya S R 2012 Surface-enhanced raman spectral analysis of substrates for salivary based disease detection Proc. - 2012 IEEE 8th Int. Colloq. Signal Process. Its Appl. CSPA 2012 (Melaka, Malaysia) p 505–509

Fan M, Andrade G F S and Brolo A 2011 A review on the fabrication of substrates for surface enhanced Raman spectroscopy and their applications in analytical chemistry Anal. Chim. Acta 693 7–25

Wu X, Chen J, Li X, Zhao Y and Zughai S M 2014 Culture-free diagnostics of Pseudomonas aeruginosa infection by silver nanorod array based SERS from clinical sputum samples Nanomedicine Nanotechnology Biol. Med., 10 1863–1870

Han H W, Yan X L, Dong R X, Ban G and Li K 2009 Analysis of serum from type II diabetes mellitus and diabetic complication using surface-enhanced Raman spectra (SERS) Appl. Phys. B Lasers Opt. 94 667–672

Siak F, Lee K Y, Mansor W, and Radzol A R M 2012 Raman spectra of drop coating deposition Raman for melamine solution IEEE-EMBS Conf. Biomed. Eng. Sci. IECBES (Langkawi, Malaysia) p 815–819

Feng S et al. 2010 Gastric cancer detection based on blood plasma surface-enhanced Raman spectroscopy excited by polarized laser light Biosens. Bioelectron. 26 3167–3174

Mahmood T et al. 2018 Raman spectral analysis for rapid screening of dengue infection Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 200 136–142

Bilal M et al. 2016 Raman spectroscopy based discrimination of NS1 positive and negative dengue virus infected serum Laser Phys. Lett. 13 1–7

Jolliffe I T 2002 Principal Component Analysis, Second Edition (Spinger-Verlag New York)

Radzol A R M, Lee K Y, Mansor W and Azmin N H R 2015 SVM-RBF Model PCA Criterion Selection for Detection of NS1 Molecule From Raman Spectra of Salivary Mixture IET International Conference on Biomedical Image and Signal Processing (Beijing, China) p 1–6

Sahak R, Lee K Y, Mansor W, Yassin A I M and Zabidi A 2010 Detection of asphyxiated infant cry using support vector machine integrated with principal component analysis Proceedings of 2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (Kuala Lumpur, Malaysia) p 485–488

Radzol A R M, Lee K Y, Mansor W and Tawi F M T 2016 Signal Processing for Raman Spectra for Disease Detection Int. J. Pharm. Pharm. Sci. 8 4–10

Ismaael S, Mira A and Chourishi D 2015 Using the Extreme Learning Machine (ELM) Technique for Heart Disease Diagnosis IEEE Canada International Humanitarian Technology Conference (Ottawa, ON, Canada) p 1–3

Zhou H, Lan Y, Soh Y C, Huang G B and Zhang R 2012 Credit risk evaluation with extreme learning machine IEEE Int. Conf. Syst. Man, Cybern., (Seoul, South Korea) p 1064–1069

Radzol A R M, Lee K Y, Mansor W and Saadun N 2014 Baseline Correction Customized for NS1 Salivary Raman Spectra with Piecewise Cubic Hermite Interpolation IEEE Symp. Ind. Electron. Appl. (Sabah, Malaysia) p 1–5

Landis J R and Koch G G 1977 The Measurement of Observer Agreement for Categorical Data Biometrics 33 159–174