Africa Journal of Physical Sciences Vol. 6, pp. 45-56, June 2021
http://journals.uonbi.ac.ke/index.php/ajps/index
ISSN 2313-3317

Concurrent Machine learning Assisted Raman Spectroscopy of Whole Blood and Saliva for Breast Cancer Diagnostics

Githaiga John Irungu[1,a*], Angeyo Hudson Kalambuka[1], Kaduki Kenneth Amiga[1], Bulimo Wallace Dimbuson.[2] and Ojuka Daniel Kenyoru[3]

[1]Department of Physics, University of Nairobi, 30197 – 0100, Nairobi, Kenya.
[2]Department of Biochemistry, University of Nairobi, 30197 – 0100, Nairobi, Kenya.
[3]Department of Surgery, University of Nairobi, 30197 – 0100, Nairobi, Kenya.
[a]gjirungu@uonbi.ac.ke*, gjirungu@gmail.com*
*Corresponding author email

ARTICLE INFO

Available online: 30th June, 2021

Keywords:
Raman spectroscopy
Breast cancer
Machine learning
Whole blood
Saliva

ABSTRACT

Highly sensitive and unique biomarkers are needed for early cancer detection. In particular, biomarkers in biofluids can be useful in detecting the existence of a tumor early in the body. The utility of biofluid markers for cancer detection can be enhanced when multiple biofluids are simultaneously biochemically analyzed in order to acquire complementary information for diagnostic purposes. This work aimed at investigating the universal human whole blood and saliva biomarkers for breast cancer screening using machine learning-assisted Raman spectroscopy. Raman spectroscopy was performed in the 393 – 2063 cm^{-1} region using 785 nm laser excitation. Machine learning-assisted Raman spectroscopy was implemented by performing principal component analysis, independent component analysis, and support vector machine modeling on the Raman spectra in order to extract the underlying multivariate relationships between the observed biochemical alterations. Ten spectral regions were determined: 612 ± 1.44 cm^{-1}, 785 cm^{-1}, 968 ± 2.02 cm^{-1}, 1000 ± 0.86 cm^{-1}, 1248 cm^{-1}, 1340 cm^{-1}, 1371 ± 0.57 cm^{-1}, 1448 ± 1.73 cm^{-1}, 1500 ± 2.88 cm^{-1}, and 1661 ± 1.44 cm^{-1}, which can be regarded as universal biomarkers of breast cancer using both whole blood and saliva samples. The diagnostic models based on principal component analysis followed by support vector machine achieved mean sensitivity of 95.83 ± 2.48%, specificity of 99.16 ± 0.65%, and accuracy of 98.50 ± 0.65% when differentiating healthy blood samples from diseased blood samples. Further, this model yielded mean sensitivity of 73.0 ± 6.20%, specificity of 97.50 ± 0.67%, and accuracy of 93.66 ± 0.80% when differentiating the healthy saliva samples from diseased saliva samples. The determined biomarkers could be used to establish a spectral system for detection of breast cancer. Further work, including large sample sizes, has to be done to figure out how proteins and nucleic acids behave in their conformational states in human blood and saliva before translating the findings to actual clinical application.

ISSN 2313-3317

1. Introduction

Breast cancer is the most frequently diagnosed cancer in women, as well as the leading cause of cancer death [1]. The routine diagnostic procedures include the histopathological examination of

biopsy samples and imaging tests which are often subjective, time-consuming, costly [2] and generally applicable in revealing cancers at already advanced stages [3]. This emphasizes the importance of developing less intrusive, real time, highly sensitive, and cost-effective screening methods for patients in the early stages of cancer development. One such method may be based on Raman microspectroscopy due to its many advantages such as use of less harmful near-infrared radiation, being reagent free, is less intrusive, requires less sample preparation, has high resolution (≤ 1 µm) and minimally affected by water bands [4].

In the last decades, preference of liquid biopsy, for example, blood and saliva to tissue biopsy in biological studies has generated interest in biomedical sciences. Biofluids have distinct properties that can be objectively assessed and tested as measures of both natural and pathogenic biologic processes - so-called biomarkers [5]. In oncology, the fluid biomarkers would be components in patient's fluids that reveal the presence of cancer. This includes macromolecules that originate from tumor cells, for example, lipids, proteins, RNA, microRNA, and DNA [6]. Other fluid biomarkers include circulating cells such as circulating tumor cells, immune cells, stromal cells, and endothelial cells [5], [6].

The majority of tumors are vascularized; hence cancer biomarkers can be shed into the bloodstream [5]. Therefore, blood as one of the major fluids would be convenient for cancer biosignature detection. This convenience is further enhanced by its ease of availability, already established collection and ethical protocols, and the ability to replicate the test as many times as required to monitor disease progression or therapeutic response [5], as observed in reported breast cancer Raman studies [3], [7], [8]. Previous blood-based Raman studies on breast cancer observed Raman spectral differences in healthy and diseased breast cancer samples can be mainly attributed to biochemical changes of DNA, proteins, and lipids [3], [7], [8], [9]. However, it should be noted that many compounds found in blood, for example, growth factors, hormones, antibodies, and enzymes are also found in saliva. These compounds can move from bloodstream into salivary glands through processes of ultrafiltration, passive diffusion, or active transportation [10]. Using saliva for cancer diagnosis would be relatively practical because of its noninvasive safe collection and high-speed sampling [10]. Indeed, transcriptomic and proteomic studies aimed at discovering and pre-validating saliva biomarkers for the noninvasive detection of breast cancer have largely shown salivary proteins can be prognostic markers for breast cancer screening [11].

There are already two Raman studies available in literature in which biochemical components of salivary proteins were found significantly important for noninvasive discrimination of various grades of breast tumors [12], [13], but these studies do not focus on utility of salivary nucleic acids, lipids, and carbohydrates biochemical alterations for breast cancer diagnostics. However, it should be noted that reliability of using biofluids for disease diagnostics is a major challenge in Raman spectroscopy. For instance, the biomarker levels in saliva and biofluids exist in very low concentrations [5], [14] well below the detection limits of most analytical spectroscopic techniques, including Raman. Therefore, the underlying biomarker alterations in healthy and diseased samples can be easily understood when Raman microspectroscopy is combined with machine learning techniques.

Machine learning techniques are intelligent computational methods for extracting maximum analytical information from measured data [15]. To achieve better diagnostic accuracies in cancer diagnostics, optimized machine learning techniques may be used to identify, isolate, amplify and process the subtle analyte peaks. This enhances sensitivity, accuracy, precision and speedy non-destructiveness for in situ and in vitro diagnostic capabilities of Raman microspectroscopy.

Currently, there is a lack of research interrogating universal whole blood and saliva Raman biomarkers for breast cancer screening. Therefore, this study set out to determine the universal prominent biochemical spectral marker differences in the whole blood and saliva samples of healthy and breast cancer patients using Raman spectroscopy and machine learning techniques. The singular value decomposition principal component analysis [16] and independent component analysis [17] were used as unsupervised methods for the multivariate exploration of primary sources of heterogeneity in spectral data sets. The support vector machine [18] was added as a supervised condition on the principal component analysis to give underlying spectral features more weight in the classification. The support vector machine classifier's output was evaluated using the overall error rate estimate from cross-validation and the receiver operating characteristic parameters.

2. Materials and methods

2.1 Blood and saliva sampling

Whole blood and saliva were taken from the consenting 20 breast cancer patients and 23 healthy volunteers, all-female, at Kenyatta National Hospital, Kenya. To take part in the study, all of the participants signed a written consent form. This study was approved by the Kenyatta National Hospital-University of Nairobi Ethics and Research Committee (ERC certificate number: P112 / 03 / 2018). About 2 mL volumes of non-stimulated saliva and peripheral venous blood were collected from participants during morning hours. For blood samples, ethylenediaminetetraacetic acid was used as an anticoagulation agent, and stored at -20^{o}C. Saliva was centrifuged at 6,500 revolutions per minute for 10 minutes to remove the food debris and oral mucous epithelial cells, and stored at -4^{o}C. Based on histological pathological reports, the samples were categorized according to the reported level of breast malignancy in recruited patients - healthy (non-malignant), stage 1 (early-stage cancer), stage 2 (the cancer tumor has increased in size and has most likely spread to surrounding tissues or lymph nodes), and stage 3 (metastatic) cancer.

2.2 Raman spectroscopy measurements

A 10 µl aliquot was deposited onto conductive silver paint smeared substrates. Raman spectra were collected using a confocal Raman system from Seki Technotron Corporation, Japan. Raman measurements were performed in the 393-2063 cm^{-1} region with a 785 nm excitation laser, an exposure time of 120 seconds, and a spectral resolution of 1.35 cm^{-1} using the microscope's 80x objective. The excitation beam's measured spot size was ≈ 31.25 µm. Spectral calibration was performed using a silicon semiconductor at 520.5 cm^{-1} band. About 15-20 spectra were measured

from five random points for each sample to obtain a mean spectrum.

2.3 Spectral analysis

2.3.1 Data preprocessing

The analysis of measured spectra was performed in the 500-1800 cm^{-1} region, the considered fingerprint region for biological specimens [19]. Data denoising, smoothing, normalization, and baseline restoration were performed to remove artificial interference during the sample preparation and spectral acquisition [3], [7], [8], [9]. Spectral smoothing and autofluorescence background subtraction were done using the Vancouver algorithm based on the fifth-order polynomial fitting method [20]. Spectra were normalized to the CH_2 deformation band near 1445 cm^{-1}, using the OriginLab software.

2.3.2 Multivariate statistical analysis of biomarkers alterations in whole blood and saliva for breast cancer diagnostics

The preprocessed Raman spectra of healthy and diseased samples were combined to form single matrices, that is, X_{wxn} where w and n represented total number of wavenumbers and spectra, respectively. For whole blood samples, three matrices labeled $stage_1$ ($X_{781\,x\,518}$), $stage_2$ ($X_{781\,x\,573}$), and $stage_3$ ($X_{781\,x\,671}$) were obtained, while matrices for saliva samples were $stage_1$ ($X_{781\,x\,559}$), $stage_2$ ($X_{781\,x\,611}$), and $stage_3$ ($X_{781\,x\,775}$). The subscripts 1, 2, and 3 identified the stages of breast malignancy. To investigate the universal prominent spectral markers amongst whole blood and saliva samples, all the six matrices were combined into one spectral matrix ($X_{781\,x\,3707}$) then subjected to independent component analysis by maximum likelihood fast fixed-point estimation algorithm. First, the algorithm worked by mean centering and whitening the spectral data, then followed by several iterations until convergence leading to determination of independent components was attained. The independent component analysis algorithm was based on the following expression [21]:

$$W^+ = W + diag(\alpha_i)[diag(\beta_i) + E\{g(y)y^T\}]W \qquad (2.1)$$

where $y = Wx,\ \beta_i = -E\{y_i g(y_i)\}$, and $\alpha_i = -1/(\beta_i - E\{g'(y_i)\}$. In this case, the matrix W needs to be orthogonalized after every step in a symmetrical manner. The convergence speed can be optimized by careful choice of matrices $diag(\alpha_i)$ and $diag(\beta_i)$. In this study, the performance of the method was optimized by choosing a suitable nonlinearity g, where the nonlinearity g function; $g(u) = u^3$ was chosen due to its optimal performance. The decorrelation approach based on deflation technique was used where the independent components were estimated one-by-one. Moreover, the stabilized capability of the fixed-point algorithm was incorporated to ensure algorithm convergence. The maximum number of iterations were set at 1000.

For discriminating between Raman spectral data sets, singular value decomposition principal component analysis by cross-validation [16] followed by support vector machine [18] was used. With singular value decomposition principal component analysis, three matrices; U representing scores, S

representing scores' size, and V representing loadings were obtained, according to the following equation [22]:

$$M_{mn} = U_{mm} S_{mn} {V_{nn}}^T \qquad (2.2)$$

where $U^T U = 1$; $V^T U = 1$, subscripts *m*, and *n* represented number of samples and wave numbers, respectively. In this study, matrices U, S, V yielded information regarding spectra discrimination, number of principal components, and the wavenumbers, respectively.

For support vector machine, the general fitting function used was [23]:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \, K(x, x_i) + \beta_o \qquad (2.3)$$

where $K(x, x_i)$ is a kernel function; x_i is the training sample eigenvector; and x is the recognizing sample eigenvector [18]. The parameter α_i is restricted to $0 \leq \alpha_i \leq C$ and can be estimated by maximizing a Lagrangian, while parameter C is the cost parameter that determines the amount of regularization, that is, the classification error term [24]. In this study, we used the linear and radial basis support vector machines models based on equations (2.4) and (2.5), respectively [18], [23].

$$K(x, x_i) = x \cdot x_i \qquad (2.4)$$
$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2) \qquad (2.5)$$

The data analysis looked for the best function parameter γ, as well as the best regularization parameter, C, for the efficient optimization process. The optimal γ and C values were sorted in such a way that the effect was sufficient to create a decision surface without misclassifying the training set, hence reducing over-prediction. The γ and C were obtained from the grid search for the highest cross-validation accuracy. In order to increase model accuracy, the 10 folds cross-validation was used during data analysis.

The linear and radial basis support vector machines models based on 5 to 10 principal components, cost = 100, and kernel parameters = 0.57 to 0.8, were used to generate the scatterplots of the control and diseased groups. To verify the effectiveness of the support vector machine diagnostic models in predicting diseased and healthy samples, two parameters; sensitivity and specificity were calculated. In this study, the ability of a test method to detect disease in diseased subjects was described by sensitivity parameters, while the ability to detect the absence of disease in healthy subjects was described by specificity parameters [9]. All multivariate analysis procedures were done in the MATLAB 2018b scripting environment.

3. Results and Discussions

The photomicrographs of typical blood and saliva samples taken at 50x magnification are shown in Figure 1 (a) and 1 (b), respectively. The characteristic pinkish-red color of the blood smear can be seen, a factor attributed to hemoglobin due to iron compounds. The picture of saliva depicts crystalline structures conjoined with each other in a tree or fern-like shape from the center of the

drop, as shown in Figure 1 (b). Other research [25] has indicated that morphological changes in dried oral saliva fluid are an indicator of quantitative and consistency molecular composition deviation.

For independent component analysis, 10 eigen vectors (variance ≈ 98%) of spectral datasets were fed to the maximum likelihood fast fixed-point estimation algorithm. Three spectra; also termed as independent components were obtained, all shown in Figures 1 (c), (d) and (e). These independent components revealed the positive as well as negative peaks' positions that explained the wavelengths representing biomarkers that could be associated with breast cancer.

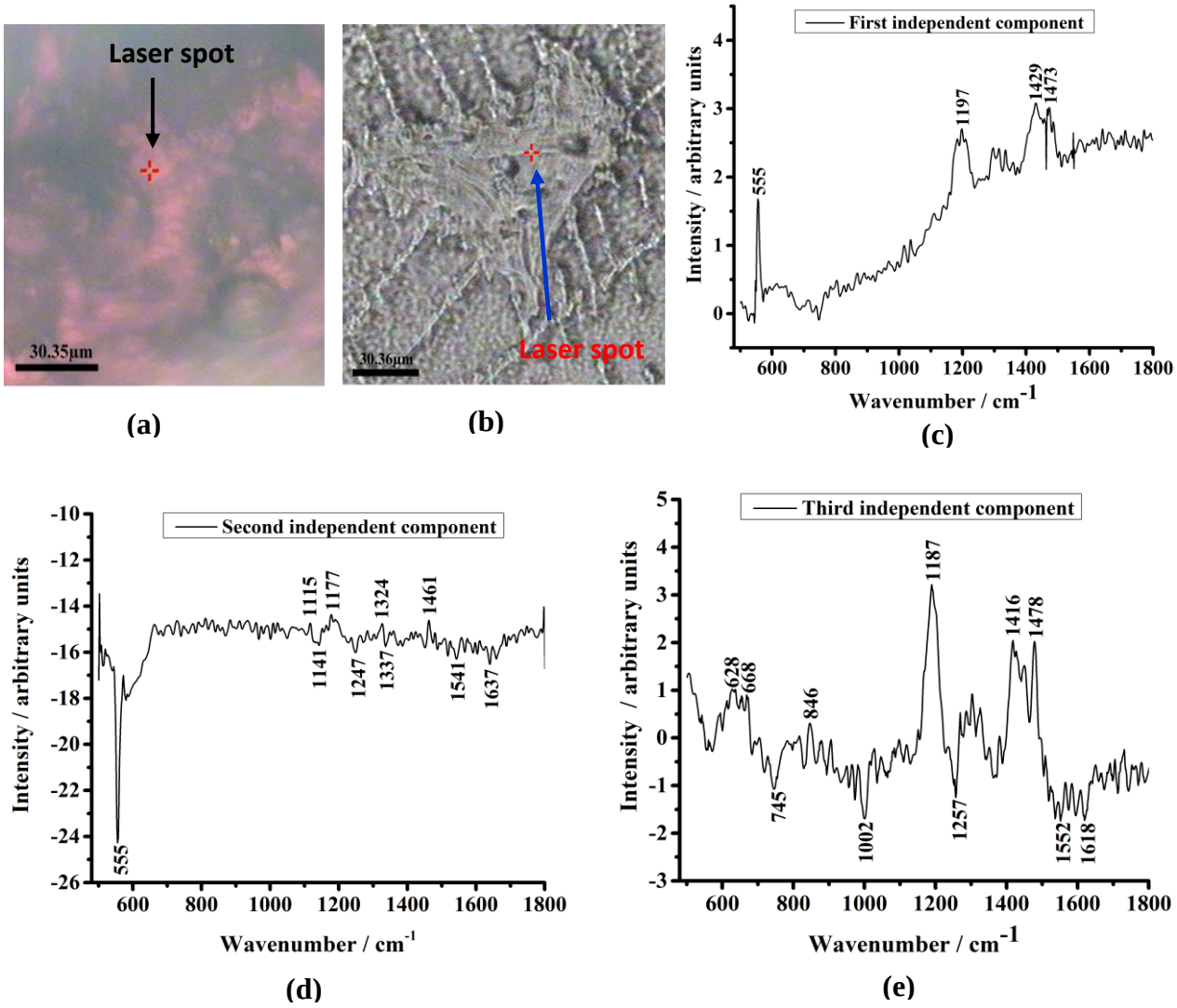

Figure 1. The photomicrographs of (a) dried blood sample, (b) dried saliva sample, and (c), (d), (e) the first three independent components that reveal breast cancer markers' wavelengths.

In addition, the difference spectra between the normalized mean baseline corrected spectrum of diseased samples (n = 20) and that of healthy samples (n = 23) were determined and plotted. The difference spectrum of blood and saliva samples are shown in Figure 2(a) and Figure 2 (b), respectively. The positive bands in the difference spectrum explained biochemical changes found in higher concentrations in breast cancer samples, while negative bands explained biochemical

changes found in higher concentrations in healthy samples. The biochemical assignments of peaks were done in accordance with the Raman spectroscopy of tissues, body fluids, or bio-molecules as reported in the literature. This was done in consideration of position and possible wavelength differences of each particular Raman band. Due to possibility of varying detection conditions and experimental errors, the Raman bands within ± 10 cm^{-1} were considered to represent the same Raman peak [26]

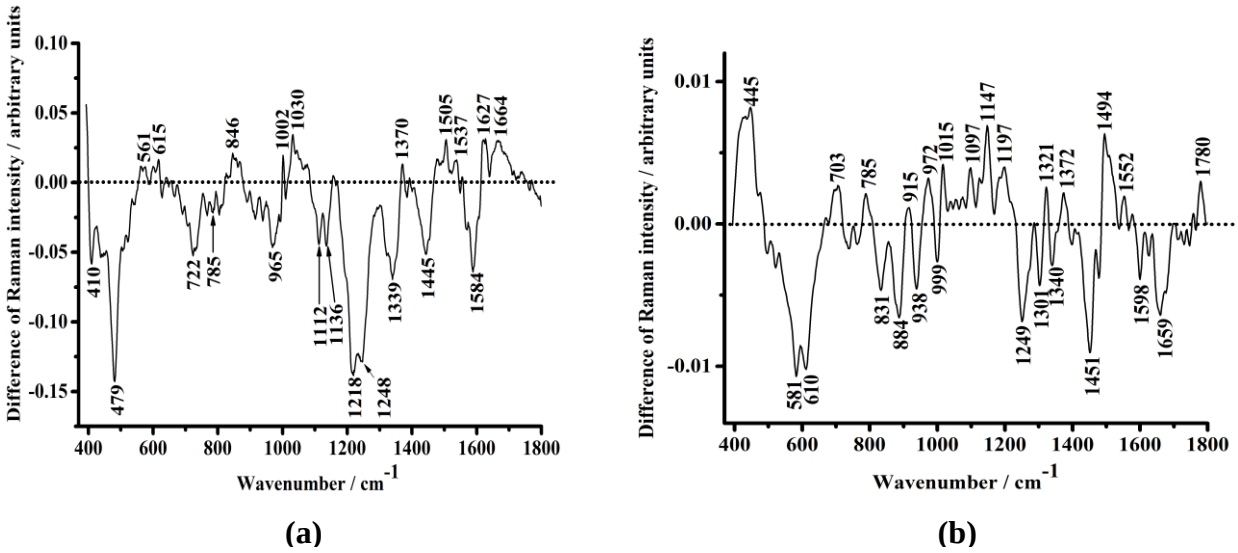

Figure 2. The difference Raman spectrum of (a) whole blood and (b) saliva samples.

It can be seen that the first and second independent components revealed a peak at 555 cm^{-1}. This spectral marker pointed to the presence of biochemical changes due to nucleic acid bases which include adenine, thymine, uracil, and guanine [27]. This band is also observed as a positive peak in the difference spectrum of blood samples in Figure 2 (a) but blue shifted to 561 cm^{-1}, and can therefore be associated with specific biochemical changes occurring in blood samples of diseased patients. This finding imply there was pronounced biochemical changes due to nucleic acid components in whole blood samples of breast cancer patients. Examination of the first and third independent components showed strong peaks at 1197 cm^{-1} and 1187 cm^{-1}, respectively. This peak is also observed at 1197 cm^{-1} as a positive peak in the difference spectrum of saliva samples in Figure 2 (a). These markers are linked to the presence of anti-symmetric phosphate vibrations associated with nucleic acids compounds [3]**,** [27], indicating that changes in salivary nucleic acid components were a dominant factor during breast cancer progression. It is observed the 1473 cm^{-1} and 1478 cm^{-1} peaks are pronounced bands in the first and third independent components, respectively. These bands indicated presence of CH_2 and CH_3 deformations associated with lipids and collagen proteins [28], [29], though it is unclear if the biochemical changes occurred in the healthy or diseased samples. Although the 1416 and 1429 cm^{-1} peaks had > ± 10 cm^{-1} difference, both bands can be attributed to CH2 scissoring vibrations of lipids in healthy samples [7]**,** [29], as evidenced in negative spectrums of Figures 2(a) and 2(b). Further analysis of all independent components revealed presence of mixed biochemical alterations due changes arising from nucleotide conformations at

628 cm^{-1}, 668 cm^{-1}, 745 cm^{-1}, 1141 cm^{-1}, 1247 cm^{-1}, 1257 cm^{-1}, 1324 cm^{-1}, 1337 cm^{-1} [3], [27], [28], [29], [30], tryptophan, tyrosine and phenylalanine proteins at 1002 cm^{-1}, 1171 cm^{-1}, 1247 cm^{-1}, 1337 cm^{-1}, 1541 cm^{-1}, 1552 cm^{-1}, 1618 cm^{-1}, 1637 cm^{-1} [7], [9], [27], [30], phospholipids at 745 cm^{-1}, 1324 cm^{-1}, and saccharides at 846 cm^{-1}, 1115 cm^{-1} [27], [30]. It can therefore be concluded that alterations of nucleic acids and proteins were predominant in whole blood and saliva of healthy and breast cancer patients.

To better understand the biochemical alterations and their differences in the studied samples, the spectral differences of whole blood and saliva bio-samples of the breast cancer group and the healthy group were compared. In consideration that the focus of this study was to determine the universal prominent biochemical spectral marker differences in the whole blood and saliva samples of healthy and breast cancer patients, we examined the difference spectrum of blood and saliva samples in Figures 2 (a) and 2 (b). Ten major peaks at the wavelengths 612 ± 1.44 cm^{-1}, 785 cm^{-1}, 968 ± 2.02 cm^{-1}, 1000 ± 0.86 cm^{-1}, 1248 cm^{-1}, 1340 cm^{-1}, 1371 ± 0.57 cm^{-1}, 1448 ± 1.73 cm^{-1}, 1500 ± 2.88 cm^{-1}, and 1661 ± 1.44 cm^{-1} were identified. We consider these bands to be the universal spectral markers that differentiated diseased and healthy groups using both whole blood and saliva samples. These Raman peaks were attributed to proteins and nucleic acid bases alterations [3], [7], [9], [27], [30]. The protein bands were based on aromatic amino acids which include tryptophan, phenylalanine, and tyrosine, while nucleic acid alterations were mainly due to the nucleic acid bases of adenine, guanine, thymine and cytosine.

Examination of Figure 2(a) and 2(b) shows the peak intensities were greater for the breast cancer group than for the control group at 1371 ± 0.57 cm^{-1} and 1500 ± 2.88 cm^{-1}, while the bands at 1248 cm^{-1}, 1340 cm^{-1}, 1448 ± 1.73 cm^{-1}, and 1591 ± 4.04 cm^{-1} were more intense in the spectra of control patients than for the breast cancer group. It can be concluded that nucleic acid levels were greater in blood and saliva samples from breast cancer patients as compared to the control group. Other serum-based studies have shown spectral features of nucleic acids points to presence and progression of breast cancer [3], [7]. In contrast, lipids, amide and phenylalanine protein levels were greater in blood and saliva samples from control patients as compared to the diseased group. This is in agreement with previous studies [7], [8], [31] where spectral markers attributed to biochemical changes of proteins and lipids at 1340 cm^{-1} and 1440-1460 cm^{-1} were found more intense in the spectra of control patients than for the breast cancer group. It is also noted that biochemical alterations corresponding to cholesterol esters at 612 ± 1.44 cm^{-1}, phenylalanine at 1000 ± 0.86 cm^{-1}, and amide I at 1661 ± 1.44 cm^{-1} were prominent in blood samples of breast cancer patients and in saliva samples of control patients, which agrees with other studies [9], [31]. Similarly, biochemical changes due to ring breathing modes in the nucleic acid bases at 785 cm^{-1}, proteins and nucleic acids at 968 ± 2.02 cm^{-1}, and glycogen at 1141 ± 3.17 cm^{-1} were predominant in saliva samples of breast cancer patients and blood samples of control patients. Detailed Raman assignments regarding other Raman markers in Figures 2(a) and 2 (b) are provided elsewhere [27], [28], [29], [30].

Multivariate analysis based on support vector machine algorithm benefits from feature extraction techniques such as principal component analysis, which serves to reduce the curse of dimensionality

[7]. In this study, 5 to 10 principal components were found suitable for developing an efficient support vector machine diagnostic model. Furthermore, classification errors were found to decrease as the number of principal components increased, implying that the remaining number of principal components accounted for a substantial portion of the variance, which explained the most significant biochemical information [32]. The diagnostic results of singular value decomposition principal component analysis followed by the selected support vector machine models are shown in Table 1.

Table 1. Diagnostic results of singular value decomposition principal component analysis followed by support vector machine diagnostic model on the Raman spectra of whole blood and saliva samples from healthy and breast cancer patients

		Blood			Saliva		
Disease status	**SVM function**	**Accuracy**	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**
Stage 1	Linear	100%	100%	100%	94%	46%	100%
	RBF	100%	98%	100%	94%	69%	97%
Stage 2	Linear	100%	100%	100%	94%	77%	98%
	RBF	100%	98%	100%	94%	78%	98%
Stage 3	Linear	93%	84%	96%	90%	76%	95%
	RBF	98%	95%	99%	96%	92%	97%

Abbreviations: RBF, Radial basis function; SVM, support vector machine; Stage 1, early-stage cancer; stage 2, cancer tumor has increased in size and has most likely spread to surrounding tissues or lymph nodes; Stage 3, metastatic cancer.

It can be seen that various values of diagnostic accuracy, sensitivity, and specificity were achieved depending on the samples and SVM functions in consideration. The developed SVM model achieved mean sensitivity of 95.83 ± 2.48%, specificity of 99.16 ± 0.65%, and accuracy of 98.50 ± 0.65% when differentiating healthy blood samples from diseased blood samples. Similarly, the SVM model yielded mean sensitivity of 73.0 ± 6.20%, specificity of 97.50 ± 0.67%, and accuracy of 93.66 ± 0.80% when differentiating the healthy saliva samples from diseased saliva samples. These results confirm the outstanding accuracy of the principal component analysis followed by support vector machine diagnostic algorithm for breast cancer detection.

The sensitivity and specificity diagnostic parameters were greater when blood spectral data sets were used for diagnosis, indicating there was a reasonably good differentiation of the healthy and diseased sample's spectra. The radial basis support vector machine kernel model performed better than the linear support vector machine kernel model in diagnosing late breast malignancy. This finding implies that the linear separable characteristic nature of spectral datasets [33] decreased with malignancy, hence the radial basis function was appropriately useful for handling non-linear data [34], leading to better diagnostic accuracy. Moreover, support vector machine model based on radial basis function had greater sensitivity diagnostic parameters, particularly in diagnosing breast cancer based on saliva samples. It should be noted that, in comparison to linear kernels which are parametric functions,

radial basis function is a squared exponential function that is a non-parametric function. The non-parametric function defines a function space that is a lot larger than that of the linear kernel or the polynomial kernel, and can therefore be viewed as powerful as an infinite order polynomial kernel [34]. Therefore, the good diagnostic performance of radial basis kernel function in our study can be attributed to its nonparametric capability in handling complex data [33], [34]. This makes it suitable for handling a high degree of collinearity in the spectroscopic datasets.

4. Conclusion

The spectral analysis of blood and saliva spectrum revealed ten wavelengths attributed to proteins and nucleic acid markers at 612 ± 1.44 cm^{-1}, 785 cm^{-1}, 968 ± 2.02 cm^{-1}, 1000 ± 0.86 cm^{-1}, 1248 cm^{-1}, 1340 cm^{-1}, 1371 ± 0.57 cm^{-1}, 1448 ± 1.73 cm^{-1}, 1500 ± 2.88 cm^{-1}, and 1661 ± 1.44 cm^{-1}, which can be regarded as universal dominant biomarkers of breast cancer. The diagnostic models based on principal component analysis followed by support vector machine achieved mean sensitivity of 95.83 ± 2.48%, specificity of 99.16 ± 0.65%, and accuracy of 98.50 ± 0.65% when differentiating healthy blood samples from diseased blood samples. The same model yielded mean sensitivity of 73.0 ± 6.20%, specificity of 97.50 ± 0.67%, and accuracy of 93.66 ± 0.80% when differentiating the healthy saliva samples from diseased saliva samples. The overall diagnostic accuracies of radial basis support vector machine kernel model were greater than that of linear support vector machine classifier models, proving that the diagnostic performance of radial basis function is superior to that of a linear function in support vector machine algorithms. The findings of this study show that Raman spectroscopy combined with multivariate analysis diagnostic algorithms has a lot of potential for non-invasive, label-free breast cancer detection using whole blood and saliva samples.

Acknowledgments

We acknowledge the Swedish International Development Cooperation Agency (SIDA), through the International Science Programme (ISP), Uppsala University, for financial support. The first author thanks the National Research Foundation of Kenya (NRF) for financial support, and Peninah Kabethi (Kenyatta National Hospital) for patient recruitment.

References

[1] A. Bouzalmat and J. Kharroubi, “Comparative Study of PCA, ICA, LDA using SVM Classifier,” *Emerging Technologies in web Intelligence*, vol. 6, no. 1, pp. 64–68, 2014.

[2] A. Chandra, S. Talari, Z. Movasaghi, and S. Rehman, “Raman Spectroscopy of Biological Tissues,” *Applied Spectroscopy Review*, vol. 50, no. 1, pp. 46–111, 2015.

[3] A. Hyvärinen, and E. Oja, “Independent Component Analysis: Algorithms and Applications,” *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.

[4] A. J. Freddie Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, and L. A. Torre, “Global cancer statistics 2018 : GLOBOCAN Estimates of incidence and mortality worldwide for 36 Cancers in 185 Countries,” *Cancer Journal of Clinicians*, vol. 68, pp. 394–424, 2018.

[5] B. Han, Y. Du, T. Fu, Z. Fan, S. Xu, C. Hu, L. Bi, T. Gao, H. Zhang, and W. Xu, “Differences and Relationships Between Normal and Atypical Ductal Hyperplasia, Ductal Carcinoma in situ, and Invasive Ductal Carcinoma Tissues in the Breast Based on Raman Spectroscopy,” *Applied Spectroscopy*, vol. 71, no. 2, pp. 300–307, 2017.

[6] C. B. Y. Cordella, "PCA : The Basic Building Block of Chemometrics," pp. 1–46, 2012.

[7] E. Vargas-Obieta, J. C. Martínez-Espinosa, B. E. Martínez-Zerega, L. F. Jave-Suárez, A. Aguilar-Lemarroy, and J. L. González-Solís, "Breast cancer detection based on serum sample surface enhanced Raman spectroscopy," *Lasers in Medical Science*, vol. 31, no. 7, pp. 1317–1324, 2016.

[8] H. Byrne, L. Kerr, and B. M. Hennelly, "Optimal choice of sample substrate and laser wavelength for Raman spectroscopic analysis of biological specimen," *Analytical Methods*, vol. 7, no. 1, pp. 5041–5052, 2015.

[9] H. Chen, Z. Lin, and C. Tan, "Cancer Discrimination Using Fourier Transform Near-Infrared Spectroscopy with Chemometric Models," *Journal of Chemistry*, vol. 2015, pp. 1–9, 2015.

[10] H. F. Nargis, H. Nawaz, A. Ditta, T. Mahmood, M.I. Majeed, N. Rashid, M. Muddassar, H. N. Bhatti, M. Saleem, K. Jilani, F. Bonnier, and H.J. Byrne, "Raman spectroscopy of blood plasma samples from breast cancer patients at different stages," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 222, p. 117210, 2019.

[11] I. Rehman, Z. Movasaghi, and S. Rehman, "Vibrational spectroscopy for tissue analysis," CRC Press, pp. 214–287, 2013.

[12] I. T. Jolliffe, Principal Component Analysis, Second Edition. New York: Springer - Verlag, pp. 111-118, 2002.

[13] J. De Gelder, K. De Gussem, P. Vandenabeele, and L. Moens, "Reference database of Raman spectra of biological molecules," *Raman Spectroscopy*, vol. 38, pp. 1133–1147, 2007.

[14] J. Zhao, H. Lui, D. I. McLean, and H. Zeng, "Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy," *Applied Spectroscopy*, vol. 61, no. 11, pp. 1225–1232, 2007.

[15] K. J. Martin, M. V Fournier, G. P. V. Reddy, and A. B. Pardee, "A need for basic Research on fluid-based early detection biomarkers," *Cancer Research,* vol. 70, pp. 5203–5207, 2010.

[16] K. Varmuza, and P. Filzmoser, *Introduction to multivariate statistical analysis in Chemometrics*. Boca Raton: CRC Press, pp. 1-20, 2008.

[17] L. McCreely, Raman Spectroscopy for Chemical Analysis. John Wiley and Sons, Ltd, pp. 73-83, 91-92, 2001.

[18] M. Bilal, S. Tabassum, M. Saleem, H. Mahmood, U. Sarwar, H. Bangush, F. Munir, M. A. Zia, M. Ahmed, S. Shahzada, and E. U. Khan, "Optical Screening of Female Breast Cancer from Whole Blood Using Raman Spectroscopy," *Applied Spectroscopy*, vol. 71, no. 5, pp. 1004-1013, 2017.

[19] M. Boiret, D. N. Rutledge, N. Gorretta, Y. M. Ginot, and J. M. Roger, "Application of independent component analysis on Raman images of a pharmaceutical drug product: pure spectra determination and spatial distribution of constituents," J*ournal of Pharmaceutical and Biomedical Analysis, Elsevier*, vol. 90, pp. 78–84, 2014.

[20] M. J. Baker, S. R. Hussain, L. Lovergne, V. Untereiner, C. Hughes, R. A. Lukaszewski, G. Thie´finbg, and G. D. Sockalingum, "Developing and understanding biofluid vibrational spectroscopy: A critical review," *Chemical Society. Reviews*, vol. 45, no. 7, pp. 1803–1818, 2016.

[21] M. V. P. Chowdary, K. K. Kumar, J. Kurien, S. Mathew, and C. M. Krishna, "Discrimination of Normal, Benign, and Malignant Breast Tissues by Raman Spectroscopy," *Biopolymers*, vol. 83, pp. 556–569, 2006.

[22] M.-G. C. Pichardo-Molina J. L., Frausto-Reyes C., Barbosa-GarcÍa O., Huerta-Franco R., González-Trujillo J. L., RamÍrez-Alvarado C. A., and Gutiérrez-Juárez G., "Raman spectroscopy and multivariate analysis of serum samples from breast cancer patients," *Lasers in Medical Science*, vol. 22, pp. 229–236, 2007.

[23] N. Christodoulides, S. Mohanty, C. S. Miller, M. C. Langub, P. N., Floriano, P. Dharshan, M. F. Ali, B. Bernard, D. Romanovicz, E. Anslyn, P. C. Fox, and J. T. McDevitt, "Application of microchip assay system for the measurement of C-reactive protein in human saliva," *Lab on a Chip*, vol. 5, no. 3, pp. 261–269, 2005.

[24] R. Singla, B. Chambayil, A. Khosla, and J. Santosh, "Comparison of SVM and ANN for classification of eye events in EEG," Biomedical Science and Engineering, vol. 4, pp. 62– 69, 2011.

[25] S. Feng, S. Huang, D. Lin, G. Chen, Y. Xu, Y. Li, Z. Huang, J. Pan, R. Chen, and H. Zeng, "Surface-enhanced Raman spectroscopy of saliva proteins for the noninvasive differentiation of benign and malignant breast tumors," International Journal of Nanomedicine, vol. 10, pp. 537–547, 2015.

[26] S. Gonchukov, A. Sukhinina, D. Bakhmutov, and S. Minaeva, "Raman spectroscopy of saliva as a perspective method for periodontitis diagnostics," *Laser Physics Letters*, vol. 9, no. 1, pp. 73–77, 2012

[27] S. Mika, and K. Tsuda, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on neural networks*, vol. 12, no. 2, pp. 181–201, 2001.

[28] S. Varghese, A. Reveendran, V. S. Kumar, K. Tm, and V. Ranganathan, "Micro Raman Spectroscopic Analysis on Blood Serum Samples of Ductal Carcinoma Patients," *Asian Journal of Pharmaceutical and Clinical Research*, vol. 11, no. 9, pp. 176–178, 2018.

[29] T. Pfaffe, J. Cooper-white, P. Beyerlein, K. Kostner, and C. Punyadeera, "Diagnostic Potential of Saliva: Current State and Future Applications," *Clinical Chemistry*, vol. 57, no. 5, pp. 675-687, 2011.

[30] W. L. Martinez and A. R. Martinez, Exploratory Data Analysis with MATLAB ®, London, United Kingdom: CRC Press, pp. 39, 42-49, 2005.

[31] W. Wu, H. Gong, M. Liu, G. Chen, and R. Chen, "Noninvasive Breast Tumors Detection based on Saliva Protein Surface Enhanced Raman Spectroscopy and Regularized Multinomial Regression," *2015 8th International Conference on Biomedical Engineering and Informatics*, no. 1, pp. 214-218, 2015.

[32] Y. U. N. X. Ci, T. I. Y. U. Gao, J. U. N. Feng, and Z. Q. Guo, "Fourier Transform Infrared Spectroscopic Characterization of Human Breast Tissue: Implications for Breast Cancer," *Applied Spectroscopy*, vol. 53, no. 3, pp. 312–315, 1999.

[33] Y. Zhang, J. Sun, C. Lin, E. Abemayor, M. B. Wang, and D. Tw, "The Emerging Landscape of Salivary Diagnostics," *OHDM*, vol. 13, no. 2, pp. 200–210, 2014.

[34] Z. Movasaghi, S. Rehman, and I. U. Rehman, "Raman Spectroscopy of Biological Tissues," *Applied Spectroscopy Review*, vol. 42, pp. 493–541, 2007.